Decoding Attentional Control from Noninvasive Measures in Humans

Submitted in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Wenkang An

B.Eng. Electronic and Information Engineering, The Hong Kong Polytechnic University
M.S. Electrical and Electronic Engineering, The University of Hong Kong
M.S. Biomedical Engineering, The Chinese University of Hong Kong

Carnegie Mellon University
Pittsburgh, PA

December 2021

ACKNOWLEDGEMENTS

This dissertation is not a one-man job. It belongs to a whole community of mentors, colleagues, friends and family who helped me along the journey. I cannot come this far without you.

To my advisor, **Prof. Barbara Shinn-Cunningham**, thank you for being a great scientist, mentor and cheerleader. You have given me all the patience, motivation and support that I ever need, and you have the magic to restore my inner peace every time I talk with you. You are my role model and everything I want to be. It is an honor and my great fortune to have worked with you.

I would like to express my sincere gratitude to the rest of my thesis committee, **Dr. Abigail Noyce**, **Prof. Lori Holt** and **Dr. Pulkit Grover**, for being more than generous with their expertise and precious time. This work would not have been in its current shape without your insightful comments and encouragement. You bring out the best in me. Thank you for making this happen.

I would like to thank the **Office of Naval Research** (under project number N00014-18-1-2069) for funding the research in this thesis.

I would like to thank my colleagues at Microsoft Research, **Dr. Ivan Tashev**, **Dr. Hannes Gamper** and **Dr. Dimitra Emmanouilidou**, for hosting me for two summers. Each internship experience with you was so memorable and rewarding. And thank you for giving me the opportunity to conduct research on brain-computer interface, my very first interest in neuroengineering.

I would like to thank my co-workers who contributed immensely to this project. I am extremely grateful for having **Alexander Pei** as a close collaborator on this thesis project. Thank you for your great help in data collection and brainstorming together for new ideas. I owe special thanks to **Dr. Scott Bressler** and **Dr. Lia Bonacci** for recording the stimuli used in my experiments; to **Dr. Jason Bohland** for helping me setup fMRI data collection;

and to **Yangyang Raymond Xia** for giving me a lot of advice on deep learning. This thesis would not happen without your help.

I would also like to thank my lab mates at CMU and BU, past mentors and colleagues, and dearest friends for giving me feedbacks and comments on my work and supporting me in life. Each one of you taught me something special. Thank you for being in my life.

Last but not the least, I would like to thank my family. Thank you, my parents, for always being supportive. You raised me right, and gave me the freedom to pursue my dreams. I wish I have made you proud. Thank you, **Eva**, my daughter, for bringing new purposes to my life and distracting me from stress and anxiety. Your "paper-thesis-rock" joke really lightened my mood and helped me get through the final stages of thesis writing.

I dedicate this work to my wife, **Linxi Wang**. Meeting you and marrying you is the single best thing that ever happened to me. I would not have achieved all these without your sacrifice, motivation, caring and love. You are my best friend, my shelter, and my doctor. You make me a better person, day by day, and it is impossible for me to pay you back. I love you.

ABSTRACT

Auditory selective attention enables us to focus on one sound within a mixture of noises. Unlike in vision, where we can easily shift attention by moving the eyes, auditory attention can only be achieved covertly via cognitive control. That is, we change our internal attentional state in the brain in order to attend to or ignore a sound object. This control is effortless and swift for healthy people, but can be challenging for people with neurological conditions such as attention deficit/hyperactivity disorder or autism. This makes it important to study the neural mechanisms that underlie auditory selective attention.

Neuroimaging technologies allow scientists to study brain activity without invasive procedures. Electroencephalography (EEG), a measure of electrical potential at the scalp, has become a popular imaging modality for neuroscience studies due to its simple setup, low cost, and high sampling rate, allowing us to capture rapid changes in electrical signals given off by the brain. EEG can thus track brain dynamics with high temporal resolution, but it is difficult to localize the location in the brain that is generating the measured activity. Another neuroimaging modality, functional magnetic resonance imaging (fMRI), measures blood oxygenation levels, which change locally when metabolic activity in a particular brain region increases. However, this change takes several seconds. fMRI signals provide millimeter-level spatial resolution, but poor temporal resolution. Neither of these two modalities, nor any other noninvasive neuroimaging methods currently available, can simultaneously achieve both high temporal resolution and high spatial resolution.

Effectively combining the information from EEG and fMRI could allow one to determine both when and where in the brain control of auditory attention happens. One technique for fusing neuroimaging modalities is representational similarity analysis (RSA). First, the information in brain signals is summarized via the difference among all pairs of experimental conditions, reflecting the information carried by the underlying neural representation. The resulting similarity matrix has the same dimension and scale regardless of whether it was derived from EEG or fMRI, and thus can be used to integrate information in these two neuroimaging modalities.

Here I report results of an experiment that required different types of auditory selective attention (attention to space and attention to acoustic features). I collected EEG and fMRI data from healthy young adults performing these tasks. From EEG, I discovered that the event-related time course of both raw EEG voltages and alpha oscillations (8 – 14 Hz) change reliably as subjects engage in auditory attention; these show that neural representations of attentional state change as a function of time. From fMRI, I identified several brain regions in the frontal (including superior and inferior precentral sulcus and inferior frontal sulcus), parietal (including intraparietal sulcus and superior parietal lobule), temporal (superior temporal gyrus) and occipital (primary visual cortex) lobe that are actively engaged during auditory attention. This allowed me to extract neural representations of attentional control as a function of location in the brain. I then conducted an RSA to fuse EEG and fMRI results and reveal the dynamics of information in different brain regions across the course of each trial. Finally, as an attempt to translate these neuroscience findings into real-life applications, I explored the feasibility of decoding attention from single-trial EEG signals in order to develop an attention-based brain-computer interface (BCI) system.

This dissertation identified important neural signatures in EEG signals that are evoked or induced by auditory selective attention, as well as a brain network that seems to be associated with these signatures. It is among the very first studies to adopt RSA fusion techniques to study information flow through the brain during attentional control, which could be an important reference for future studies in this field. Finally, this work examined several different ways to decode attention from single-trial EEG signals, achieved promising results from these attempts, and suggested possible ways to improve for future BCI development.

TABLE OF CONTENTS

Chapter

LIST OF TABLES

Chapter 5 **90**

LIST OF FIGURES AND ILLUSTRATIONS

Chapter 1 **1**

Chapter 2 **11**

Chapter 3 **44**

Chapter 4 **76**

Chapter 5 **90**

Chapter 1

Introduction

1.1 Overview of document

In our daily life, we are exposed to a mixture of all kinds of sounds — the conversation you are in with a colleague, people chatting in the background, noise of a bus running by, etc. This creates a seemingly overwhelming auditory scene, in which the signal we desire is buried in abundant unwanted noises. However, most people find it effortless to focus on one target sound and ignore the distractors, thanks to our auditory selective attention. Unlike in vision, where we can easily shift attention by moving the eyes, auditory attention can only be achieved covertly via cognitive control. That is, we change our internal attentional state in the brain in order to attend to or ignore a sound object [1]. This control is natural and swift for healthy people, but can be challenging for people with neurological conditions such as attention deficit/hyperactivity disorder [2], [3] or autism spectrum disorder [4], [5]. This makes it important to study the neural mechanisms that underlie auditory selective attention.

When we are in a busy acoustic scene, there are two major strategies that can help us better focus. First, if we know the location of a target sound (e.g., cars running past us from the left when we are riding a bike), we can tune our attention more to that location for a better perception of the sound. And we call this strategy attention to space, or spatial attention. Alternatively, if we are unsure about the target's location, but are certain about the acoustic features of the target sound (e.g. a person with a known voice chit-chatting in a noisy cafe), we can selectively attend to a sound object with that acoustic feature. And we call this strategy attention to acoustic features, or more generally, non-spatial attention. The neural mechanism behind spatial and non-spatial auditory attention has been studied via different methods, and there is a broad agreement that there exists a dorsal pathway, where the spatial information of sound (the "where" information) is primarily processed, and a ventral pathway, where the non-spatial information of sound (the "what" information) is encoded [6]. However, the exact role of brain areas along these pathways in spatial and non-spatial auditory attentional control remains unclear. In this dissertation, I aim to 1) use two noninvasive neuroimaging modalities, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), to unveil the role of brain regions in spatial and non-spatial auditory attention, and their dynamics over the course of the experiment, and 2) decode attentional control from single-trial EEG data for the design of a brain-computer interface (BCI) system.

In **Chapter 2**, I design a condition-rich experiment that requires spatial or non-spatial auditory attention, and record EEG signals while the listeners participate in these tasks. I decode attention from EEG signals and extract representational dissimilarity features from the EEG time course and alpha oscillation power. Then I compare these features with ideal conceptual models or behavioral performance. I identify time intervals in which particular contrasts in attentional state, such as the difference between attention types or between attention to different locations, have strong representation in the EEG time course or in its alpha power. I also reveal that the listener's behavioral performance in the attention task is significantly and positively correlated with the P2 amplitude evoked by the target syllable.

In **Chapter 3**, I adopt the same experimental design as in **Chapter 2**, and record fMRI data while the listeners participate in the task. I identify an extended attention network, in which individual brain regions show different specialization in spatial or nonspatial attention. I also extract representational dissimilarity features from each voxel, and compare these features with ideal conceptual models or behavioral performance. I discover that the regions in the medial occipital lobe encode the spatial information of auditory attention, and the right IFS is the sole region that encodes information of the gender / pitch of the attended talker. The neural representation of the parietal regions are correlated with the behavioral performance, demonstrating their important role in spatially demanding tasks.

In **Chapter 4**, I correlate the time-wise EEG representations (derived from **Chapter 2**) with the voxel-wise fMRI representations (derived from **Chapter 3**) to search for corresponding information between these two imaging modalities. This analysis depicts the dynamics of attentional control in each brain region during the experiment, i.e. a spatially and temporally resolved information flow map during auditory attention.

In **Chapter 5**, I present four different studies in which I explored the feasibility of decoding attention from EEG signals for the design of an attention-based BCI system. In these studies, I explore the BCI design with different features (e.g., the EEG time course or induced oscillations), classifiers (e.g., support vector machine or convolutional neural network), stimuli (e.g., human-voiced syllables, sequence of tones or music) and form factors (e.g., full-sized gel-based EEG system or EEG “headphones”). Together, these studies demonstrate means to improve an auditory BCI design with better accuracy, efficiency and user-friendliness.

1.2 Background: Noninvasive neuroimaging

Noninvasive neuroimaging refers to the techniques that can acquire signals from the brain for its structural or functional information, without breaking the skin. It offers a convenient way for neuroscientists to study brain activities in healthy populations. To date, several noninvasive neuroimaging modalities have been popularly used in neuroscience research, which sample signals of different properties to track brain activities. Electroencephalography (EEG), for example, monitors neural activity through sampling the electrical potential at the scalp. These signals originate from the influx and efflux of ions through the membrane of large cortical pyramidal neurons in deep cortical layers [7]. When an action potential happens, ions move in and out of neurons, which causes the electrical potential in the extracellular space to change. If this change is coherent within a local population of neurons (tens to hundreds of millions of neurons), it can be captured by an electrode placed at the scalp. The movement of ions forms an electrical current, which also caused the magnetic field to change around the neurons. This change in magnetic field can be captured by supra-sensitive magnetometers, which constitutes the signals in magnetoencephalography (MEG).

Another approach to trace neural activity is through metabolic dynamics. Blood flow delivers oxygen to different parts of the brain, and it is coupled with neuronal activation — when an area of the brain is recruited, blood blow to that region also increases [8]. And this oxygen level change induced by the blood flow is what functional magnetic resonance imaging (fMRI) is tracing in its blood oxygen level dependent (BOLD) signal. fMRI data can be acquired from a MRI scanner, which usually has a main magnet to polarize the nuclear spin of molecules and a gradient system for signal localization. It scans one slice of the brain at a time and provides millimeter-level spatial resolution. Another imaging modality, functional near-infrared spectroscopy (fNIRS), also relies on the oxygen level in cortical blood flow to monitor brain activities. It sends near-infrared light into the cortex using transmitters placed at the scalp. Due to photon diffusion, part of this light will change the direction of path and come back to the scalp along a U-shape trajectory. The intensity of this return light, captured by detectors at the scalp, is modulated by the oxygenation of blood: the oxygenated and deoxygenated hemoglobin absorb light at different levels for different wavelengths. Therefore, by tracking the intensity change of the return light, we can estimate the oxygen level change underneath the scalp, which is coupled with neuron activation.

A major distinction among these noninvasive imaging modalities (in addition to their cost, mobility and availability) is the spatial and temporal resolution in their signals. Figure 1.1 shows where these modalities are in a temporal-spatial resolution map. The time course of EEG and MEG carries phase-locked evoked responses and can be sampled at a supra-kilohertz rate. There is also oscillation in EEG and MEG that contains information about time-locked induced responses, whose temporal resolution is approximately 10^{-2} to 10^{-1} seconds. The spatial resolution of these two modalities, however, is considerably low — usually in a few centimeters. The opposite is fMRI, which provides exquisite spatial resolution, but is fairly sluggish in tracing temporal dynamics. fNIRS lies between EEG and fMRI in this spectrum. Therefore, with the current noninvasive neuroimaging techniques, there is always a trade-off between knowing "when" and knowing "where" a neural activity happens.

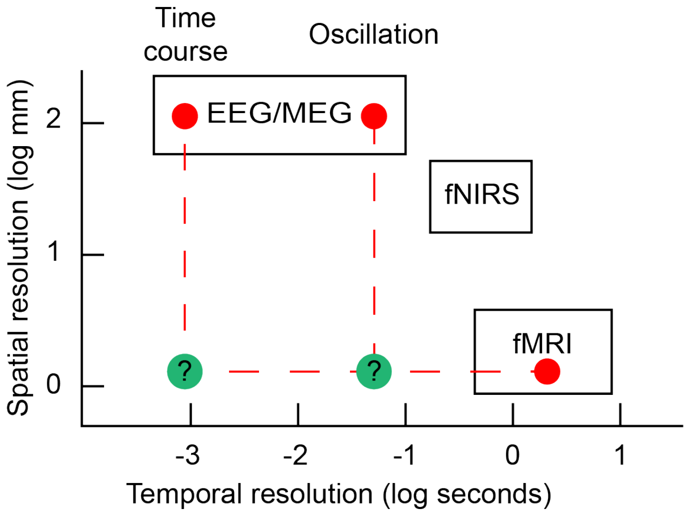


Figure 1.1: Temporal and spatial resolution of common noninvasive neuroimaging modalities. Electroencephalography (EEG) and magnetoencephalography (MEG) have the finest temporal resolution, but the poorest spatial resolution. Functional magnetic resonance imaging (fMRI) has the opposite resolution characteristics. Functional near-infrared spectroscopy (fNIRS) has medium resolution in both spatial and temporal domains. Red dots represent the resolution of EEG and fMRI techniques adopted in this study. The green dots are the ideal resolution we can possibly achieve through EEG-fMRI fusion, if we can find perfect correspondence between information we learned from these two modalities.

One possible workaround to break this resolution barrier is through multimodal neuroimaging, in which more than one imaging techniques are used. For example, we can learn information with fine temporal resolution from EEG and information with fine spatial resolution from fMRI, and find some correspondence between these two through a fusion analysis. In this way, the two modalities will compensate for each other's weakness and reveal a full picture of neural activity in terms of both when and where things happen. In this study, I demonstrate how we can fuse EEG with fMRI through a representational similarity analysis (to be discussed next), and apply this method to explore human auditory selective attention.

1.3 Background: Representational similarity analysis

Classical EEG and fMRI studies, such as analysis of event-related potentials (ERPs) and general linear model (GLM) analysis, directly compare EEG or fMRI features between conditions to make inferences about the neural processes underpinning a cognitive function. Another important cognitive construct, the neural representation, however, is less studied and understood [9]. Neural representation refers to the description of the information being encoded in a neural unit, which is typically a neuron or a brain area [10], or in some other cases, the brain at a certain time point. Kriegeskorte et al. [11] proposed a framework named representational similarity analysis (RSA) to study neural representations from neuroimaging data. Under this framework, representation can be studied through the differences among all pairs of experimental conditions — we can extract some features from each condition and calculate the difference between conditions using a distance metric. These representational dissimilarities, summarized in a matrix called representational dissimilarity matrix (RDM), can then be compared with conceptual models or with representations derived from other imaging modalities, revealing different properties of a cognitive function (e.g., its control model, dynamics, behavioral correlate, etc.). The RSA approach has been used in EEG studies to trace the dynamics of various cognitive functions, including audiovisual integration [12], visual object processing [13] and memory [14], and in fMRI studies to investigate the semantic processing of words [15], language comprehension [16], and computational models of reading [17].

One fascinating property of RSA is that it utilizes the RDMs to study what information is being encoded, which is modality free — given the same condition-rich experimental design, we can calculate RDMs from EEG, fMRI, the observed behavioral performance, or any other data we collected from these conditions. As long as the differences between each pair of conditions are quantifiable, we can derive RDMs that have the same dimension and scale for all the modalities. This is immensely useful for multimodal neuroimaging fusion, because signals like EEG and fMRI have different dynamics, scales, noise levels, etc. Thus, it is difficult, if not impossible, to find a direct correspondence between the EEG signal at one time point and the fMRI signal at one brain location. With RSA, however, we can collect EEG and fMRI data with the same experimental design, calculate RDMs from EEG at each time point and from fMRI at each voxel, and conduct a time-by-location correlation analysis to search for when and where EEG and fMRI share common information. This technique has already been applied to study the cognitive functions underlying visual object recognition. Cichy et al. [18] designed an experiment in which they presented participants with images of objects. They extracted neural representations at each time point from magnetoencephalography (MEG), and at each voxel from functional magnetic resonance imaging (fMRI). Then, they searched for time instances and voxels that shared similar patterns in their neural representation, which depicted a spatiotemporally resolved information flow during visual object recognition.

One common character of these past works (using unimodal or multimodal imaging) is the use of different stimuli across conditions — they focused more on brain's response to different categories of sensory inputs than on its internal cognitive state. This dissertation is among the very few pioneering studies that adopted a RSA framework to investigate the neural representation of attentional states. In later chapters, I will present how I conduct the study, what I learn from it, and what are the potentials and challenges of this approach.

1.4 Background: Brain-computer interface

Brain-computer interface (BCI) systems offer a non-verbal and covert way for humans to communicate a control signal to a computer. They are designed to monitor the brain signals, extract features, and output a decision based on a pre-trained classification model. During the past two decades, we have witnessed rapid growth in BCI research, thanks to the development of advanced sensor technology, machine learning algorithms, and new experimental paradigms [19]. BCI, as an assistive technology, has proved its efficacy in various applications such as communication [20], [21], movement control [22], [23] and rehabilitation [24], [25].

Among the noninvasive neuroimaging modalities that are currently available, electroencephalography (EEG) has become the most popular choice for BCI applications due to its noninvasiveness, mobility, and low cost [26]. EEG monitors brain activity through sampling the electrical potential along the scalp at a very high rate. The high temporal resolution of EEG oscillations allows capturing certain neural signatures of a brain state or of mental efforts, which can be used to decode users' intention. Previous studies have demonstrated great success in building EEG-based BCI systems using visual or auditory stimuli. Chen et al. [27] designed a high-throughput visual BCI system using flickering objects. When the user focuses on one of them, a neural signature known as the steady-state visual evoked potential (SSVEP) appears in EEG signals. However, SSVEP requires a stable line of sight, which may not be available due to permanent or situational impairment (e.g., while driving). As an alternative solution, researchers applied a similar idea to designing auditory BCI systems, where the users were presented with multiple streams of pure tones modulated at different frequencies. The modulation frequency of the attended stream may result in a strong EEG component known as the auditory steady-state response (ASSR) [28].

One major disadvantage of SSVEP or ASSR paradigms is the use of flickering objects or modulated pure tones, which can cause fatigue in users. The sinusoidal carriers used in an ASSR paradigm were perceived by users to be annoying [29]. Past studies have endeavored to use more naturalistic and pleasant stimuli to improve the user-friendliness of BCI systems. For example, Huang et al. [30] used drip-drop sounds in their BCI design, creating a relaxing auditory scene for the users. They decoded attention based on a neural marker called P300, a large positive deflection in an event-related potential (ERP). It is a typical neural signature observed in an oddball paradigm — the subjects are presented with a sequence of identical "standard" stimuli, where, occasionally, the standard stimulus is replace by a target "oddball" [31]. The P300 wave usually occurs around 300 - 500 ms after the oddball onset. In their study, Huang et al. extracted P300 features from EEG signals and trained a linear discriminant analysis (LDA) model to determine the user's attentional state. They achieved a 73.5% decoding accuracy for binary classification, and a 2.75 bit/min equivalent information transfer rate (ITR, the amount of information transferred per unit time, which is a standard method for measuring the performance of a BCI system). In another study, Treder et al. [32] embedded oddballs in streams of real music and used the P300 feature for attention decoding. They achieved a 91% decoding accuracy, an amazing result for a binary classification problem. However, since their system averaged 40 seconds of data to generate one output, its overall efficiency of information transfer is merely 0.8 bits/min.

In summary, attention-based auditory BCI is a concept with great potential. It does not require visual attention for use, a cognitive resource that is heavily demanded in our daily life and is not always available. However, it still has a few major obstacles before it can become practical for everyday use. First, we need to improve its user-friendliness to make it unobtrusive and pleasant to use. This concerns many aspects of a BCI design, such as the stimuli we play, the form factor we use to record EEG signals, and the overhead time to setup the system. Second, the throughput of most current auditory BCI designs is not as high as that of a visual BCI. Previous visual BCIs built on SSVEP or P300 could easily achieve an ITR of 20 - 30 bits/min [27], [33], which is at least 10 times greater than most auditory BCI designs we have today. This is a combined effect of relatively lower decoding accuracy and longer processing time in auditory BCIs compared to their visual counterparts.

Lastly, an auditory BCI system should be designed to offer more practical value in daily life. For example, visual BCIs have been developed to assist people with typing [33] or spatial navigation [27]. These systems can be used naturally and intuitively by users, because they are coherent with how these functions are usually executed in daily life — we look when we type or navigate. However, it would be awkward if we intend to replace these functions with an auditory BCI system. New ideas on what benefits an auditory BCI can offer are therefore warranted.

Chapter 2

Neural representation of auditory attention in EEG[1]

2.1 Introduction

Auditory selective attention enables us to focus on one sound within a mixture of noises. Unlike in vision, where we can easily shift attention by moving the eyes, auditory attention can only be achieved covertly via cognitive control. That is, we change our internal attentional state in the brain in order to attend to or ignore a sound object [1]. This control is effortless and swift for healthy people, but can be challenging for people with neurological conditions such as attention deficit/hyperactivity disorder [2], [3] or autism spectrum disorder [4], [5]. This makes it important to study the neural mechanisms that underlie auditory selective attention.

Electroencephalography (EEG), for its simple setup, low cost, and high sampling rate has become a popular tool to study the dynamics of cognitive functions in human brain. It monitors neural activity through sampling the electrical potential at the scalp at a very high rate. It carries rich information in the spectral and temporal domain, and can show fast

[1] This chapter is adapted with permission from a manuscript: Winko W. An, Alexander Pei, Abigail Noyce, Barbara Shinn-Cunningham, "Neural representation of auditory attention in EEG", in preparation

dynamics of brain response to a stimulus. It has been used to capture modulation of neural signatures, such as event-related potentials (ERPs) and induced oscillations, as evidence of top-down attentional control [34], [35]. Choi et al. [36] created an auditory attention task with three competing melodies and asked the listeners to identify the pitch contour of the attended stream. They discovered that the magnitude of N1, which is an early component (100 - 150 ms after stimulus onset) in an ERP waveform, is modulated by attentional control. In another study, Giuliano et al. [37] demonstrated that the magnitude of the P1 component (50 - 100 ms after stimulus onset) is also modulated by attention. Deng et al. [38] designed an experiment using spatialized streams of syllables, and directed the participant's attention to different locations. They found that the alpha oscillation, which is an oscillatory induced response whose frequency ranges from 8 to 14 Hz, is modulated by auditory spatial attention. They observed an alpha power increase in the parieto-occipital region, ipsilateral to the side to which the participant attended. And this parietal alpha wave was also proved to have a causal relationship with suppression of the representation of contralateral auditory space [39].

Classical EEG studies, such as the aforementioned ERP analysis and time-frequency analysis, directly compare EEG features (e.g., ERP or band power) between conditions to make inferences about the neural processes underpinning a cognitive function. Another important cognitive construct, the neural representation, however, is less studied and understood [9]. Neural representation refers to the description of the information being encoded in a neural unit, which is typically a neuron or a brain area [10], or in some other cases, the brain at a certain time point. Kriegeskorte et al. [11] proposed a framework named representational similarity analysis (RSA) to study neural representations from neuroimaging data. Under this framework, representation can be studied through the differences among all pairs of experimental conditions. These representational dissimilarity features can then be used to compare with conceptual models or dissimilarity features derived from other modalities to reveal different properties of a cognitive function (e.g., its control model, dynamics, behavioral correlate, etc.). The RSA approach has been widely used to study the dynamics of cognitive function underlying visual object recognition. For example, Cichy et al. [18] designed an experiment in which they presented participants with images of objects. They extracted neural representations at each time point from magnetoencephalography (MEG), and at each voxel from functional magnetic resonance imaging (fMRI). Then, they searched for time instances and voxels that shared similar patterns in their neural representation, which depicted a spatiotemporally resolved information flow during visual object recognition. Besides, RSA has also been applied to EEG studies to trace the dynamics of various cognitive functions, including audiovisual integration [12], visual object processing [13] and memory [14].

One common character of these past works is the use of different stimuli across conditions — they focused more on brain's response to different categories of sensory inputs than on its internal cognitive state. In this study, we deploy the idea of RSA to investigate the neural representation of attentional states. We designed an experiment in which input auditory signals were identical, but varied whether listeners engaged spatial or non-spatial auditory attention. We used features in EEG time course and its alpha oscillations for neural decoding to understand the degree to which the listener's internal attentional states differed across conditions. We then compared these neural representations to several conceptual models or behavioral performance to reveal the kind of information being encoded at each time point during the experiment.

2.2 Materials and methods

All study procedures were approved by the Institution Review Board of Boston University.

2.2.1 Participants

Thirty adults (19 – 44 years old, 14 women) participated in this study. No participant reported hearing loss or any history of neurological disorders. All participants gave written informed consent, and were paid for their time.

2.2.2 Stimuli and task

Participants used either spatial or non-spatial attention to listen for the identity of a target syllable (/ba/, /da/, or /ga/) among three distractor syllables differing from the target in both time and spatial position. Raw stimuli were recordings of these syllables spoken by four native English speakers (two men, two women). The syllables were spatialized, via a generic head-related transfer function [40], to five simulated locations in azimuth: 90° from the left (L90), 30° from the left (L30), center, 30° from the right (R30), or 90° from the right (R90). (Figure 2.1a).

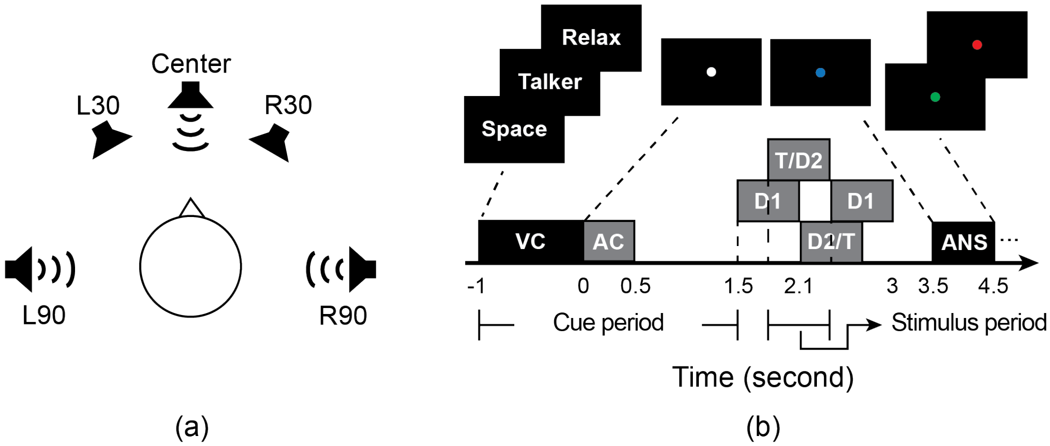


Figure 2.1: Experimental task. (a) Spoken syllables were spatialized to center, 30° left (L30) or right (R30), and 90° left (L90) or right (R90), always in the horizontal plane. This figure is showing one possible scenario where sounds come from L90, R90 and center. (b) Illustration of the events within a trial. A visual cue (VC) showed the type of attention required for this trial, followed by an auditory cue (AC) specifying the desired value of attention (a specific direction or talker). A 4-syllable mixture was played one second after the AC. The first and the last syllables were always distractors (D1). Of the second and the third syllables, one was the target (T), and the other one was the second distractor (D2). All syllables were 600ms long, and their onsets were 300ms apart. The participants were instructed to respond when the fixation dot turned blue. Feedback was given at the end of each trial via a green (correct) or red (incorrect) fixation dot.

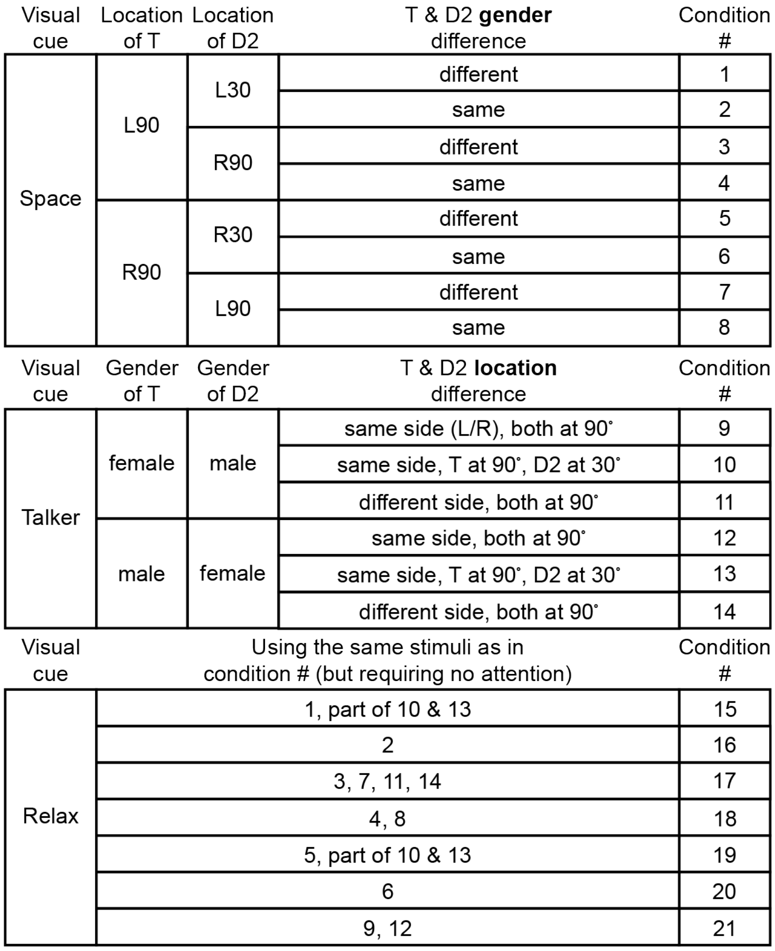
At the beginning of each trial, a visual cue (VC) was presented for 1 s, which instructed the participant to prepare for one of three types of trials (Figure 2.1b). "Space" indicated that participants should use spatial attention to report the syllable at a particular location. "Talker" indicated that participants should report the syllable spoken by a particular talker. "Relax" represented a no-attention control trial. The VC was immediately followed by an auditory cue (AC) that conveyed the relevant feature of the target. The AC was always an /a/ sound. In "Space" trials, it was spoken by a synthesized gender-neutral voice and spatialized to either L90 or R90, specifying the target location. In "Talker" attention trials, the AC was spatialized to the neutral center location and spoken by one of the four talkers, specifying the target talker. In "Relax" trials, the AC was spatialized to center and spoken by the synthesized neutral voice.

After a one-second silent preparatory period, a 4-syllable mixture was played. All syllables were 600 ms long, and their onsets were separated by 300 ms. The first and last syllables were always distractors (D1) spatialized to center and spoken by a gender-neutral talker. Of the second and third syllables, one was the target and the other was the second distractor (D2). Syllables /ba/, /da/, and /ga/ were randomly permuted among target, D1 and D2. Subjects were instructed to identify and report the identity of the target syllable via keypress ("1" for /ba/, "2" for /da/, and "3" for /ga/). Feedback, a green signal for a correct answer or a red signal for an incorrect answer, was given after each response.

In "Relax" trials, participants were were asked to ignore all syllables, and give a random answer at the end. The inter-trial interval was 2 seconds with jitters (0 – 0.5 second).

2.2.3 Experiment Design and Procedures

The full experiment consisted of 21 conditions (Figure 2.2). These conditions are distinguished by (1) the kind of attention required, as indicated by the VC, and (2) by the characteristics of the target and of D2. The stimuli used in "Space" and "Talker" conditions were also used in "Relax" conditions, to control for stimulus-driven effects. Participants completed 36 trials of each condition for a total of 756 trials. The order of trials was shuffled and evenly divided into 12 blocks with short breaks between for participant comfort.



Visual cue	Location of T	Location of D2	T & D2 **gender** difference	Condition #
Space	L90	L30	different	1
			same	2
		R90	different	3
			same	4
	R90	R30	different	5
			same	6
		L90	different	7
			same	8

Visual cue	Gender of T	Gender of D2	T & D2 **location** difference	Condition #
Talker	female	male	same side (L/R), both at 90°	9
			same side, T at 90°, D2 at 30°	10
			different side, both at 90°	11
	male	female	same side, both at 90°	12
			same side, T at 90°, D2 at 30°	13
			different side, both at 90°	14

Visual cue	Using the same stimuli as in condition # (but requiring no attention)	Condition #
Relax	1, part of 10 & 13	15
	2	16
	3, 7, 11, 14	17
	4, 8	18
	5, part of 10 & 13	19
	6	20
	9, 12	21

Figure 2.2: The 21 experimental conditions in this study. These conditions differ by their VC and the location or gender of talkers for T and D2. The "Relax" conditions used the same stimuli as in "Space" and "Talker", but required no attention.

After collecting informed consent, subjects started with practicing the attention tasks on a laptop. They performed a 21-trial test run of the experiment (i.e., one trial for each of the 21 conditions) repeatedly until their response accuracy reached 75%. This ensured that they fully understood the instructions and could correctly deploy Spatial and Talker attention as required during the actual experiment. After training, subjects were asked to sit in a sound-treated booth in front of a LCD computer monitor. All experimental audio was presented through a pair of insert earphones (ER1, Etymotic Research).

2.2.4 EEG recording and preprocessing

EEG was continuously recorded from 64 electrodes arranged according to the international 10/20 system. Signals were digitized at 2048 Hz using the ActiveTwo system (BioSemi B.V.) The data were first passed through a sinc windowed FIR band-pass filter (0.1 - 50 Hz) to remove slow drifts and line noise. This also served as an anti-aliasing filter. The signals were then downsampled to 256 Hz. An independent component analysis was conducted using EEGLAB [41] to remove eye blinks and muscle artifacts. The continuous EEG signals were then split into epochs. We isolated a preparatory attention period from -1000 – 1500 ms relative to the onset of the auditory cue, and a peristimulus attention period from -300 – 500 ms relative to the onset of the target syllable.

2.2.5 Measuring representational dissimilarity

Auditory attention was decoded from EEG signals to allow us study the neural representation of attentional states. Similar to previous studies [18], [42], [43], decoding was based on analyzing signals in the time domain, i.e., from the EEG time course. Additionally, since alpha band oscillations (8 – 14 Hz), especially those in parietal cortex, carry rich information about attentional state, we also decoded auditory attention from the time course of alpha power. Figure 2.3 summarizes the steps in extracting neural representations for attention decoding. Two features — the EEG time course and the instantaneous power of alpha band oscillations (8 – 14 Hz) — were studied independently. Oscillatory power was calculated from a continuous wavelet transform (CWT), which was averaged within the alpha frequency range to estimate band power [44]. For each subject, at each time instance t_0, we concatenated the feature of interest from 64 EEG channels to form a feature vector.

Representational dissimilarities were measured from these vectors via machine learning classification. If two conditions are dissimilar (e.g. a spatial attention condition compared with a no-attention condition), the classifier may be able to identify distinguishing information from their EEG features, and thus yield an above-chance classification accuracy (i.e.,

>50%). At each time instance, for each pair of conditions, we trained a linear support vector machine (SVM) using leave-one-trial-out cross-validation. On each of 100 iterations, one trial from each of the two conditions under consideration was selected to be left out; all others were used to train the SVM. The resulting model was applied to the held-out trials, and its classification accuracy was recorded. The average classification accuracy across iterations served to estimate the difference between conditions. This process yields a representational dissimilarity matrix (RDM) for each time point, where each entry reflects the dissimilarity of the neural responses across a particular pair of conditions.

Specifically, each row and column of an RDM corresponds to a condition index, and each element stores the dissimilarity value between two conditions: the classification accuracy when discriminating between condition i and j is saved at (i , j) and (j , i) of the RDM. With this approach, the major diagonal of an RDM, representing the difference between one condition and itself, is filled with zeros and is excluded from any analysis. Two sets of RDMs for each subject were generated from this attention decoding process, one for the EEG time course and one for the alpha-band power. Because there is one RDM at each time instance, each set is a function of time, summarizing how the differences in neural responses to each pair of conditions evolves over the course of a trial. These time-varying RDMs can be used to study how neural representation of attentional states changes in the course of a trial.

It needs to be stressed that this neural decoding analysis was conducted for each subject independently. This is because, there is abundant inter-subject variability, such as differences in their neural anatomy and electrode placement, that may hinder classification using data from all subjects. For example, a neural signature observed at channel CPz in one subject may appear at channel POz in another. As close as these two channels are physically, the classifier may only recognize two different effects in two separate dimensions, and thus will not enhance its learning. Training subject-specific classifiers can ensure that the topographic pattern of the same neural signature is consistent within all the samples, so that the distinguishing features between classes can be easily captured. These subject-specific RDMs can be studied through correlation with some ideal conceptual model RDMs (to be discussed in Section 2.2.6) to find patterns embedded in these EEG RDMs, or through correlation with subject-specific behavioral RDMs (to be discussed in Section 2.2.7) to link EEG with behavioral performances.

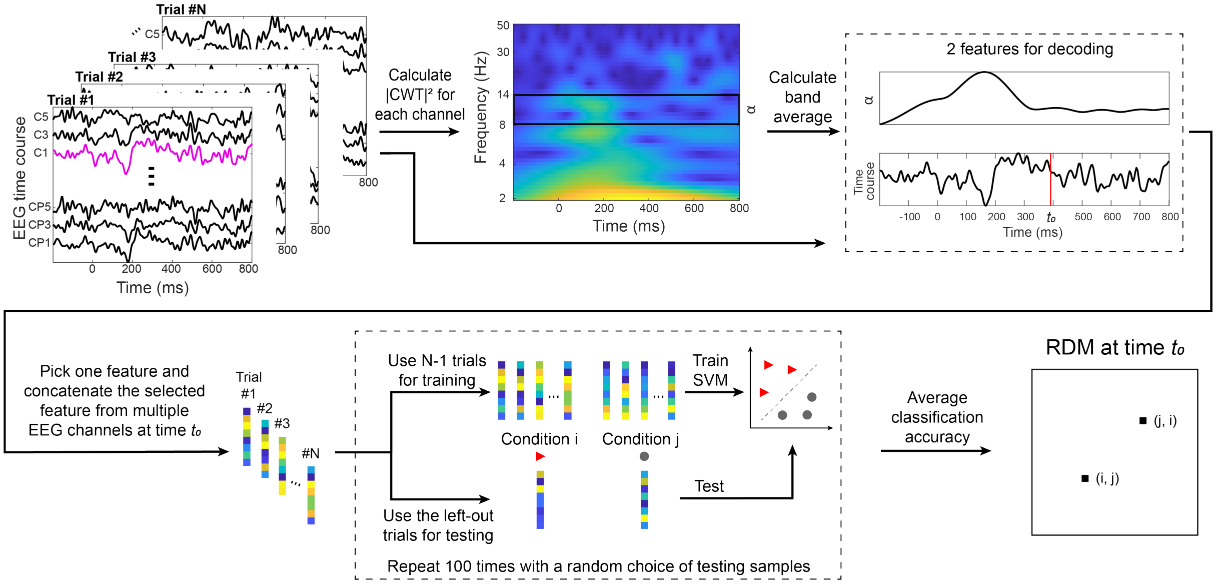
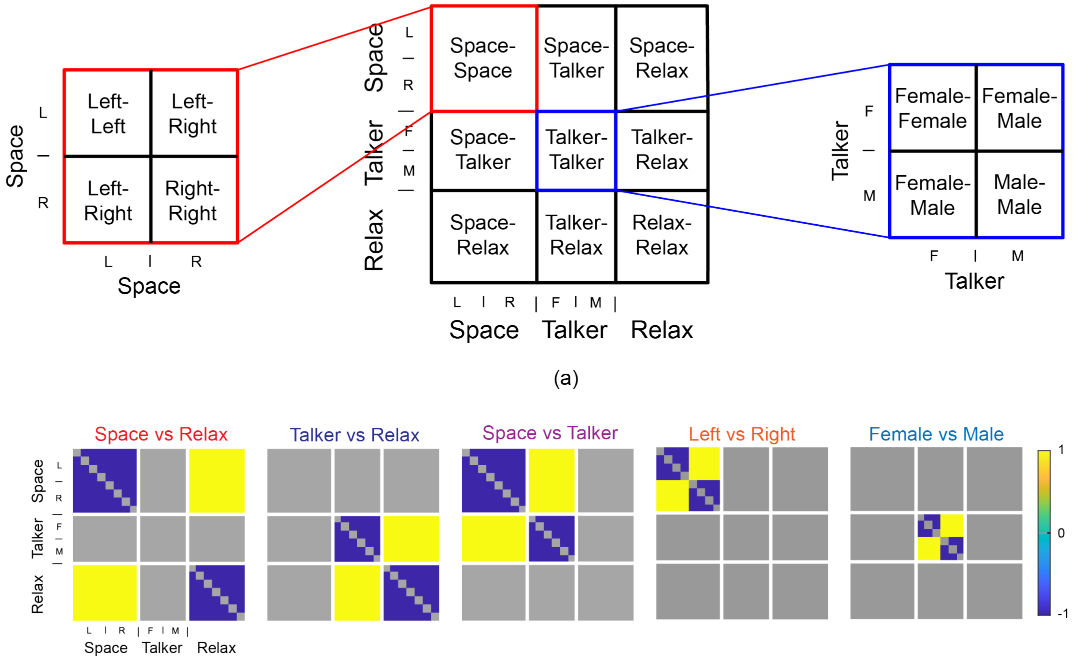

Trial #N
Trial #3
Trial #2
Trial #1
C5
C3
C1
CP5
CP3
CP1
EEG time course
0 200 400 600 800
Time (ms)
Calculate |CWT|² for each channel
Frequency (Hz)
50 30 14 8 4 2
0 200 400 600 800
Time (ms)
α
Calculate band average
2 features for decoding
α
Time course
-100 0 100 200 300 t₀ 500 600 700 800
Time (ms)
Pick one feature and concatenate the selected feature from multiple EEG channels at time t₀
Trial #1
#2
#3
#N
Use N-1 trials for training
Condition i
Condition j
Train SVM
Test
Use the left-out trials for testing
Repeat 100 times with a random choice of testing samples
Average classification accuracy
RDM at time t₀
(j, i)
(i, j)

Figure 2.3: Schematic diagram for how a representational dissimilarity matrix (RDM) was derived. Either the EEG time course or power of a specific frequency band (calculated from a continuous wavelet transform, or CWT) was used as the feature to decode attention at a particular time point. Dissimilarity between each pair of conditions was estimated through 100 iterations of training-and-test a support vector machine (SVM) with a random choice of test samples in each iteration. The average classification accuracy was saved in an RDM at its corresponding positions. The output of this attention decoding procedure is a series of RDMs as a function of time for each feature of choice (the EEG time course or power of a frequency band).

2.2.6 Comparison to conceptual models

The RDM feature summarizes how one experimental condition differs from another. In our study, since the conditions could be categorized into certain condition groups (e.g., Space, Talker, etc.), this RDM feature also contained information about how a specific aspect of attention differed from another. For example, if we are interested in the difference between two attention types, or the difference between conditions within the same attention type, we can refer to certain portions of an RDM for this information — the square cells along the major diagonal of an RDM (Space-Space, Talker-Talker and Relax-Relax in Figure 2.4a) represent how one condition differs from another within the same attention type, while the cells off the diagonal (Space-Talker, Space-Relax and Talker-Relax) show the difference between attention types. Moreover, there were also micro-structures within Space-Space and Talker-Talker. Four conditions in Space were about left attention, and the other four were about right attention. They formed micro-cells that showed differences within attention to the same side (Left-Left and Right-Right), and cells for attention to different sides (Left-Right, Figure 2.4a). Similarly, the six talker-attention conditions were separated between attention to a female talker and attention to a male talker. Thus, the Talker-Talker cell could be further divided into Female-Female and Male-Male for within-gender comparisons, and Female-Male for between-gender comparisons.

We created several conceptual models that capture specific patterns of discrimination among conditions. Each was an ideal RDM model consisting entirely of 0s and ±1s: positive for conditions that are very different in the dimension of interest, negative for conditions that are very similar, and zero for conditions that are irrelevant. We then compared each subject's EEG RDMs over time to the ideal model RDMs by computing the correlation between the computed RDM at each time point and the ideal model. The conceptual models are shown in Figure 2.4b. For example, the "Space vs. Relax" model has low dissimilarity within Space and Relax conditions (the Space-Space cell in the top left and the Relax-Relax cell in the bottom right portions of the RDM), and high dissimilarity between these conditions (the Space-Relax cell in the bottom left and top right portions of the RDM). All cells of the model that include a Talker condition are set to zero, effectively excluding them from the analysis, since these conditions give no information about the difference in brain activity patterns for Space and Relax conditions. Similar conceptual models were constructed for "Talker vs. Relax", "Space vs. Talker", "Left vs. Right", and "Female vs. Male" (Figure 2.4b).



(a)

(b)

Figure 2.4: (a) A representational dissimilarity matrix (RDM) in this study can be divided into several cells, and each cell represents different information. The top left portion of the RDM, Space-Space (the red cell), shows the difference between conditions within the spatial attention group. Similarly, the other two square cells along the major diagonal, Talker-Talker and Relax-Relax, also show differences within their respective condition groups. The cells off the diagonal (Space-Talker, Space-Relax and Talker-Relax), on the other hand, store dissimilarity values between conditions groups. In addition, within Space-Space and Talker-Talker, there are also micro-structures showing comparisons between attention to different directions (left or right), or between attention to talker of different genders (female or male). (b) Five conceptual model RDMs used to study patterns in EEG RDMs. In Space vs Relax, Talker vs Relax and Space vs Talker, cells representing comparisons between attention types are given values of +1s, and cells for comparisons within each attention type are given values of −1s. These models can be used to study the representation of attention types. In Left vs Right, the Left-Right cells are filled with +1s, and the Left-Left and Right-Right cells are filled with −1s — this model RDM is built to study the representation of "where to attend". In Female vs Male, we assigned +1s to the Female-Male cells, and −1s to the Female-Female and Male-Male cells – this model is built to study the representation of "what to attend".

For each EEG feature and experimental interval (cue and stimulus period), we tested the correlation between the conceptual model RDMs and the EEG RDMs derived from our data. A high correlation at a time point indicates that the EEG RDM at that time is similar to the model being tested, which implies that the brain may have elicited signals that can distinguish different attentional states. The time series of correlation coefficients from each subject were used as the input for a cluster-based permutation test (CBPT) for statistical inferences.

CBPT (for details, see [45]) is a non-parametric statistical test that aims at finding significant effects that are continuous in feature space (e.g., in time, space, frequency, etc.). In this study of correlation with conceptual models, we setup the CBPT to identify continuous time intervals in which the average correlation across all subjects were significantly above zero. Specifically, we first calculated the t-statistics at each time point by comparing all subjects' correlation coefficients at each time point with zero using a one-sample t-test. We set the cluster formation alpha to 0.05, which led to a series of above-threshold t-values. These t-values formed clusters that were continuous in time, and we used the sum of t-values within each cluster as the cluster-level test statistics. We then derived a null distribution through randomly permuting data labels. For a one-sample t-test, this permutation process could be effectively achieved by negating the sign of data for a randomly selected subset of all subjects. With the permutation dataset, we again conducted a one-sample t-test at each time point, and calculated the sum of t-values for each found cluster. Only the maximum cluster-level test statistics from each permutation was used to form the null distribution. We repeated this permutation process for 10,000 time, which yielded a null distribution with 10,000 cluster-level test statistics. Then, we compared the observed cluster-level test statistics with the null distribution, and assigned a p-value to each cluster — this p-value equaled the proportion of the null distribution that were greater than the observed test statistics of the cluster. For example, if the cluster-level test statistics of one cluster was 300, and only 10 out of 10,000 permutations yielded a cluster-level test statistics over 300, this cluster was assigned with a p-value of 0.001. Any cluster with a p-value less than 0.05 was deemed as significantly above zero in this study.

These correlation analyses identify time intervals in which certain information (e.g., type of attention or direction of attention) can be decoded from EEG signals. However, it does not reveal what aspects of the signal are driving the decoding, such as which channel contains information that distinguishes one condition from another, or which condition has greater value in that channel during this period of time. To answer these questions, we compared the event-related potential (ERP) or band powers between conditions of interest. For ERP, we used bootstrapping to estimate condition means in each channel for comparison [46]. For band power, we calculated the average of conditions within time intervals in which we observed a high correlation. For example, if the alpha power EEG RDM and the Space vs Relax model RDM are strongly correlated during a period of time, we took the average alpha power of all spatial attention conditions within this period, and compared it with the average alpha power of all no-attention conditions within the same period. This computation complements the study of neural representation, which only discovers the existence of differences, by identifying information about what and where these differences are, which could help us better understand the mechanisms of attention.

2.2.7 Comparison to behavioral performance

Participants' behavioral performance varies across task conditions. The RSA approach allows us to ask whether brain activity during any time periods is particularly relevant to the subject's behavioral performance. We constructed behavioral RDMs that comprise the absolute difference in behavioral performance between conditions. Only active attention conditions (8 spatial and 6 talker conditions) were used in this analysis. In addition, we excluded subjects with more than two perfect scores (i.e., 100% accuracy), a result suggesting that the task was not challenging enough for a particular participant (thus hindering the manifestation of behavioral correlates). 19 out of 30 participants survived this exclusion criteria. Their average behavioral RDM is shown in Figure 2.11a. We looked for time intervals in the stimulus period where each individual's behavioral RDM correlates with their own EEG RDMs better than chance, which is indicative of a strong correspondence between EEG signals and the observed behavioral performance.

As with the analysis of correlation with conceptual models, correlation with the behavioral RDM does not indicate any individual channel's contribution to behavior. Therefore, we conducted a channel-wise behavioral correlation analysis — we calculated the average scalp voltage or alpha power for each condition at each channel across the period where a subject's EEG RDM correlates significantly with their behavioral RDM. For each subject individually, these channel-wise averages were first z-scored across all conditions, and then were correlated with that subject's corresponding z-scored behavioral performance. This process yielded a set of channel-wise correlation coefficients, which reveal how EEG measures at each channel covaried with behavioral performance.

2.3 Results

In this work, we studied the neural representation of auditory selective attention in EEG signals. We designed an experiment with multiple conditions that used similar stimuli, but required the listeners to adopt different listening strategies and different cognitive states. We recorded EEG from 30 subjects when they were performing the attention task and extracted neural representation features from their EEG time course or alpha oscillation power. These representation features are essentially the differences between each pair of conditions, which were estimated via a machine learning classification approach — EEG features, either the EEG time course or its alpha oscillation power, were used to train and test a linear SVM. The average classification accuracy of each binary SVM, estimated by cross-validation, was used as the dissimilarity value between the corresponding conditions. These pair-wise differences were summarized in a representational dissimilarity matrix (RDM) — each row and column of an RDM represent a condition index, and the value stored at element (i,j) and (j,i) of an RDM represents the estimated difference between condition i and condition j. This process yielded a set of RDMs as a function of time;each time point has one specific RDM.

2.3.1 Representational dissimilarity matrices

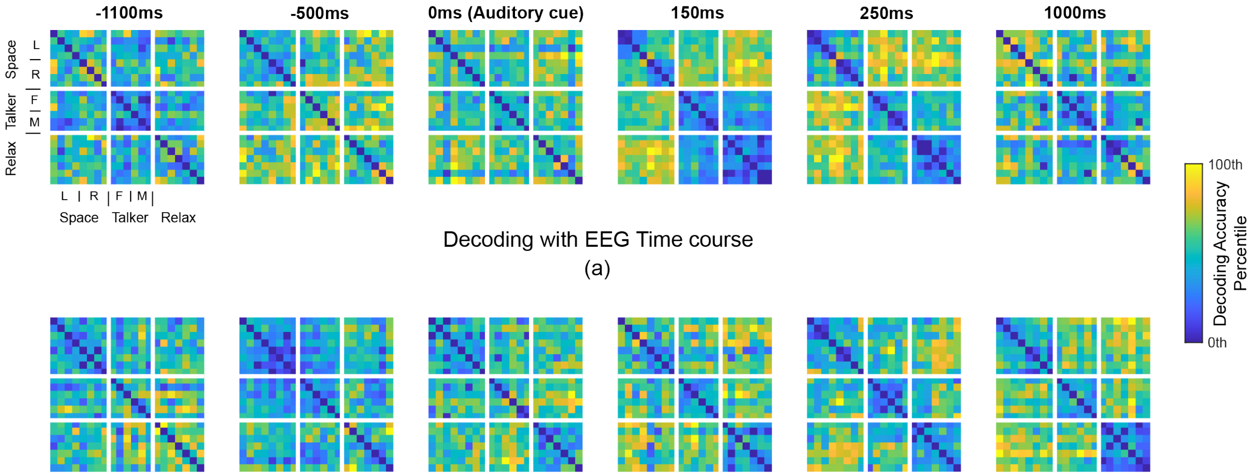
2.3.1.1 Cue period

Figure 2.5a shows several examples of EEG RDMs yielded by decoding attention from the EEG time course during the cue period. Each row and column in an RDM represent a condition index, and these conditions are arranged in the same way as in Figure 2.2.6 (i.e., the first 8 conditions are Space, followed by 6 conditions of Talker and 7 conditions of Relax). The RDMs did not show any particular pattern until approximately 150 ms after an AC was played. Spatial attention then separated from talker and no-attention in feature space, as indicated by the high decoding accuracies in the bottom left cell and the low accuracies in the bottom right cell at and after this time point. Moreover, at 150 ms, subjects' attentional state in conditions with a left cue could be decoded from conditions with a right cue, indicated by the micro-structure within the Space-Space cell in the top left corner. At 250 ms after AC onset, differences between attended to a female versus a male talker emerged in the EEG responses (indicated by the micro-structure within the center square). These differences between conditions in EEG time course were transient — the RDM returned to its random pattern for the rest of cue period (as an example, see the RDM for 1000 ms).

Information in the alpha band persists compared to that in the EEG time course. Differences between active attention conditions and no-attention conditions could be decoded from alpha power as early as when an AC was presented (Figure 2.5b). And such difference lasted through the end of the cue period, suggesting that auditory attention modulated alpha oscillations when subjects cognitively prepared themselves for the upcoming task.

2.3.1.2 Stimulus period

We also examined the RDMs during the stimulus period. The EEG time course RDMs showed discernable patterns only after the target syllable was played (Figure 2.6a). Space and talker attention conditions were distinguishable from no-attention conditions immediately



Decoding with EEG Time course
(a)

Decoding with alpha power
(b)

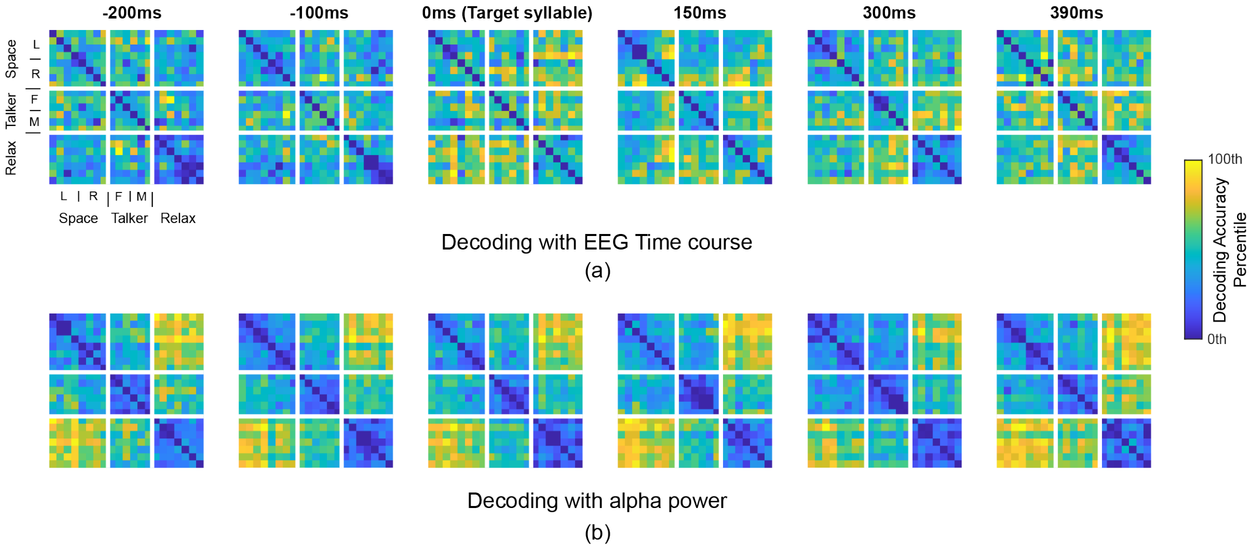
Figure 2.5: Examples of representational dissimilarity matrix in the cue period, derived from (a) EEG time course, and (b) alpha power. 0 ms denotes the onset of auditory cue. Color map was scaled to the 0th and 100th percentile value within each RDM.

(0 ms). In contrast, the alpha power RDMs showed a persistent pattern throughout the stimulus period (Figure 2.6b), even before the onset of the target syllable — spatial and talker attention conditions were consistently different from no-attention conditions in alpha power. In addition, we also observed dissimilarity between left and right attention during this period.

2.3.2 Correlation with conceptual models

2.3.2.1 EEG time course during cue period

We identified two important intervals in the cue period where the EEG time course RDMs correlate with ideal RDM models of attention type (i.e., Space vs Relax, Talker vs Relax and Space vs Talker) better than chance. The first interval occured between the VC and the AC (-750 to -300 ms, Figure 2.7a, top panel) — the average correlation for all three models peaked at around 350 ms after the onset of VC (i.e., -650 ms on the time axis), and had relatively small correlation (ρ <0.1) throughout this period. We examined ERP difference waves at several electrodes in order to understand the brain activity contributing



Decoding with EEG Time course
(a)

Decoding with alpha power
(b)

Figure 2.6: Examples of representational dissimilarity matrix in the stimulus period, derived from (a) EEG time course, and (b) alpha power. 0 ms denotes the onset of target syllable. Color map was scaled to the 0th and 100th percentile value within each RDM.

to these differences. The analysis of EEG time courses revealed that that there was a minor increase in correlation that could be attributed to subtle differences in late ERP components observed at certain EEG channels (e.g., N2 at C4 and P3 at POz, Figure 2.7a). The second interval, however, exhibit a strong, salient peak in correlation for Space vs Relax and Space vs Talker, indicating that difference in EEG time courses between attention types could be well decoded from the EEG time course. Spatial attention in Space conditions evoked stronger N1 and P2 responses at C4 compared to Talker and Relax conditions, as well as a stronger P2 response at POz. We also observed greater N1 responses at C4 and POz in Talker than in Relax. This difference between the neural responses for Talker and Relax conditions contributed to a significant increase in discriminability of these conditions at these times, and an increase in the correlation of the measured responses with the Talker vs Relax ideal model at approximately 130 ms after the AC onset.

The EEG time course RDMs did not correlate significantly with Left vs Right and Female vs Male conceptual models until after an AC was played (Figure 2.7b). Correlation with the Left vs Right model has an early peak (153 ms). In contrast, the Female vs Male model correlated most strongly with the computed RDMs sat a later time point (270 ms). The former peak is aligned with differences in N1 response at channels like C4, while the latter aligns with P2 responses at channels around FCz (Figure 2.7b). Both correlation traces returned to their baseline soon after the offset of the AC (500 ms).

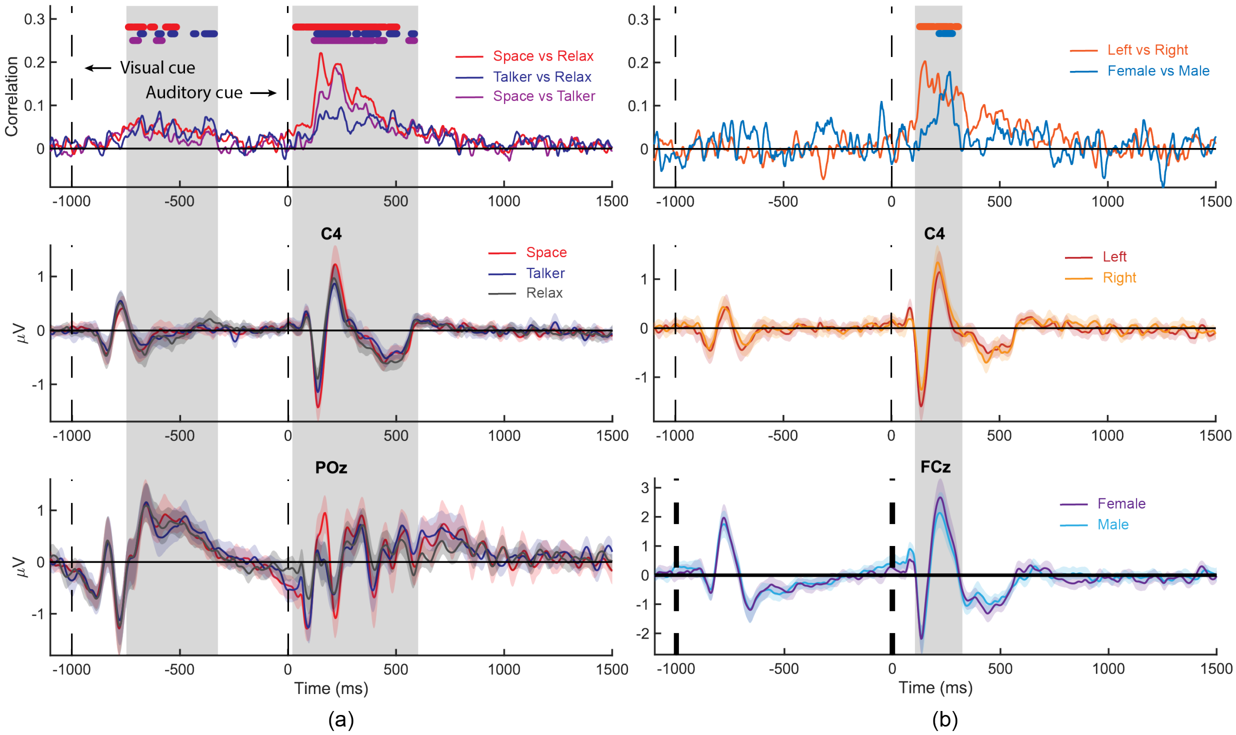


Figure 2.7: Correlation between EEG time course RDMs and conceptual model RDMs in the cue period. (a) Five conceptual model RDMs consisted of 0s and ±1s. Zeros (masked by grey) were not used in calculating correlation. (b) and (c) Top row shows the correlation time course of each conceptual model, averaged across all subjects. The bars above indicate time intervals in which their corresponding correlation trace is significantly above zero. The second and third rows show EEG time courses from selected channels, averaged within each condition group and across all subjects. The shaded region associated with each trace indicates the 95% confidence interval of the respective mean, estimated with bootstrap resampling.

2.3.2.2 Alpha power during cue period

The correlation between the measured alpha power RDMs and the ideal model RDMs in the cue period yielded values consistently above zero (Figure 2.8a). This is notably different from the transient effects observed in the time course analysis described above. After an AC is given, all three attention types could be decoded from each other using alpha power features alone. We compared whole-scalp alpha power differences from 900 ms to 1200 ms and found that, compared to the no-attention condition, Space and Talker focused attention induced slightly, but not significantly, stronger alpha power in central and parietal channels. We also observed significantly higher alpha power in the right parietal region in spatial attention than in talker attention during the same window. Direction of attention or of the AC could also be decoded from alpha power during the cue period. The Left vs Right model correlates with alpha power RDMs above chance in a few discrete windows after the offset of the AC (Figure 2.8b). Between 500 ms and 645 ms, we observed strong alpha power in parietal sensors ipsilateral to the direction of attention or of the AC. This pattern was retained in later windows, but became weaker and less significant over time. No strong correlation was found between the Female vs Male model and alpha power RDMs.

2.3.2.3 EEG time course during stimulus period

In the stimulus period, correlations between the measured EEG RDMs and ideal model RDMs were significantly above zero after the onset of target (Figure 2.9a), indicating that differences between attention types could be decoded from the EEG time course during this time window. ERP analysis revealed some differences between attention types at channels like C4 and POz. However, unlike in the cue period, where we could attribute the observed high correlation to differences in specific ERP components (e.g., N1 or P2, etc.), it is difficult, if not impossible, to do so in the stimulus period; specifically, ERPs of different syllables overlapped with each other in time, as the the syllables were only separated by 300 ms, shorter than the duration of a typical ERP. We also observed significant correlations between the the measured RDMs and the ideal Left vs Right model roughly 150ms after target syllable onset, and for the Female vs Male model at around 380 ms (Figure 2.9b).

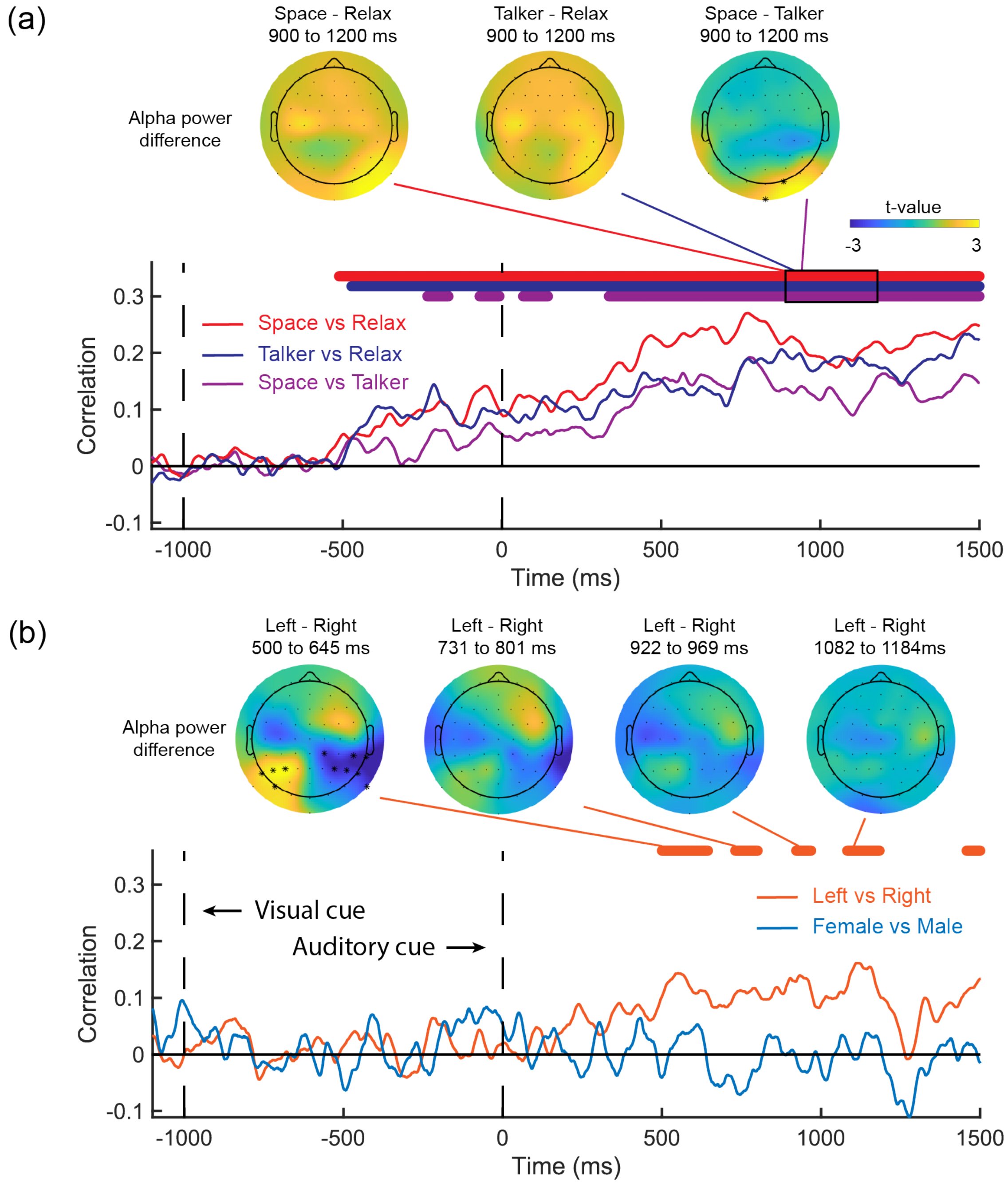


Figure 2.8: Correlation between EEG alpha power RDMs and conceptual model RDMs (shown in Figure 2.7a) in the cue period. (a) and (b) Traces represent correlation time courses, averaged across all subjects. The bars above indicate time intervals in which their corresponding correlation trace is significantly above zero. The topographic maps show alpha power differences between condition groups within certain time intervals, in the form of t values. Asterisks indicate channels where the observed difference is significant.

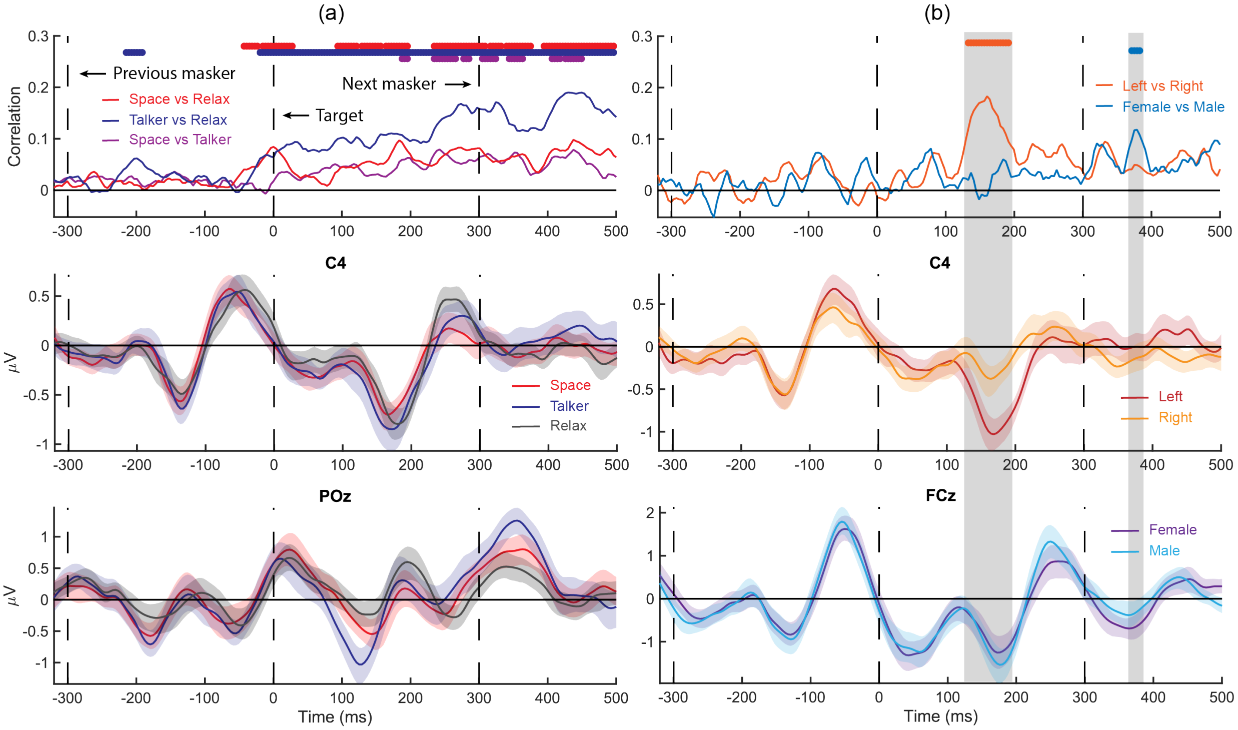


Figure 2.9: Correlation between EEG time course RDMs and conceptual models (shown in Figure 2.7a) in the stimulus period. (a) and (b) Top row shows the correlation time course of each conceptual model, averaged across all subjects. The bars above indicate time intervals in which their corresponding correlation trace is significantly above zero. The second and third rows show EEG time courses from selected channels, averaged within each condition group and across all subjects. The shaded region associated with each trace indicates the 95% confidence interval of the respective mean, estimated with bootstrap resampling.

2.3.2.4 Alpha power during stimulus period

Alpha power in the stimulus period carries abundant information about the internal type of top-down attention a listener engages. The correlation traces for all three attention type models were above chance throughout the whole stimulus period (Figure 2.10a). Space and Talker attention induced substantially stronger alpha power than Relax conditions in frontal and parietal regions from 0 ms to 300 ms after the onset of the target syllable. For Space vs Talker, we observed strong alpha power in left temporal and left parietal regions during the same window. Direction of attention could be decoded from alpha power in the stimulus period, evidenced by the high correlation between alpha power RDMs and the Left vs Right model (Figure 2.10b). The alpha power difference between left and right attention conditions exhibit the same topographic pattern as in the cue period (Figure 2.8b) — there is stronger alpha power in the parietal region ipsilateral to the direction of attention. This pattern grew stronger and more significant over time, especially after the onset of the target syllable. In addition, we also noticed stronger alpha power in frontotemporal regions contralateral to the direction of attention, which was also hinted from, but not significant in, the results for the cue period. No strong correlation was shown between the Female vs Male model and alpha power RDMs during the stimulus period.

2.3.3 Similarity to behavioral performance

The average behavioral RDM shows a "checkerboard" pattern in the top-left Space-Space cell and a stripy pattern in the Space-Talker cells (Figure 2.11a). This is because, in our experimental design, conditions 1, 2, 5 and 6 present the target and late distractor syllables on the same sides (both left or both right), while conditions 3, 4, 7 and 8 have target and late distractor syllables on opposite sides. Since a distractor on the same side as the target is more likely to interfere with perception of the target, conditions 1, 2, 5 and 6 are more difficult (Space_Hard), while the other conditions in Space are relatively easier (Space_Easy). A paired t-test to compare the average behavioral performance in different condition groups found that Space_Easy (92.18%) and Talker (93.18%) have comparable behavioral results (p=0.42), which are both significantly lower than that of Space_Hard (82.20%, p<0.001). These differences between condition groups could explain the patterns observed in Figure 2.11a.

We identified one time interval in which EEG time course correlates significantly with behavioral performance — around 230 ms after the target onset (Figure 2.11b). Channel-wise correlation analysis revealed that, during this interval, the EEG time course in sevearal centro-frontal channels (Cz, FCz, Fz, C2, FC2 and F2) correlates significantly and positively with behavioral performance (Figure 2.11c).

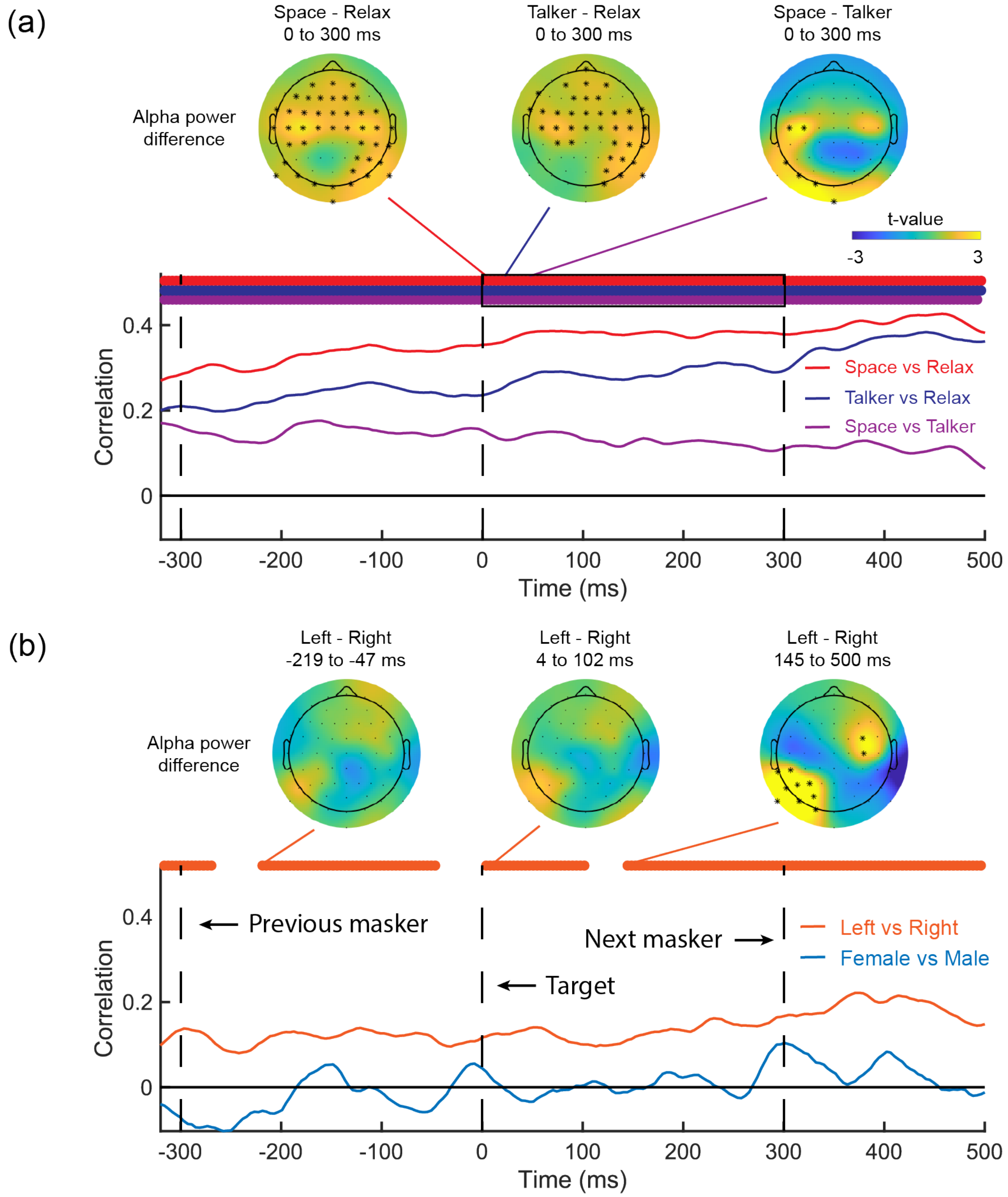


Figure 2.10: Correlation between EEG alpha power RDMs and conceptual models (shown in Figure 2.7a) in the stimulus period. (a) and (b) Traces represent correlation time courses, averaged across all subjects. The bars above indicate time intervals in which their corresponding correlation trace is significantly above zero. The topographic maps show alpha power differences between condition groups within certain time intervals, in the form of t values. Asterisks indicate channels where the observed difference is significant.

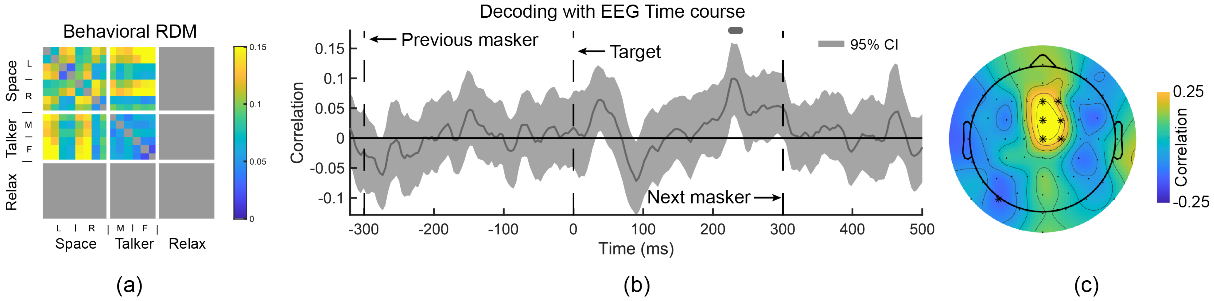


Figure 2.11: Correlation between EEG RDMs and the behavioral RDMs in the stimulus period. (a) Average behavioral RDM across all subjects. (b) Correlation time courses, derived from EEG time course. Correlation coefficients were averaged across all subjects, and the shaded area shows 95% confidence interval of the mean, estimated with bootstrapping. The bars above indicate time intervals in which their corresponding correlation trace is significantly above zero. (c) Channel-wise correlation between the EEG time course, averaged within the found significance interval in (b), and behavioral performance. Channels with a significant correlation ($p<0.05$) are denoted with asterisks.

2.4 Discussion

In this work, we studied the neural representation of auditory selective attention in EEG signals. We designed an experiment with multiple conditions that used similar stimuli, but required the listeners to adopt different listening strategies and different cognitive states. We recorded EEG from 30 subjects when they were performing the attention task and extracted neural representation features from their EEG time course or alpha oscillation power, via a neural decoding approach. These neural representation features, or representational dissimilarity matrices (RDMs), quantified the difference in brain signals between each pair of experimental conditions, and were calculated from each time point during the experiment. We examined the patterns in RDM features and how these patterns evolved over time through correlating these EEG RDMs with specific conceptual model RDMs. In addition, we also explored the link between EEG signals and the observed behavioral performance in attention tasks by correlating EEG RDMs with individual behavioral RDMs. These analyses identified time intervals during which brain signals are significantly modulated by the type of attention a listener deploys, or have strong correspondence with a listener's performance in the attention task.

2.4.1 Neural representation of brain states

The method of studying neural representations to understand brain dynamics, or representational similarity analysis (RSA, [11]), has been applied in multiple domains in neuroscience research, including visual object recognition [18], [42], [47], visual object processing [13], scene perception [48], audiovisual integration [12], and semantic categorization of sound [49]. In all of these studies, researchers presented various visual or auditory stimuli to subjects and extracted RDM features from their EEG, MEG or fMRI signals to investigate how stimuli of different categories were processed during the experiment, or how they were represented in different brain regions. In other words, these studies applied RSA to decode certain properties of the stimulus. Our study, on the other hand, deployed RSA to track the dynamics of brain states instead of input categories. The auditory stimuli we presented to subjects were either identical or only slightly different between experimental conditions. Instead, we varied the attentional states of subjects across different tasks, which may be the major property that is being characterized by the RDM features in this study. To our knowledge, only one prior work has attempted to explore the neural representation of brain states. [43] conducted an RSA study to investigate the dynamics of the fronto-parietal attention network during an audiovisual attention task. They extracted RDM features from EEG time course, and compared them with RDMs calculated from fMRI signals to track the spatiotemporal dynamics of attentional control. One challenge faced in this previous study was the trade-off between a condition-rich experimental design and the sample size for each condition: due to the fact that trials in attention studies are usually long due to multiple presentations of stimuli, subject response and feedback to answers, very limited amounts of data were collected in this prior work (20 trials per condition for 6 conditions, and 5 trials per condition for the other 12 conditions), which inevitably hindered an accurate estimation of dissimilarities. In our study, we designed an experiment comprising short trials that used fast repetitions of short syllables as stimuli. As a result, we managed to collect 36 trials for each of the 21 conditions. With these data, we could afford estimating representational dissimilarities with cross-validation (e.g., the classification approach adopted in this study), which is highly recommended as noise in the dataset can make EEG signals more dissimilar than they are in reality [50], [51]. Moreover, we decoded attention not only from EEG time courses, but also from its oscillatory band powers. The induced alpha oscillation used for attention decoding in this study is a robust neural signature of selective attention [38], [39], [52] — when our brain shifts its cognitive state, the alpha oscillation power observed in EEG is strongly modulated and shows distinctive topographic patterns, which underscores the need to examine both time course and oscillation features in brain state studies.

2.4.2 Correlation with conceptual model RDMs

2.4.2.1 Cue period

The correlation between EEG RDMs and the ideal conceptual model RDMs reveals time intervals during which the difference between attention types has strong representation in EEG signals. When using EEG time course for decoding, we identified two intervals in the cue period that are significantly correlated with the attention type RDMs (Figure 2.7a, top panel). The first interval, from 250 ms to 700 ms after the VC onset (i.e., -750 to -300 ms as shown in the figure), happened after the participants received an instruction for the type of attention (i.e., spatial, talker or no-attention) they need to use in the coming trial. The ERP analysis for channels C4 and POz revealed that though there is no discernable difference in early ERP responses among the three attention types, some minor differences in late responses (e.g., around -500 ms at POz) may have contributed to the observed above-chance correlation. The second interval (around 0 to 600 ms) contains a correlation peak for the Space vs Relax and Space vs Talker models (Figure 2.7a, top panel), indicating that the brain responses contrastingly to an AC in Space than one in Talker or Relax. This difference may be largely caused by the fact that the AC in Space comes from either left or right, while the AC in Talker and Relax always comes from the center. Therefore, we should expect a difference in neural processes during this window for different cue perception and/or AC-evoked bottom-up attention. Correlation with the Left vs Right model consolidates this view — its peak amplitude and latency is almost identical to the ones in Space vs Relax, indicating that the information being encoded during this interval is mostly the perceived location of the sound, or the bottom-up attention aroused by it. In addition, model Female vs Male also strongly correlated with EEG RDMs during the same window, but with a relatively later peak than the correlation with the Left vs Right model (Figure 2.7b, top panel). It might suggest that the spatial information of sound is encoded by the brain earlier than the encoding of its acoustic features (e.g., pitch, talker voice or gender, etc.).

The correlations between conceptual model RDMs and alpha power RDMs in the cue period exhibit a different temporal pattern — the correlation with all three attention-type models increases over time and stays high till the end of the cue period (Figure 2.8a). During this period, the listeners are expected to adjust their internal attentional state according to the given VC and AC. The fact that all three correlation traces stayed consistently and significantly above zero towards the end of the cue period may have indicated that the participants successfully switched their attentional state during this preparatory period and were getting ready for the coming task. Correlation with the Left vs Right model suggests that the spatial information of sound is represented in alpha power only after 500 ms following the AC onset (Figure 2.8b), which is much later than that in the EEG time course (Figure 2.7b, top panel). The acoustic information of sound, however, is not shown in alpha power RDMs, as suggested by their lack of correlation with the Female vs Male model. The CWT analysis revealed that Space induced significantly higher alpha power than Talker in the parieto-occipital region (Figure 2.8a), and the difference between Left and Right yielded a contrasting pattern in the same region (Figure 2.8b). These results match with what has been reported in previous EEG studies [39], and show the significant role of alpha oscillation, as a gating mechanism, in auditory spatial attention.

2.4.2.2 Stimulus period

The correlation between EEG time course RDMs and the attention-type model RDMs in the stimulus period revealed that information about attention-type is mostly encoded in EEG time course after the target onset (Figure 2.9a, top panel). However, since ERPs of different syllables are overlapping in this interval, it is difficult, if not impossible, to disentangle the effects of different ERP components (e.g., N1 or P2). Therefore, we can not confirm whether the observed post-target effects were caused by attention modulation on the N1 or P2 magnitude. Correlation with the Left vs Right and the Female vs Male model (Figure 2.9b, top panel) showed a similar pattern as in the cue period — the spatial information of sound is represented in EEG earlier than its acoustic information. The gap between these two correlation peaks in the stimulus period (around 220 ms) appears to be greater than that in the cue period (around 100 ms). This might be caused by the existence of competing sounds during this period, which increased cognitive load and slowed down the neural processes for target identification.

The correlation between EEG alpha power RDMs and all conceptual model RDMs in the stimulus period yielded results (Figure 2.8) that are akin to what we observed towards the end of the cue period (Figure 2.7). The only difference is the significantly stronger alpha power observed in frontal and parietal channels in Space and Talker than in Relax (Figure 2.8a). It re-emphasizes the important role of alpha oscillation in auditory attention.

2.4.3 Comparison to behavioral RDMs

The RSA approach also offers a way to link behavioral performance with EEG signals via the use of RDM features. In a condition-rich experiment, subjects' performance may vary substantially across different conditions. The relative difference between each pair of conditions, or in other words, the information summarized by a behavioral RDM, can help reveal time points in which EEG features covary with the observed behavioral performance. We identified an interval in the stimulus period, around 230 ms after the target syllable onset, that shows significant correlation between the EEG time course RDM and the behavioral RDM (Figure 2.11a). A subsequent channel-wise behavioral correlate analysis revealed that EEG signals in several centro-frontal channels (Cz, FCz, Fz, C2, FC2 and F2) are positively correlated with the observed performance (the topographic map in Figure 2.11b). The window around 230 ms after stimulus onset is generally accepted as when the second major positive deflection in an ERP waveform, or the P2 wave, occurs. The functional role of P2 is less understood compared to other major components in an ERP [53]. A limited number of previous studies have associated auditory P2 with various aspects of cognitive functions, such as stimuli perception [54], language learning [55], and selective attention [56], [57]. In these studies, an increase in P2 amplitude was observed after completion of musical training or language learning, or when attention was focused on the frequency or timing of sound. In other words, a stronger P2 is generally associated with a better perception of sound — this improvement in perception could be acquired through training, or through top-down attentional control. In our study, the listeners focused their attention on a particular location or talker to complete the task. Their behavioral performance greatly depended on how well they could shift their brain state to suppress the perception of distractors and enhance the perception of the target. Therefore, a positive correlation between the centro-frontal P2 amplitude and behavioral performance may have implicated a better perception of the target syllable in trials when the listeners gave a correct answer.

2.4.4 Compare RSA to classical EEG analysis

RSA is a powerful tool to study neural correlates with a condition-rich experimental design. It is different from classical EEG studies, such as ERP analysis or time-frequency analysis, where we directly compare the signal or some measures calculated from the signal (e.g., frequency band power) between two conditions or condition groups. Instead, RSA operates on the relationship among multiple conditions — it quantifies the dissimilarity between each pair of conditions, and uses these dissimilarity features as an abstraction to study brain state. In contrast to a classical analysis, RSA omits the sign of difference when accumulating evidence across samples — any difference is informative in an RSA and can be added to another regardless of their sign. This is particularly powerful when there are individual differences in the effect under test. For example, in an auditory salience study, one subject may find certain sounds to be more salient than the others, while another subject may feel the opposite. If the question being asked is which brain region encodes the perceived salience of sound, data from these two subjects will boost the observed effect under an RSA framework, but will neutralize each other in a classical analysis. Therefore, RSA can be more sensitive in capturing certain effects than a classical analysis.

Another issue in EEG studies that RSA can possibly mitigate, which is also due to individual differences of some kind, is the variability in where an effect happens across subjects. People have differences in neuroanatomy, which may cause the signals observed from the scalp to be different even if they come from two identical neural sources. In addition, factors like head size, EEG cap size and cap montage may also introduce extra variability in spatial alignment of effects between subjects. This could be a potential issue for a classical EEG study — two subjects may end up having major effects in separate channels, even if they were caused by the same neural process. When this happens, these two effects can not be easily combined, unless the misalignment is small and a high-density EEG is used. However, this is less problematic in RSA, if the timing of events is the only inference to make. As demonstrated in this study, RSA uses all the data from space (i.e., data from all EEG channels) to estimate dissimilarity between conditions at one time point. In this way, any spatial information is confined within each subject, and we only compare temporal dissimilarity measures across subjects. Therefore, as long as the effects of interest can be captured by some electrodes, the misalignment issue is irrelevant to an RSA study.

2.4.5 Limitations

In this study, we estimated the power of alpha oscillation in EEG signals using a continuous wavelet transform (CWT). CWT calculates oscillatory power within a series of sliding windows, whose length is dependent on the frequency of the oscillation at test — a longer window for a low-frequency oscillation, and a shorter window for a high-frequency oscillation. As a result, the CWT output at one time point was calculated using information both before and after that time, which essentially smeared the temporal resolution of results in our analyses. Moreover, the "alpha" band (8 – 14 Hz) oscillation being analyzed in this study spans several frequency bins in our CWT outputs, each of which had a slightly different window length. Since we averaged values across these bins to yield the estimated alpha power measure, there is an ambiguity in the exact amount of information leakage from past or future data points. Therefore, even though the results of decoding with EEG time course shares the same temporal resolution as the preprocessed EEG data (i.e., 256 Hz), the results of decoding with alpha power should be interpreted with caution — the "onset" of an effect (e.g., the moment when correlation starts to increase) should not be taken as the exact moment when certain cognitive function happens. Instead, it actually happens after a delay defined by the window length of the exact wavelet that captures this event.

Another limitation of this study is the relatively small number of conditions in our experimental design compared to previous RSA studies [18], [42]. A greater number of conditions can effectively reduce the chance of having false discoveries. We tried to control the false discovery rate by adopting a non-parametric statistical method (i.e., the cluster-based permutation test), which does not make assumptions on noise distribution and makes statistical inferences on a cluster level. We also recruited a relatively large number of participants (n = 30) for this type of study, for a greater statistical power and a less biased estimation of population means. However, a greater number of conditions and trials per condition are always preferred in RSA studies. Our design was constrained by the use of an attention task that requires subject response, which is not necessary for most previous RSA studies, since they focused on decoding stimulus rather than differences in the internal processing state across different listenting conditions. Future studies should consider designing experiments in a way that data from multiple trials and/or conditions could be acquired per subject response. This will effectively reduce the time to acquire a dataset that is ideal in size (i.e., number of conditions and trials per condition) for an RSA study.

2.4.6 Future directions

In this study, we tested the representation of a set of conceptual models in EEG through their correlation with EEG RDMs. These tests were conducted separately for each individual model, so that we could examine one aspect of attentional control at a time. For example, correlating EEG RDMs with the Space vs Relax model RDM only revealed time points in which EEG signals show differences between spatial attention and no-attention. As simple and straightforward as it is, this approach does not account for the fact that our brain is essentially a multi-tasking machinery — the cognitive functions encoded by these conceptual models may in fact happen simultaneously. One way to acknowledge the interaction between these functions during the task and examine them in a more integrative manner is to conduct a multi-model fitting analysis [9]. We can treat the conceptual models in Figure 2.4b as competing models during the task, and fit the observed EEG RDMs as a linear combination of the conceptual models at each time point. This process will yield a set of model coefficients (or weights) as a function of time, which can implicate the dynamics of cognitive functions encoded by these models, and offer a rigorous framework to compare their relative strength at a certain time point.

2.5 Conclusions

This paper is among the very few pioneering studies that adopted a representational similarity analysis framework to investigate the neural representation of attentional states. We designed a condition-rich experiment and recorded EEG signals while the listeners participated in an auditory attention task. We extracted representational dissimilarity features from the EEG time course and alpha oscillation power, and compared these features with ideal conceptual models or behavioral performance. We identified time intervals in which particular contrasts in attentional state, such as the difference between attention types or between attention to different locations, have strong representation in the EEG time course or in its alpha power. We also revealed that the listener's behavioral performance in the attention task is significantly and positively correlated with the P2 amplitude evoked by the target syllable.

Chapter 3

Neural activation and representation of auditory attention in fMRI[1]

3.1 Introduction

In our daily life, we are exposed to a mixture of all kinds of sounds — the conversation we are in with a colleague, people chatting in the background, noise of a bus running by, etc. This creates a potentially overwhelming auditory scene, in which the signal we desire is buried in abundant unwanted noises. However, most people find it effortless to focus on one target sound and ignore the distractors, thanks to auditory selective attention. Unlike in vision, where we can easily shift attention by moving the eyes, auditory attention can only be achieved covertly via cognitive control. That is, we change our internal attentional state in the brain in order to attend to or ignore a sound object [1]. This control is natural and swift for healthy people, but can be challenging for people with neurological conditions such as attention deficit/hyperactivity disorder [2], [3] or autism spectrum disorder [4], [5]. This makes it important to study the neural mechanisms that underlie auditory selective

[1] This chapter is adapted with permission from a manuscript: Winko W. An, Abigail Noyce, Alexander Pei, Barbara Shinn-Cunningham, "Neural activation and representation of auditory attention in EEG", in preparation

attention.

When we are in a busy acoustic scene, there are two major strategies that can help us enhance a target sound and suppress distractors. First, if we know the location of a target sound (e.g., cars passing us on the left when we are riding a bike), we can tune our attention to that location for a better perception of the sound. This strategy is attention to space, or spatial attention. Alternatively, if we are unsure about the target's location, but are certain about the acoustic features of the target sound (e.g. a person with a known voice chit-chatting in a noisy cafe), we can selectively attend to a sound object with that acoustic feature. This strategy is attention to acoustic features, or more generally, non-spatial attention. The neural mechanism behind spatial and non-spatial auditory attention has been studied via different methods, and there is a broad agreement that there exists a dorsal pathway, where the spatial information of sound (the "where" information) is primarily processed, and a ventral pathway, where the non-spatial information of sound (the "what" information) is encoded [6]. However, the exact role of brain areas along these pathways in spatial and non-spatial auditory attentional control remains unclear.

Functional magnetic resonance imaging (fMRI) has exquisite spatial resolution and has become a popular tool for studying cognitive function. fMRI monitors the blood-oxygen-level-dependent (BOLD) signal during an experiment, and uses it as a measure of brain activities. It has been applied to study differences between spatial and non-spatial auditory attention. Hill and Miller [58], for example, used fMRI to compare attending to a speech stream by its location or by its pitch. They identified several brain regions that are biased to attention to space, including inferior parietal sulcus (IPS), superior parietal lobule (SPL), and dorsal precentral sulcus (DPreCS). The inferior frontal gyrus (IFG) and superior temporal sulcus (STS), on the other hand, showed more activity during a pitch attention task. In another study, Michalka et al. [59] identified a visual-attention network including superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), and parietal and occipital cortex. The study also found an auditory-attention network including transverse gyrus intersecting precentral sulcus (tgPCS), caudal inferior frontal sulcus (cIFS), and superior temporal cortex. Interestingly, the study revealed that spatial and temporal attention, respectively, recruit visual and auditory attention networks in the frontal lobe, independent of sensory modality. The visual-attention network can be called during an auditory task that requires spatial information, and vice versa. These studies show that spatial and non-spatial attention may recruit different brain regions or attention networks, and such differences can be shown in fMRI data.

Classical fMRI studies, such as the general linear model (GLM) analysis, directly compare conditions to make inferences about the neural processes underpinning a cognitive function. Another analysis approach is to infer the neural representation [9]. A neural representation is a description of the information encoded in a neural unit, such a brain area [10]. Representational similarity analysis (RSA) is a framework to study neural representations from neuroimaging data [11]. Under this framework, representation can be studied through the differences among all pairs of experimental conditions. These pair-wise dissimilarities, summarized in a matrix called a representational dissimilarity matrix (RDM), can then be compared with conceptual models or with dissimilarity features derived from other neuroimaging modalities to reveal different properties of a cognitive function (e.g., its control model, dynamics, behavioral correlate, etc.). The RSA approach has been used to study the dynamics of cognitive function underlying visual object recognition. For example, Cichy et al. [18] designed an experiment in which they presented participants with images of objects. They calculated RDMs at each time point from magnetoencephalography (MEG), and at each voxel from fMRI. Then, they searched for time instances and voxels that shared similar patterns in their RDMs, which depicted a spatiotemporally resolved information flow during visual object recognition. RSA has also been applied in other fMRI studies to investigate the semantic processing of words [15], language comprehension [16], and computational models of reading [17].

One common character of these past works is the use of different stimuli across conditions

— they focused more on brain's response to different categories of sensory inputs than on its internal cognitive state. In this study, we deploy RSA to investigate the neural representation of attentional states. We matched the auditory stimuli across conditions but varied whether listeners engaged spatial or non-spatial auditory attention. We collected fMRI data while subjects performed this task and calculated RDMs at each voxel to understand how the listener's internal attentional states differed across conditions. We then compared the RDMs to several conceptual models and to behavioral performance to reveal the kind of information being encoded at each voxel and brain region.

3.2 Materials and methods

3.2.1 Participants

Nineteen adults (19 – 30 years old, 8 women) participated in this study. No participant reported hearing loss or any history of neurological disorders. The Institutional Review Board of Boston University approved this study. All participants gave written informed consent, and were paid for taking part in the study. They also filled and signed a MRI safety screening form before the experiment. All participants had previously participated in an EEG study using this same task (see Chapter 2).

3.2.2 Stimuli and task

The syllables /ba/, /da/ and /ga/, spoken by native English talkers (2 female and 2 male talkers), were used as the stimuli. The syllables were spatialized by a set of generic head-related transfer functions [40]. The simulated locations of these syllables were 90° from the left (L90), 30° from the left (L30), center, 30° from the right (R30), or 90° from the right (R90), in the horizontal plane (Figure 3.1a).

At the beginning of each trial, a visual cue (VC) was shown on the screen for one second, which could be one of three words: "Space", "Talker", or "Relax" (Figure 3.1b). "Space" (spatial attention) indicated that participants should attend to a particular location in the upcoming trial; “Talker” (talker attention) instructed the participants to attend to a specific talker by the acoustic features of his or her voice; “Relax” (no-attention) represented a control trial where no attention would be required. An auditory cue (AC) — a spatialized /a/ sound — was given after the VC to direct the participant’s attention. In “Space” attention trials, the AC was spoken by a synthesized gender-neutral talker from either L90 or R90, specifying the target location. In “Talker” attention trials, the AC was spoken by one talker from the center, specifying the target talker. In no-attention “Relax” trials, the AC carried a neutral value for both space (center) and talker gender (a synthesized gender-neutral talker). After a 1000 ms silent period, a 4-syllable mixture was played. All syllables were 600 ms long, and their onsets were separated by 300 ms. In “Space” and “Talker” attention trials, the first and the last syllables were always distractors (D1) spoken by a gender-neutral talker from the center. Of the second and third syllables, one was the target (T), with the other one being the second distractor (D2). Syllables /ba/, /da/, and /ga/ were randomly permuted among T, D1 and D2. The task was to ignore D1 and D2, identify the syllable of T, and answer using the response button box (the first button for /ba/, the second for /da/, and the third for /ga/). Visual feedback was given after each response to show whether the answer was correct. In no-attention trials, the participants were asked to ignore all syllables, and give a random answer at the end. The experiment was developed in MATLAB (2017a, Mathworks, MA, USA) using the Psychtoolbox package (V3.0.14, [60], [61]).

We designed an experiment consisted of 21 conditions (Figure 3.2). These conditions are distinguished by their attention type and characteristics of the stimuli — their location, gender of talkers, and a feature orthogonal to the type of attention being tested (i.e., a talker gender difference in spatial attention conditions, and a location difference in talker attention conditions). The stimuli used in active attention conditions (Space or Talker) were re-used in Relax conditions, but required no attention.

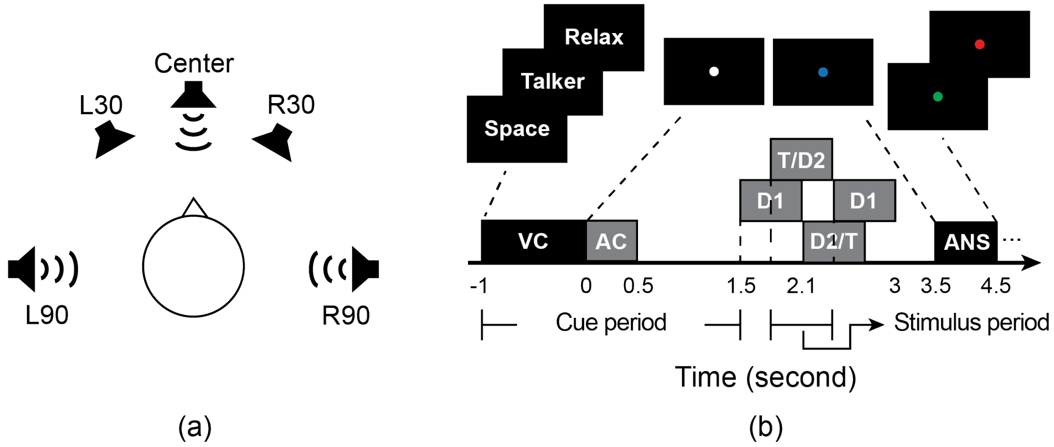
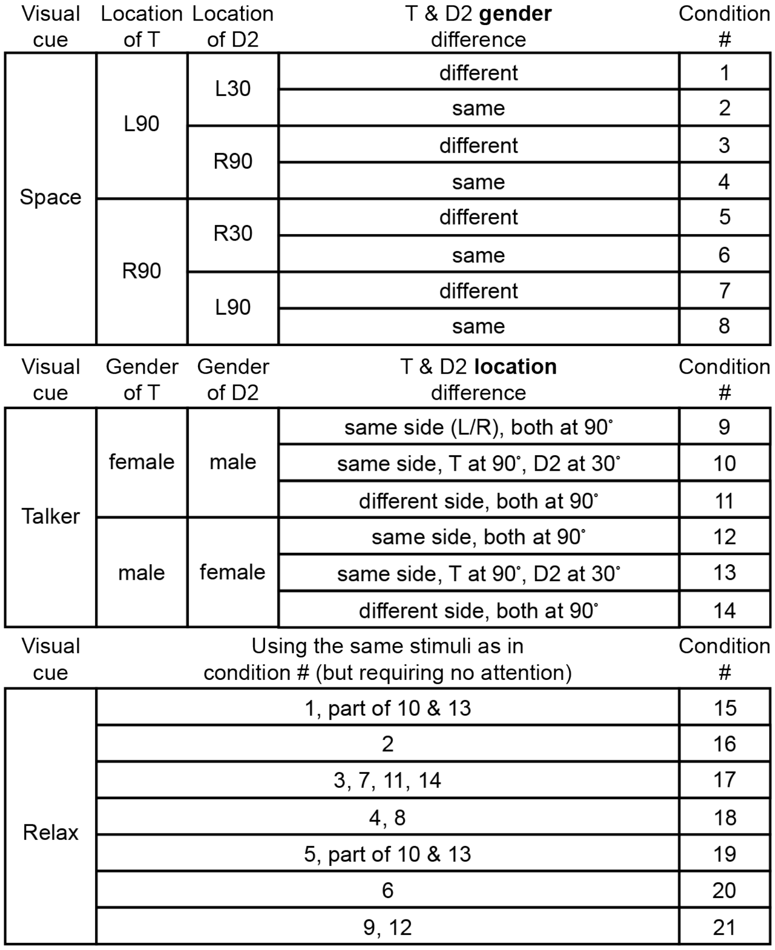


Figure 3.1: Experimental task. (a) Spoken syllables were spatialized to center, 30° left (L30) or right (R30), and 90° left (L90) or right (R90), always in the horizontal plane. This figure is showing one possible scenario where sounds come from L90, R90 and center. (b) Illustration of the events within a trial. A visual cue (VC) showed the type of attention required for this trial, followed by an auditory cue (AC) specifying the target (a specific location or talker). A 4-syllable mixture was played one second after the AC. The first and the last syllables were always distractors (D1). Of the second and the third syllables, one was the target (T), and the other was the second distractor (D2). All syllables were 600ms long, and their onsets were 300ms apart. The participants were instructed to respond when the fixation dot turned blue. Feedback was given at the end of each trial via a green (correct) or red (incorrect) fixation dot.

3.2.3 Data collection

All experiments were conducted at the Boston University Cognitive Neuroimaging Center. fMRI data were collected with a 3T Siemens MAGNETOM Prisma scanner (64-channel head coil), equipped with a PROPixx Lite Projector (VPixx Technologies, QC, Canada) for presenting visual stimuli and Sensimetrics S14 insert earphones for playing sound stimuli. Foam materials were applied between the head coil and earphones to reduce the perceived loudness of scanner noise.

The participants completed the experiment in two visits that were scheduled on different days within a week. In their first visit, the subjects started with practising the attention tasks on a laptop in a preparation room. They were required to attempt a 21-trial test run of the experiment (i.e., one trial for each of the 21 conditions) repeatedly until their



Visual cue	Location of T	Location of D2	T & D2 **gender** difference	Condition #
Space	L90	L30	different	1
			same	2
		R90	different	3
			same	4
	R90	R30	different	5
			same	6
		L90	different	7
			same	8

Visual cue	Gender of T	Gender of D2	T & D2 **location** difference	Condition #
Talker	female	male	same side (L/R), both at 90°	9
			same side, T at 90°, D2 at 30°	10
			different side, both at 90°	11
	male	female	same side, both at 90°	12
			same side, T at 90°, D2 at 30°	13
			different side, both at 90°	14

Visual cue	Using the same stimuli as in condition # (but requiring no attention)	Condition #
Relax	1, part of 10 & 13	15
	2	16
	3, 7, 11, 14	17
	4, 8	18
	5, part of 10 & 13	19
	6	20
	9, 12	21

Figure 3.2: The 21 experimental conditions in this study. These conditions differ by their VC and the location or gender of T and D2. The "Relax" conditions used the same stimuli as in "Space" and "Talker", but required no attention. Adapted with permission from An et al. [62].

response accuracy reached 75%. This is to ensure that they fully understood the instructions and would correctly allocate attention during the actual experiment. Next, we acquired structural scans using both a T1-weighted sequence (176 sagittal slices, FOV = 256 mm^2, TR = 2530 ms, TE = 1.69 ms, flip angle = 7°), and a T2-weighted sequence (176 sagittal slices, FOV = 256 mm^2, TR = 3200 ms, TE = 425 ms, flip angle = 120°). The T1 images were used for preprocessing the functional data, which will be discussed in Section 3.2.4. The T2 images were not used in this study. Following the structural scans, the participants were asked to perform 5 runs of attention tasks. Each run consisted of 63 trials with equal presence of each condition (i.e., 3 trials from each of the 21 conditions), whose order of appearance was randomly permuted, forming an event-related design. These functional scans were acquired from the whole brain with a fast sampling rate using multiband pulse sequences (TR = 650 ms with a multiband factor of 8, TE = 3.48 ms, FOV = 720 mm^2, flip angle = 52°, resolution = 2.3 mm isotropic). To address the issue of reduced tissue contrast due to multiband acquisition, a single-band reference images with full tissue contrast was acquired at the beginning of each run of functional scans for realignment and registration with structural scans (see Section 3.2.4). The inter-trial intervals in these runs were integer seconds with the value randomly chosen from 5 to 10 with equal probability. In total, 1000 frames were collected from each functional run. An optional short break was given between two runs. In their second visit, the participants completed another 7 runs of functional scans, making the total number of trials 756 (i.e., 36 trials per condition).

3.2.4 Preprocessing

The fMRI data were preprocessed using SPM12 [63], following its recommended pipeline [64]. No slice timing correction was applied due to the low TR employed in this study [65]. The single-band reference image with full tissue contrast acquired at the beginning of each functional run was used to realign the time series of images and register with the T1 structural scan [65]. Segmentation was performed on T1 scans using the voxel-based morphometry technique, and the segmented data of individual subjects were normalized to the MNI152 linear space for group-level analysis.

Figure 3.3 illustrates the analyses conducted in this study and the preprocessing required for each of them. For representational similarity analysis (RSA, the blue blocks in Figure 3.3, to be discussed in Section 3.2.6), no more preprocessing procedures were needed after normalization. For the general linear model (GLM) analysis to be discussed in Section 3.2.5 (the red block in Figure 3.3), additional spatial smoothing was applied to increase the signal-to-noise ratio. In this study, a Gaussian smoothing kernel with an isotropic 6-mm full width at half maximum was used to serve this purpose.

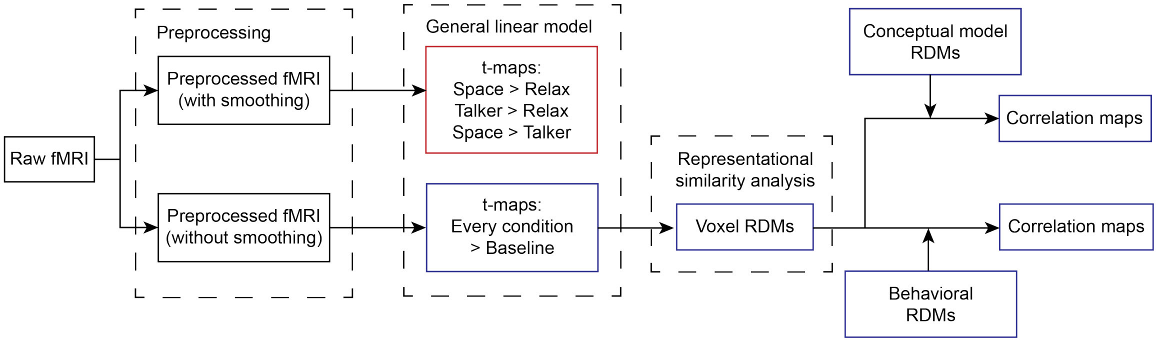


Figure 3.3: Schematic diagram for the analyses conducted in this study. Raw fMRI data were first preprocessed with and without spatial smoothing. The smoothed data were used for a general linear model (GLM; in the red block) analysis, in which three contrasts were examined — Space > Relax, Talker > Relax, and Space > Talker — to explore the relative strength of activation between different attention types. The unsmoothed data were used for a representational similarity analysis (RSA; in the blue blocks). First, we conducted another GLM analysis to derive the contrast between each of the 21 experimental conditions and an explicitly modeled resting state. Then, we extracted multivariate features from these condition-wise t-statistics, and calculated the dissimilarity between each pair of conditions at each voxel. This information was summarized in a representational dissimilarity matrix (RDM). Finally, we correlated these voxel-level RDMs with 1) some conceptual model RDMs to discover patterns in these fMRI RDMs, and 2) behavioral RDMs to explore the relationship between voxel-level activation and the observed behavioral performance in the attention tasks.

3.2.5 General linear model

We conducted a GLM analysis to study voxel-level effects of auditory attention. A set of regressors were created for GLM, including:

- Active attention, modeled as time-locked to the onset of auditory cues (duration = 3 seconds, spanning a period during which the auditory cue and stimulus were played), labelled with their corresponding condition index (#1 – #21)
- Onset of subject response and visual feedback (an impulse function)
- An explicitly modelled resting state (2 seconds after the subject response, duration = 5 seconds)

- A set of nuisance regressors for head movement
- Another nuisance regressor for the number of run

All non-nuisance regressors were convolved with a canonical hemodyamic response function (a double gamma function, 6-second delay of response relative to onset) to model blood oxygen level change following external events. The time and dispersion derivatives of the convolution were also included as regressors to allow for variations in response across subjects and voxels. A first-level autoregression model was used to account for serial correlations due to aliased biorhythms and unmodelled neuronal activity. A high-pass filter (cutoff = 128 seconds) was applied to remove slow signal drifts.

To explore the difference between each pair of attention types, three contrasts were examined: Space > Relax, Talker > Relax, and Space > Talker. For Space > Relax, regressors for the eight spatial attention conditions were assigned with a weight of 1/8, while regressors for the seven no-attention conditions were assigned with a weight of -1/7 (i.e., the regressor weights were normalized by the number of conditions in each attention type). Talker > Relax and Space > Talker were setup in the same manner. In addition, we examined two other contrasts, Left > Right and Talker Female > Talker Male, to explore the voxel-wise activation difference between attention to different directions, or between attention to different genders.

Group level statistics were carried out using the threshold-free cluster enhancement method (TFCE; for details, see [66]), with the null hypothesis being that the conditions under test caused the same level of activation, or, in other words, the contrast between these two conditions was zero. For each pair of condition groups (e.g., Space versus Relax), we first compared their GLM contrast (an output of the GLM analysis, showing the difference in voxel-wise activation between condition groups) with zero using a one-sample t-test, which generated a brain volume of t-values for each contrast. Then, we calculated the TFCE score of each voxel based on this volumetric t-statistic image following the approach proposed in Smith and Nichols [66] — basically, a voxel-wise TFCE score was calculated not only with the t-value of the voxel of interest, but also with the t-values of other voxels in the same cluster. The general idea behind this TFCE approach is to enhance areas of signal that exhibit some spatial contiguity without relying on a self-selected cluster-formation threshold [66].

To determine the significance threshold for a particular contrast, we used a permutation approach to derive a null distribution of TFCE scores for that contrast. First, we randomly permuted the data labels to create a permutation dataset. For a one-sample t-test, this permutation process could be effectively achieved by negating the sign of data for a randomly selected subset of all subjects. Next, using the method described above, we calculated the TFCE scores of all voxels, of which only the maximum was used to form the null distribution. We repeated this process for 10,000 time, yielding a distribution with 10,000 TFCE scores, which were all derived from random permutation. Then, we used the 95th percentile of this distribution as the significance threshold: all voxels with a TFCE score greater than this threshold were deemed to have significantly different levels of activation between the two condition groups at test.

3.2.6 Measuring representational dissimilarity

We extracted pairwise differences among all conditions from fMRI data for the RSA study. First, we conducted a second GLM analysis with the same set of regressors as the one used in Section 3.2.5. Different from the previous GLM analysis, here we used the preprocessed data without spatial smoothing. The purpose of using unsmoothed data is to retain as much voxel-wise independence as possible, where important features that can help distinguish conditions may exist. Next, for each subject, we calculated the contrast between each of the 21 conditions and the explicitly modelled resting state. We then extracted voxel-wise multivariate features for each condition: we concatenated t values of the Condition X > Rest contrast at a particular voxel and its surrounding neighbours (within a radius of 4 voxels) into a vector, and used this vector as the feature for condition X at that anchor voxel. An illustration of this method is shown in Figure 3.4. The voxel-wise dissimilarities between each pair of conditions was estimated as the Euclidean distance between the feature vectors of the respective conditions at each voxel. This process yielded a representational dissimilarity matrix (RDM) for each voxel, where each entry reflected the dissimilarity of the neural responses across a particular pair of conditions.

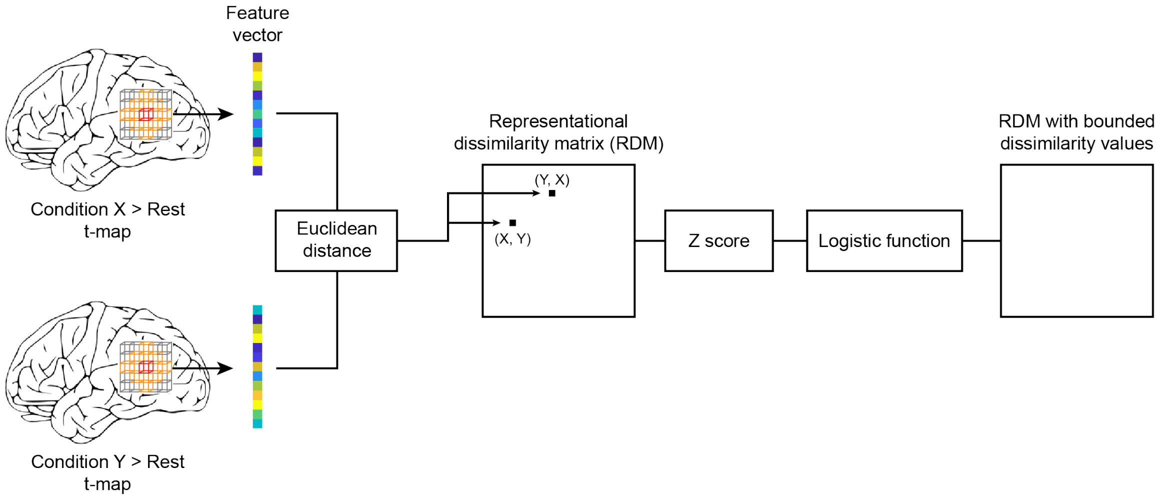


Figure 3.4: Schematic diagram for calculating representational dissimilarity. We estimated the voxel-wise dissimilarity between each pair of conditions using the multivariate feature vector that we extracted at each voxel via a searchlight approach. For a particular voxel (i.e., the "anchor" voxel, with a red color in the figure) in the t-map for Condition X > Rest, we concatenated the t-value at this voxel together with t-values at its neighbouring voxels (the ones with an orange color in the figure) into a feature vector to represent condition X at this anchor voxel.The same method could be applied to derive a feature vector for condition Y at the same voxel. Then, we calculated the Euclidean distance between these two vectors to estimate the dissimilarity between condition X and Y. This dissimilarity value was saved at (X,Y) and (Y,X) of a representational dissimilarity matrix (RDM), which could be used to summarize the difference between each pair of conditions at the anchor voxel. After an RDM was filled, we calculated the z-score of each element in this RDM and transformed these values with a logistic function, so that the dissimilarity values in an RDM were always bounded by -1 and 1.
Note: For better visualization, the searchlight method in this figure is illustrated in a 2-D plane with a radius of 2. In the actual study, this searchlight method was conducted in the 3-D space with a radius of 4.

Specifically, each row or column of an RDM corresponds to a condition index, and each element stores the dissimilarity value between two conditions: the transformed distance between condition X and Y is saved at (X , Y) and (Y , X) of the RDM. With this approach, the major diagonal of an RDM, representing the difference between one condition and itself, is filled with zeros and is excluded from any analysis. To reduce the effect of outliers on the estimation of dissimilarity (e.g., a huge distance measure in one dimension of the feature vector due to the existence of noise), we first calculated the z-score of each element in an RDM (i.e., we subtracted the mean of the RDM and divided by its standard deviation; major diagonal excluded), and transformed the z-scores with a standard logistic function. The dissimilarity values, therefore, are bounded by -1 and 1.

The ultimate output of this process was a set of fMRI RDMs as a function of space — there was one RDM at each voxel. These RDM features captured how neural activities at each voxel differ across attentional conditions, which describes the neural representation of auditory selective attention in fMRI. We analyzed the patterns in these RDMs through their correlation with conceptual models (see Section 3.2.7), and calculated the average RDMs within several regions of interest (ROIs) to study the information encoded in these regions (see Section 3.3.2).

3.2.7 Comparison to conceptual models

The RDM feature summarizes how one experimental condition differs from another. In our study, since the conditions could be categorized into certain condition groups (e.g., Space, Talker, etc.), this RDM feature also contained information about how a specific type of attention differed from another. For example, if we are interested in the difference between two attention types, or the difference between conditions within the same attention type, we can refer to certain portions of an RDM for this information — the square blocks along the major diagonal of an RDM (Space-Space, Talker-Talker and Relax-Relax in Figure 3.5a) represent how one condition differs from another within the same attention type, while the blocks off the diagonal (Space-Talker, Space-Relax and Talker-Relax) show the difference between attention types. Moreover, there were also micro-structures within Space-Space and Talker-Talker. Four conditions in Space were about left attention, and the other four were about right attention. They formed micro-blocks that showed differences within attention to the same side (Left-Left and Right-Right), and blocks for attention to different sides (Left-Right, Figure 3.5a). Similarly, the six Talker conditions were separated between attention to a female talker and attention to a male talker. Thus, the Talker-Talker block could be further divided into Female-Female and Male-Male for within-gender comparisons, and Female-Male for between-gender comparisons.

Patterns in RDM features can be quantified via their correlation with specific conceptual model RDMs. These models, consisted of 0s and ±1s, are built to capture discrimination among conditions: positive for conditions that are very different on the dimension of interest, negative for conditions that are very similar, and zero for conditions that are irrelevant. Figure 3.5b and 3.5c show two such models: Left vs. Right and Female vs. Male. The first, Left vs. Right, model was built to study the direction of attention — the elements showing the difference between left and right attention (i.e., the Left-Right micro-blocks) were assigned with a value of $+1$, while the elements showing the difference within left or right attention conditions (i.e., the Left-Left and Right-Right micro-blocks) were assigned with a value of -1. The rest of the matrix was filled with zeros, and was not used in any correlation calculation. The second model, Female vs Male, was created in a similar manner with an aim to study attention to different genders (i.e, $+1$ for Female-Male micro-blocks, and -1 for Female-Female and Male-Male micro-blocks). For each subject, we correlated these model RDMs with the fMRI RDM at each voxel. A high correlation between the RDMs indicates that this brain location contains discriminant information about the difference in attentional state (i.e., attention to different directions or genders).

Group level statistical analysis was conducted in the same way as in Section 3.2.5. Correlation coefficients were first Fisher z-transformed, and were then used as the inputs for a group-level TFCE with the null hypothesis being that the correlations under test are zero. As in Section 3.2.5, the null distribution was derived from randomly permuting the data labels for 10,000 times, and any observed effects with an TFCE score greater than the 95th percentile threshold were deemed as significant.

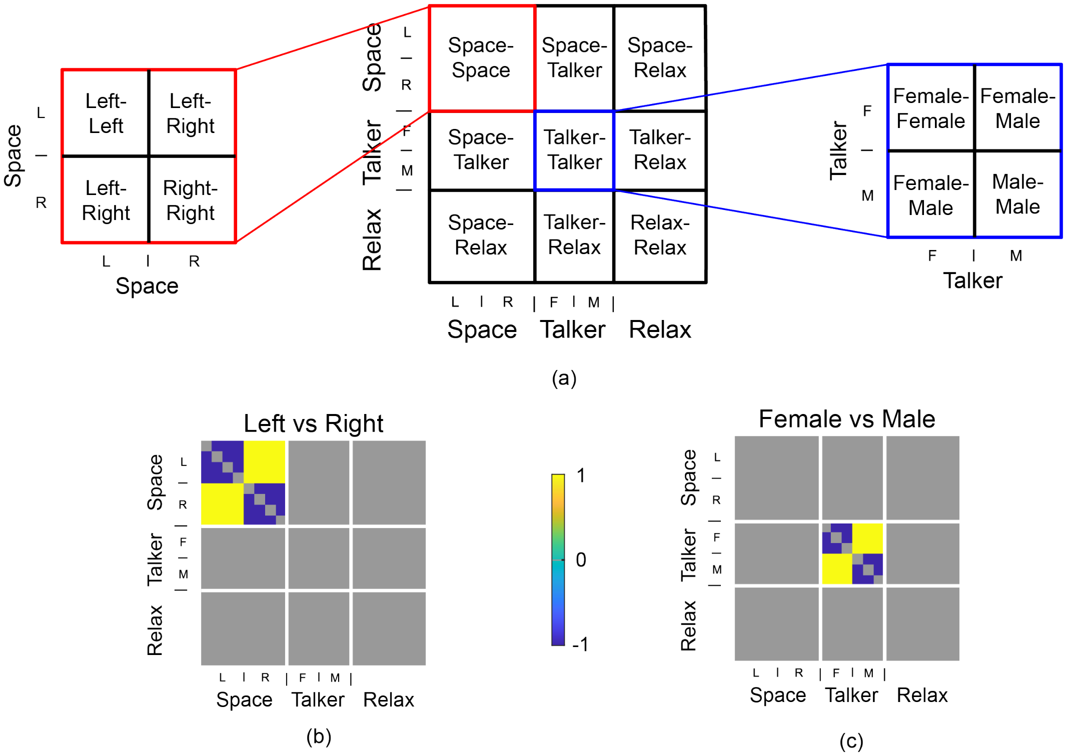


Figure 3.5: (a) A representational dissimilarity matrix (RDM) in this study can be divided into several blocks, and each block represents different information. The top left portion of the RDM, Space-Space (the red block), shows the difference between conditions within the spatial attention group. Similarly, the other two square blocks along the major diagonal, Talker-Talker and Relax-Relax, also show differences within their respective condition groups. The blocks off the diagonal (Space-Talker, Space-Relax and Talker-Relax), on the other hand, store dissimilarity values between conditions groups. In addition, within Space-Space and Talker-Talker, there are also micro-structures showing comparisons between attention to different locations (left or right), or between attention to talker of different genders (female or male). (b) & (c) Two conceptual model RDMs used to study patterns in fMRI RDMs. In Left vs Right, the Left-Right blocks are filled with +1s, and the Left-Left and Right-Right blocks are filled with −1s — this model RDM is built to study the representation of "where to attend". In Female vs Male, we assigned +1s to the Female-Male blocks, and −1s to the Female-Female and Male-Male blocks – this model is built to study the representation of "what to attend".

These correlation analyses identify voxels in which certain information (e.g., direction of attention) can be decoded from fMRI signals. However, it does not reveal what is driving the decoding, for example, whether one condition group has greater activation than another. One way to possibly answer this question is to calculate the percent signal change (%SC) between condition groups of interest within the voxels we identified as significantly correlated with the ideal model. To compare attention to different directions, we calculated the %SC of all left-attention conditions (conditions 1 – 4) relative to all right-attention conditions (conditions 5 – 8). To compare attention to different genders, we calculated the %SC of all conditions for attention to a male talker (conditions 9 – 11) relative to all conditions for attention to a female talker (conditions 12 – 14). These %SC measures were averaged within ROIs, and then were compared with zero via a one-sample t-test (alpha = 0.05). This computation complements the study of neural representation, which only discovers the existence of differences, by identifying information about what these differences are, which could help us better understand the mechanisms of attention.

3.2.8 Similarity to behavioral performance

Participants' behavioral performance varies across task conditions. The RDM feature allows us to ask whether brain activity in any brain regions is particularly relevant to the subject's behavioral performance. We constructed behavioral RDMs that comprise the absolute difference in behavioral performance between conditions. Only active attention conditions (8 spatial and 6 talker conditions) were used in this analysis. We correlated individual behavioral RDMs with their respective fMRI RDMs at each voxel and transformed the correlation coefficients using the Fisher z-transformation. The same TFCE method as the one used in Section 3.2.7 was applied here to draw group-level statistical inferences.

3.3 Results

3.3.1 General linear model

We compared the level of activation between attentional conditions at each voxel with a GLM analysis. The t-statistics of each contrast, originally in the MNI152 volume space, were projected onto an inflated cortical surface model using Nilearn (v0.8.0, [67]) for better visualization (Figure 3.6). To facilitate relating the observed effects to specific brain regions, we overlaid the contour of several selected regions of interest (ROIs) defined in Destrieux et al. [68].

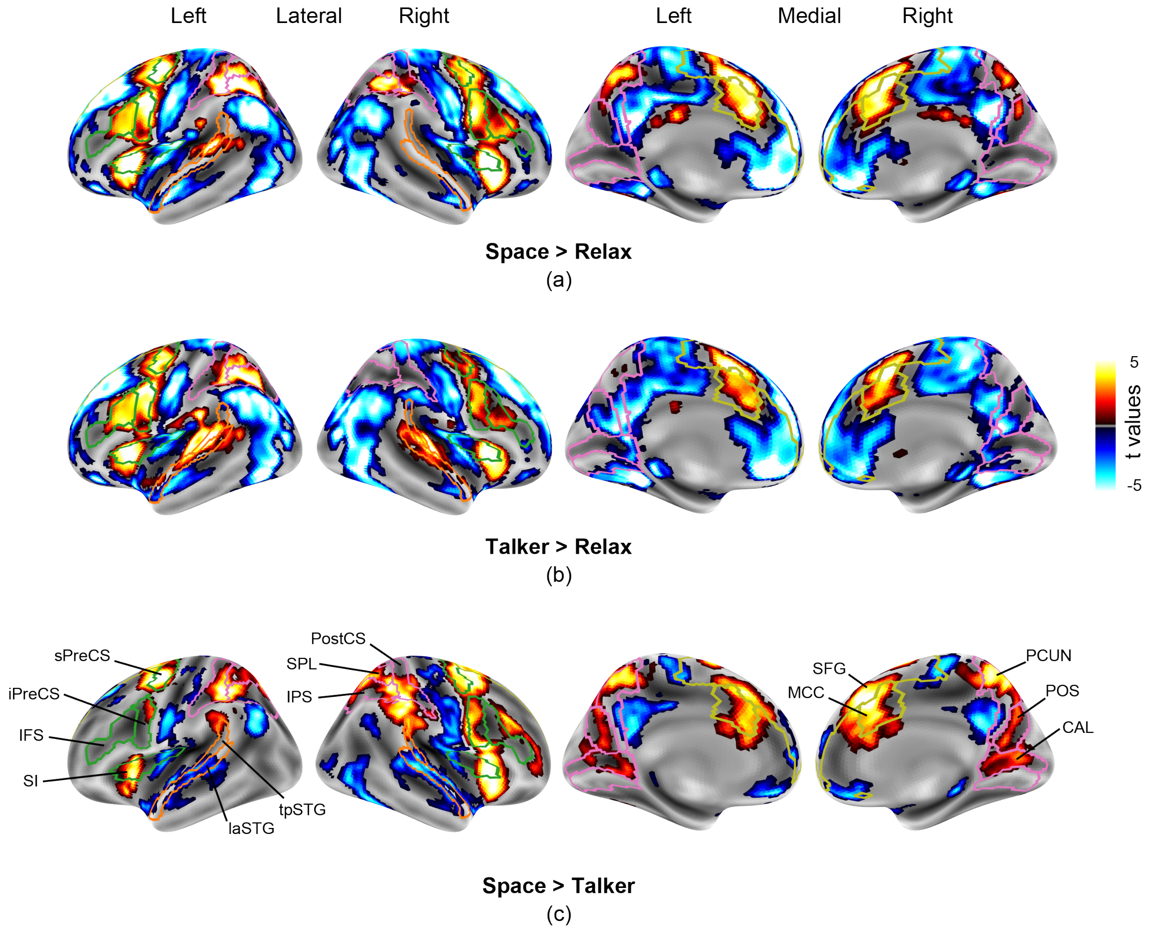


Figure 3.6: T statistics of three GLM contrasts: (a) Space > Relax, (b) Talker > Relax, (c) Space > Talker. These results are masked by their statistical significance ($p < 0.05$), which was determined by a threshold-free cluster enhancement (TFCE) method. We overlaid the contours of several regions of interest (ROIs) onto each plot for easy mapping of the effects. These ROIs were previously defined in Destrieux et al. [68]. We identified several ROIs that were actively engaged in auditory selective attention: the superior and inferior pre-central sulcus (sPreCS and iPreCS), inferior frontal sulcus (IFS), superior insula (SI); the lateral aspect and the temporal plane of the superior temporal gyrus (laSTG and tpSTG); the post-central sulcus (PostCS), superior parietal lobule (SPL), inferior parietal sulcus (IPS), precuneus (PCUN), parietal-occipital sulcus (POS), calcarine sulcus (CAL); the medial cingulate cortex (MCC), and the superior frontal gyrus (SFG). In addition, the default mode network and the precentral sulcus were significantly more active in Relax, when no attention was required, than in Space or Talker.

3.3.1.1 Space > Relax

Auditory attention activates a wide attention network that encompasses regions in the frontal, parietal, temporal and occipital lobe. We observed significantly stronger signal in Space than Relax in superior and inferior precentral sulcus (iPreCS and sPreCS), inferior frontal sulcus (IFS), superior insula (SI), postcentral sulcus (PostCS), intraparietal sulcus (IPS), precuneus (PCUN), medial cingulate cortex (MCC) and superior frontal gyrus (SFG) in both hemispheres (Figure 3.6a). In addition, the lateral aspect and the temporal plane of the superior temporal gyrus (laSTG and tpSTG) and the superior parietal lobule (SPL) showed significantly higher activation in Space than Relax only in the left hemisphere (Figure 3.6a). The default mode network (DMN) and the precentral sulcus, on the other hand, were more engaged when no attention was required.

3.3.1.2 Talker > Relax

The Talker > Relax contrast generated a similar activation pattern as Space > Relax (Figure 3.6b). The iPreCS, sPreCS, IFS, SI, MCC and SFG in both hemispheres showed significantly stronger signal in Talker than in Relax, while the DMN and the precentral sulcus behaved the opposite. Different from Space > Relax, the laSTG and tpSTG in both hemispheres were prominently more engaged in Talker than in Relax. Moreover, except for the SPL and IPS on the left hemisphere, none of the ROIs in the parietal and occipital lobe showed significant activity in Talker > Relax.

3.3.1.3 Space > Talker

We directly compared Space with Talker to investigate the specificity to attention types of each ROI. We observed significantly stronger activities in Space than in Talker in the parietal and occipital lobe (Figure 3.6c). The parieto-occipital sulcus (POS) and calcarine sulcus (CAL) in both hemispheres were highly engaged in Space, in addition to the PCUN, PostCS, SPL and IPS, which have been seen with strong activity in the Space > Relax contrast. The sPreCS, iPreCS, SI, MCC and SFG were also more active in Space than in Talker, while IFS in both hemispheres did not differ significantly in activation between attention types. The laSTG and tpSTG, on the other hand, showed the opposite pattern — signals in these regions were notably stronger in Talker than in Space.

3.3.2 Representational dissimilarity matrices

We explored the neural representation of auditory attention by studying the difference between each pair of conditions. First, we contrasted each condition with an explicitly modeled resting state. The dissimilarity between two conditions at each voxel was then estimated by the Euclidean distance between the multivariate feature vectors of those conditions in a searchlight centered on this voxel. These pair-wise differences were summarized in a representational dissimilarity matrix (RDM) — each row and column of an RDM represent a condition index, and the value stored at element (X,Y) and (Y,X) of an RDM represents the estimated difference between condition X and condition Y. This process yielded a set of RDMs as a function of brain location — each voxel has one specific RDM.

In Figure 3.7, we present several mean RDMs averaged within selected ROIs and their corresponding multidimensional scaling (MDS) plots. These MDS plots visualize the 21 conditions in a 2-D feature space and show the information encoded in each ROI. For example, if the MDS plot for one ROI shows that the Relax conditions form a cluster and separate from Space and Talker conditions, it means this ROI is encoding information about ”attention or not”. For most regions depicted, Space and Talker could be well distinguished from Relax, with the ventral posterior cingulate cortex (vPCC) being the only exception. Its MDS plot shows a random pattern, indicating that no specific attention-related information is encoded in this region. In sPreCS, IFS, and CAL, the dissimilarity values were high in the bottom left (Space-Relax) and middle (Talker-Relax) portion of the RDMs, and low in the bottom right portion (Relax-Relax), indicating that the activation pattern in these regions was dependent on attentional state. More specifically, for sPreCS and CAL, the average difference between Space and Relax was greater than that between Talker and Relax (i.e., higher values in Space-Relax than in Talker-Relax). This was not seen in IFS, where these two differences were comparable, which might have shown some variability across these ROIs in terms of their specificity to spatial or talker attention. In addition, special micro-structural patterns that discriminates the direction of attention or the gender of the attended talker were only observed in CAL and IFS. In CAL, left attention conditions were well separated from right attention conditions, indicated by the high dissimilarity between these two condition groups, and the low dissimilarity within each of them. In IFS, we observed a similar separation between attention to a female talker and attention to a male talker. These results suggest that the information about "where to attend" or "what to attend" were highly encoded in these two regions.

3.3.3 Correlation with conceptual models

The pattern of RDMs can be learned from their correlation with specific conceptual model RDMs. These conceptual model RDMs consist of 0s and ±1s, and resemble an ideal case where conditions of different groups (e.g., one condition for attending left and one condition for attending right) can be perfectly decoded (thus assigned with a value of +1), and conditions of the same group can not be differentiated at all (thus assigned with a value of -1). If the RDM at one voxel correlates with a model RDM better than chance, it indicates there is information at that brain location that can distinguish one condition group from another, or, in other words, that brain region behaves differently in these two attentional states.

We first correlated the fMRI RDMs with the Left vs Right model (Figure 3.5b) to explore if direction of attention has strong representation in any brain regions. The correlation coefficients were first Fisher z-transformed, and then tested with the TFCE method to discover voxels with a correlation coefficient significantly above zero on the group level. The t-statistics of this comparison were rendered onto a surface brain model using Nilearn (v0.8.0, [67]) and shown in Figure 3.8a. We observed above-chance correlation in the CAL,

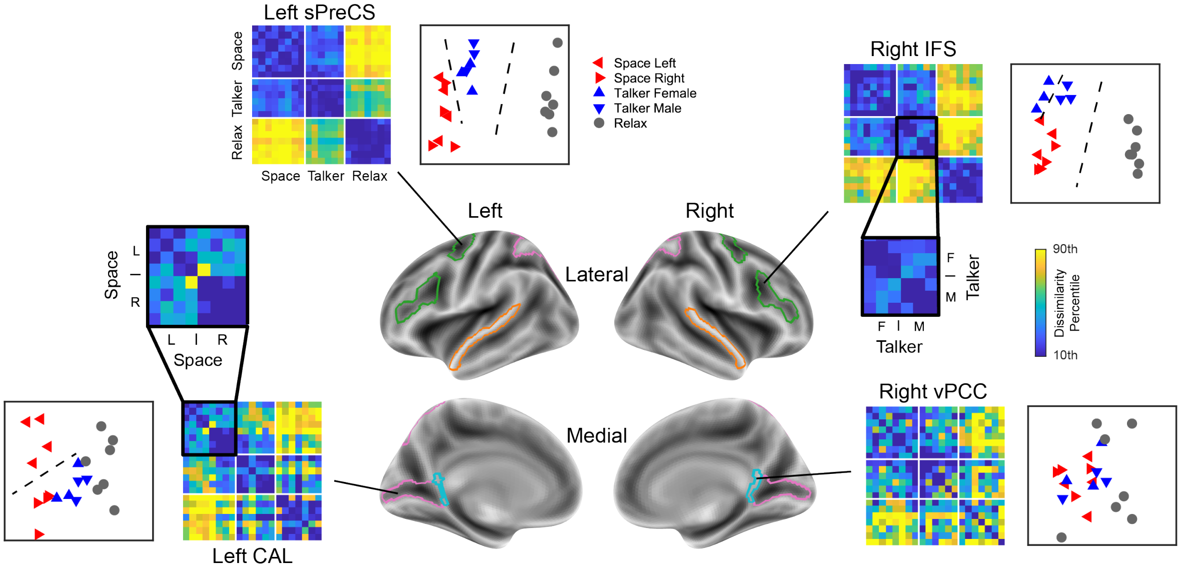


Figure 3.7: Average representational dissimilarity matrices (RDMs) of several selected regions of interest (ROIs) and their corresponding multidimensional scaling (MDS) plots. The MDS plots visualize the 21 conditions in a 2-D feature space. High dissimilarity between Space or Talker and Relax (i.e., the bottom left and bottom middle portion of the RDM) was observed in all the selected ROIs, while the dissimilarity between conditions within Relax (i.e., the bottom right portion of the RDM) was low in most regions. Differences between Space and Talker conditions (i.e., the middle left portion of the RDM) were medium to high in CAL and IFS. There were also micro-structures embedded in these RDMs, indicating that attention to different directions and attention to talkers of different genders could also be decoded in some ROIs: in CAL, the dissimilarities between left and right attention conditions were high, while the dissimilarities between conditions within left or right attention groups were low (see the enlarged view of the top left portion of the RDM); a similar dissimilarity pattern was observed in IFS for Talker vs Talker (see the enlarged view of the center portion of the RDM).
Acronyms: sPreCS, superior pre-central sulcus; IFS, inferior frontal sulcus; CAL, calcarine sulcus; vPCC, ventral posterior cingulate cortex.

POS and lingual gyrus (LING) in both hemispheres. None of the ROIs in the parietal lobe showed strong representation of the direction of attention.

To investigate whether the activation level at these above-threshold voxels was modulated by the direction of attention, we calculated the percent signal change (%SC) in left-attention conditions relative to right-attention conditions. These %SC measures were averaged within ROIs in each hemisphere, and are shown in Figure 3.8c. Significant %SC was observed in left LING (p=0.0163, Cohen's d=0.61), left CAL (p=0.0064, Cohen's d=0.71), and left

POS (p=0.0099, Cohen's d=0.66) — all ROIs in the left hemisphere. Their counterparts in the right hemisphere, however, showed no significant %SC in attention to left relative to attention to right.

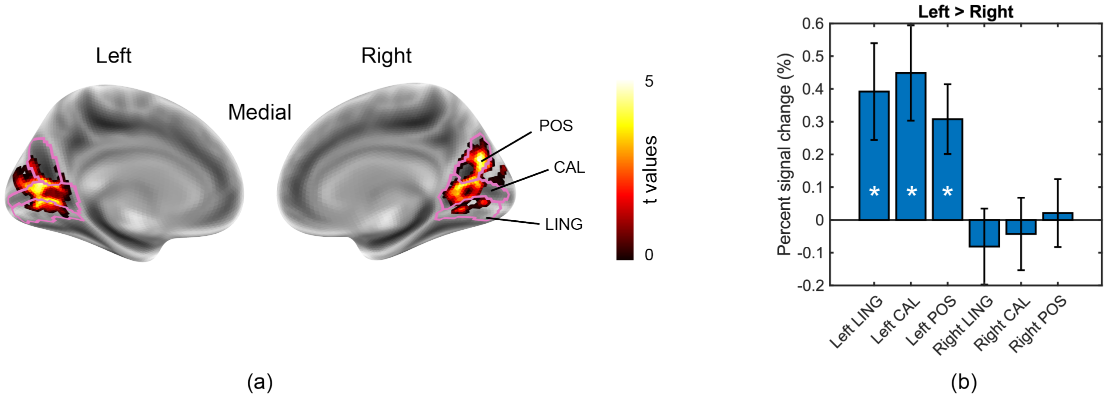


Figure 3.8: (a) Voxels in the parieto-occipital region showed significant correlation between their voxel-wise fMRI RDMs and the conceptual Left vs Right model RDM. We compared the Fisher-transformed correlation coefficients at each voxel with zero using a one-sample t-test, and rendered the t-statistics of this comparison onto a cortical surface model. The results were masked by their statistical significance derived from the threshold-free cluster enhancement method. (b) Percent signal change (%SC) in left-attention conditions relative to right-attention conditions. Regions in the left hemisphere showed significant %SC, while regions in the right hemisphere did not. Error bars indicate the standard error of the mean. $*$, $p<0.05$
Acronyms: POS, parietal-occipital sulcus; CAL, calcarine sulcus; LING, lingual gyrus.

We also correlated the fMRI RDMs with a Female vs Male conceptual model (Figure 3.5c) to explore if any brain regions have representation of the gender of the attended talker. With the same TFCE statistical analysis, we observed above-chance correlation only in the right IFS (Figure 3.9a). A similar %SC analysis as described above revealed no significant %SC in the right IFS (p) when attention was focused on a female talker relative to when it was on a male talker (Figure 3.9b).

3.3.4 Similarity to behavioral performance

The participants demonstrated varied behavioral performance across attentional conditions. The average behavioral RDM, derived from taking the absolute difference in response accu-

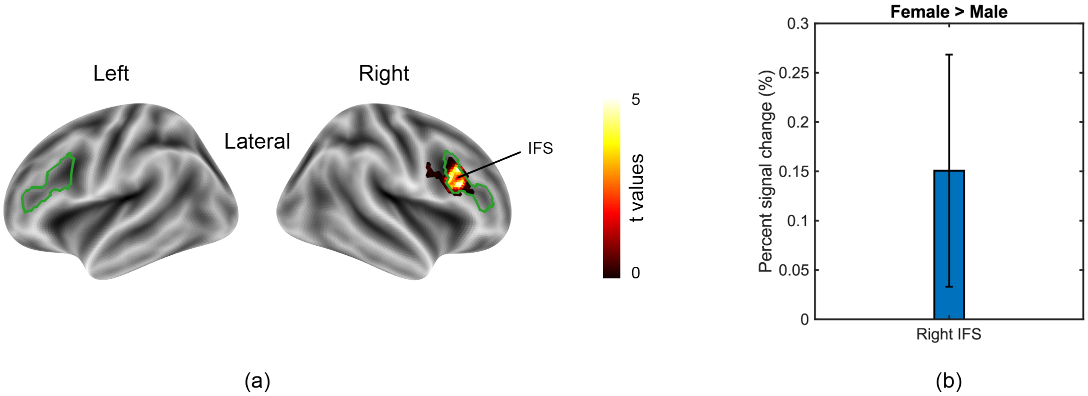


Figure 3.9: (a) Voxels in the right inferior frontal sulcus (IFS) showed significant correlation between their voxel-wise fMRI RDMs and the conceptual Female vs Male model RDM. We compared the Fisher-transformed correlation coefficients at each voxel with zero using a one-sample t-test, and rendered the t-statistics of this comparison onto a cortical surface model. The results were masked by their statistical significance derived from the threshold-free cluster enhancement method. (b) Percent signal change (%SC) in attention to a female talker relative to attention to a male talker. No significant %SC was observed in the right IFS. The error bar indicates the standard error of the mean.

racy between each pair of conditions, is shown in Figure 3.10a. We noticed a checkerboard pattern in the top left block (Space-Space), a striped pattern in the top right and bottom left blocks (Space-Talker), and a block with negligible difference values in the bottom right (Talker-Talker). This is because, in our design, condition 1, 2, 5 and 6 have their target (T) and the second distractor (D2) on the same side, while condition 3, 4, 7 and 8 have T and D2 on the opposite sides. As a result, there was more interference from the distractor, and thus a relatively lower performance score, during trials in the former condition group than the latter. The behavioral performances between the easy conditions in Space and Talker are comparable.

We correlated subject-specific behavioral RDMs with their corresponding fMRI RDMs and compared the result correlation coefficients with zero as we did in Section 3.3.3. We observed above-chance correlations in the parietal lobe — SPL, IPS and PCUN in both hemispheres (Figure 3.10b). The POS in the left hemisphere also showed significant correlation with the behavioral RDM.

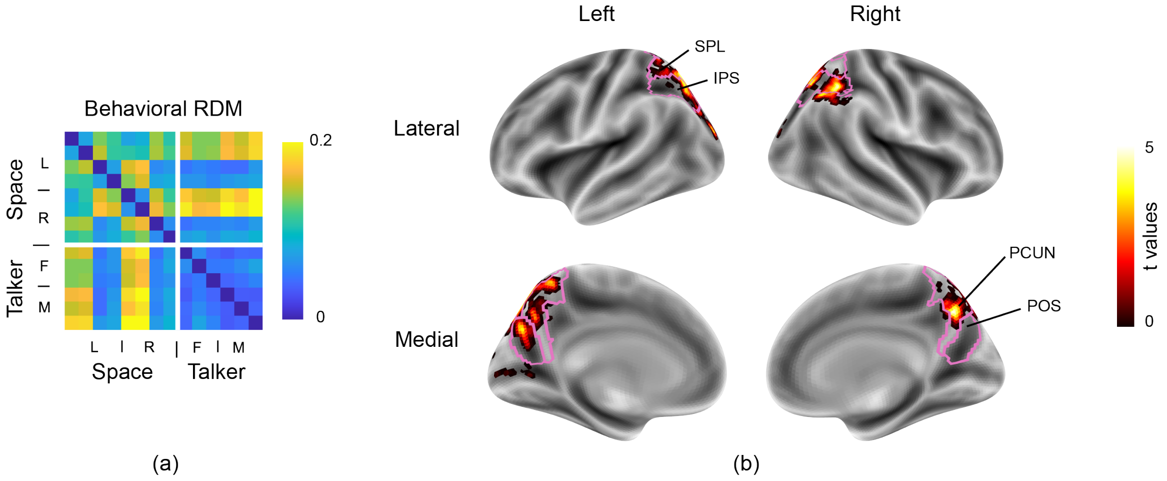


Figure 3.10: (a) The mean behavioral representational dissimilarity matrix (RDM) averaged across all subjects. Element (X,Y) in this RDM shows the absolute difference in behavioral performance (i.e., accuracy of the attention task) between condition X and condition Y. (b) Voxels in the parietal region showed significant correlation between their voxel-wise fMRI RDMs and the subject-specific behavioral RDM. We compared the Fisher-transformed correlation coefficients at each voxel with zero using a one-sample t-test, and rendered the t-statistics of this comparison onto a cortical surface model. The results were masked by their statistical significance derived from the threshold-free cluster enhancement method.
Acronyms: SPL, superior parietal lobule; IPS, intraparietal sulcus; PCUN, precuneus; POS, parietal occipital sulcus.

3.4 Discussion

In this work, we studied the neural activation and representation of auditory selective attention in fMRI. We designed an experiment with multiple conditions that used similar stimuli, but required the listeners to adopt different listening strategies and different cognitive states. We recorded fMRI from 19 subjects when they were performing tasks that required spatial or non-spatial auditory attention. We conducted a GLM analysis to compare the level of activation between attention types, from which we identified a broad brain network that is actively engaged when attention is deployed. ROIs in this network have their own characteristics of specialization in spatial or talker attention: some are more active in one than the other, and some are not. We also studied the neural representation of auditory attention: at each voxel, we estimated the dissimilarity between each pair of conditions, and compared these dissimilarity features with conceptual models or behavioral performance to reveal the cognitive control underlying auditory attention.

3.4.1 Neural representation of brain states

RSA studies neural representations, or how information is encoded in a neural unit, through the study of representational dissimilarity features [9], [11]. It has been previously employed in a variety of topics in neuroscience research, including visual object recognition [18], [42], [47], visual object processing [13], scene perception [48], audiovisual integration [12], and semantic categorization of sound [49]. In these studies, researchers presented various visual or auditory stimuli to subjects and extracted RDM features from their EEG, MEG or fMRI signals to investigate how stimuli of different categories were processed during the experiment, or how they were represented in different brain regions. In other words, these studies applied RSA to decode certain properties of the stimulus. Conversely, our study deployed RSA to track brain states instead of input categories. The auditory stimuli we presented to subjects were either identical or only slightly different between experimental conditions. What did vary, however, was the attentional states of subjects across different tasks, which may be the major property that is being characterized by the RDM features in this study. To our knowledge, only one prior work has attempted to explore the neural representation of brain states. Salmela et al. [43] conducted an RSA study to investigate the dynamics of the fronto-parietal attention network during an audiovisual attention task. They extracted RDM features from EEG time course, and compared them with RDMs calculated from fMRI signals to track the spatiotemporal dynamics of attentional control. One challenge faced in this previous study was the trade-off between a condition-rich experimental design and the sample size for each condition: due to the fact that trials in attention studies are usually long for the need of multiple presentations of stimuli, subject response and feedback to answers, very limited amount of data were collected in this prior work (20 trials per condition for 6 conditions, and 5 trials per condition for the other 12 conditions), which inevitably hindered an accurate estimation of dissimilarities. In our study, we designed an experiment comprising short trials that used fast repetitions of short syllables as stimuli. As a result, we managed to collect 36 trials for each of the 21 conditions, which is theoretically sufficient for the central limit theorem to hold, and therefore offers a good estimate of the between-condition dissimilarity.

3.4.2 Neural activation in auditory attention

The GLM analysis revealed an extended brain network that is actively engaged by auditory attention (Figure 3.6). The fronto-parietal attention networks triggered by Space and Talker are greatly overlapping with each other (Figure 3.6a & b). They both comprise sPreCS, iPreCS, IFS, SFG, MCC, IPS and SPL — a network that has been reported in past works on auditory attention [69], [70]. This result matches with a previous finding in Alho et al. [69], in which the authors discovered that attention to location and attention to pitch, in the auditory domain, activate similar brain regions. This high similarity between Space > Relax and Talker > Relax could further be explained by the identical stimuli being used across Space and Talker conditions: they both created a busy auditory scene by using spatialized syllables spoken by multiple talkers, which enhanced the neural activity in both space- and talker-specialized regions, compared to passive listening.

Each ROI's specialization in spatial or talker attention was revealed in the Space > Talker contrast. We observed stronger activation in sPreCS, iPreCS, SI, SFG, MCC, the posterior part of tpSTG and all ROIs in the parietal and occipital lobe (Figure 3.6c) in Space than in Talker, while the anterior part of laSTG and tpSTG is more active in Talker than in Space. This approximates the classic hierarchical processing model, in which the "dorsal" pathway, comprised of posterior auditory cortex, inferior parietal lobule and premotor cortex, specializes in extracting the spatial components of an auditory stimulus, whereas the "ventral" pathway, comprised of the anterior auditory cortex and the inferior frontal cortex, is specialized for non-spatial information of sound [6], [71]. A similar result was shown in Degerman et al. [72], where the authors compared attention to location and attention to pitch with fMRI, and found stronger activation in the prefrontal (including the sPreCS and iPreCS in this study) and inferior parietal cortical regions during attention to location. A recent study by Michalka et al. [59] identified an interdigitated pattern in the prefrontal cortex, where sPreCS and iPreCS are biased to visual tasks, and IFS is biased to auditory tasks. These visual regions are functionally connected to areas in the visual cortex, and have stronger activation in attention to spatial information than in attention to temporal information of an object, even when this object is purely auditory. IFS, on the contrary, is functionally connected to the auditory cortex. It is recruited during auditory attention tasks, but has no preference between attention to space and attention to timing of events. The exact pattern is observed in this study, with sPreCS and iPreCS being more active in Space than Talker, and IFS being indifferent between the two. This result consolidates the view that there are domain-specific regions in the prefrontal cortex, and these regions can be engaged to process different information dimensions of a sensory input object [73].

3.4.3 Neural representation in auditory attention

The average RDMs shown in Figure 3.7 reveal the information being encoded in each ROI. In all of these ROIs except vPCC, we observed strong encoding of "attention or no-attention" — the Space-Relax and Talker-Relax blocks, in the bottom left and bottom middle portion of the RDM, contain high dissimilarity values, while values in the Relax-Relax block in the bottom right corner are generally low. This result is consistent with our GLM analysis, in which these ROIs are more active in Space or Talker than in Relax. The vPCC is the only exception in this figure that does not encode strong information about attentional state: its RDM pattern is close to random and does not exhibit a distinction between attention and no-attention conditions. In addition, we also observed some micro-structures in these RDMs that encode different aspects of attention. The Space-Space block in left and right CAL shows a clear contrast between attention to left and attention to right; the Talker-Talker block in left and right IFS shows a difference between attention to talkers of different genders. These results demonstrate that the method we used in this study to calculate representational dissimilarities can successfully decode auditory attention, and the RDM features we derived can be used to describe the attentional state of the brain at each voxel.

3.4.4 Correlation with conceptual models

We correlated voxel-wise RDMs with two conceptual model RDMs in Figure 3.5b & 3.5c (i.e., Left vs Right and Female vs Male) to search for voxels that encode information about the attended location, or the gender of the attended talker. The correlation with the Left vs Right model revealed a cluster of ROIs in the occipital lobe that show significant encoding of the spatial information of attention (Figure 3.8a). Our visual system excels at encoding spatial information, with over 20 cortical areas that show visuospatial maps [59], [74]–[76]. Moreover, it also offers cross-sensory flexibility; portions of the visual system can be recruited for an auditory task when space is the primary information of interest [59]. Therefore, we speculate that the observed high correlation with the Left vs Right model may indicate these ROIs' active role in mapping the target sound in space. Further investigation on the dynamics of these ROIs is warranted.

We also correlated voxel-wise RDMs with the Female vs Male model to identify brain regions that are encoding information about the gender / pitch of the target talker. The right IFS is the only ROI that survived the significance test (Figure 3.9a). Previous studies have found that IFS, as an ROI along the "ventral" pathway, specializes in processing non-spatial information about the sound [6] — it has enhanced activity during tasks like attention to pitch [58] or attention to the timing of events [59]. In this study, IFS may have been recruited to process the acoustic features of the target talker's voice (e.g., pitch, timbre, etc.). However, a study of the percent signal change in this region revealed no significant difference between attention to a female talker and attention to a male talker (Figure 3.9b). This might happen due to the individual differences in how acoustic information is encoded in this ROI. For example, some participants may encode a female's voice in a way that will trigger greater response than a male's voice (e.g., if they perceive a female's voice as more salient and distinct than a male's voice), while the others may do the opposite. If this happened, the variability across subjects in this gender-specific signal enhancement would lead to an insignificant signal difference on the group level. Another possible cause of the observed result is the voxel-by-voxel variability. Some voxels within this ROI may have greater activation in one condition group than the other, while the other voxels may show the opposite pattern; calculating the average activation within an ROI ignores this variability and loses information. RSA, however, can resolve these problems, because any difference, regardless of its sign, is treated as information in an RSA — the aforementioned differences across subjects or across voxels will not cancel out each other under the RSA framework. This makes RSA a powerful tool with great sensitivity to study cognitive functions. We will compare RSA with conventional fMRI analysis in more detail in Section 3.4.6 for its advantages and disadvantages.

3.4.5 Comparison to behavioral performance

We constructed a behavioral RDM for each subject and correlated it with their individual fRMI RDMs at each voxel. This analysis revealed a few ROIs in the parietal region, including SPL, IPS, PCUN and POS, whose RDMs correlated above-chance with the behavioral RDM (Figure 3.10b). Previous studies demonstrated that auditory tasks recruit both visuotopic (i.e., IPS) and non-visuotopic parietal regions, and these regions are more active in spatial tasks than in non-spatial tasks [77]–[79]. Moreover, Michalka et al. [77] showed that some sub-regions in IPS and SPL can be flexibly recruited under high auditory spatial demands. In this study, the behavioral RDM is comprised of conditions for Space and Talker attention. The most distinctive feature of this RDM is the difference between hard spatial attention conditions and the other conditions — there is a checkerboard pattern in the Space-Space block, and a striped pattern in the Space-Talker blocks. We speculate that the observed significant correlation between parietal RDMs and the behavioral RDM is mainly caused by the performance difference between hard spatial conditions and the other conditions, and it may show an important role of these parietal ROIs in spatially demanding tasks, in which they are more actively recruited to distinguish a target from an adjacent distractor.

3.4.6 Compare RSA to conventional fMRI analysis

RSA is a powerful tool to study neural correlates with a condition-rich experimental design. It is different from conventional fMRI studies, such as GLM analysis, where we directly compare the signals between two conditions or condition groups and examine their contrast. Instead, RSA operates on the relationship among multiple conditions — it quantifies the dissimilarity between each pair of conditions, and uses these dissimilarity features as an abstraction to study brain state. In contrast to a conventional analysis, RSA omits the sign of difference when accumulating evidence across samples. This is particularly powerful when there are individual differences in the effect under test. For example, in an auditory salience study, one subject may find certain sounds more salient than the others, while another subject may feel the opposite. If the question being asked is which brain region encodes the perceived salience of sound, data from these two subjects will boost the observed effect under an RSA framework, but will neutralize each other in a conventional analysis. In this study, we also demonstrated another example in Figure 3.9, where there is no significant difference in signal change on group level, but the RSA approach revealed significant effects in right IFS. Thus, RSA could be more sensitive than conventional methods in certain cases.

The high sensitivity of RSA comes at a cost. For example, in this study, ignoring the sign of difference introduces confusion to distinguishing an area in the attention network from an area in the default mode network. For example, if Relax conditions are included as part of a conceptual model (different from the two models we showed in Figure 3.5b, where Relax conditions were not used), one RDM in the attention network would appear similar to another one in the default mode network, due to the their similar "dissimilarity values" between Space or Talker and Relax, despite the fact that they behave in an opposite manner. One direct consequence of this confusion is a possible misinterpretation of the inferences drawn by a cluster-based statistical analysis. Cluster-based statistics, like the TFCE method used in this study, give more power to effects that are continuous in space and enhance the TFCE score of one voxel with contribution from all other voxels within the same spatially connected cluster. Under this framework, if Relax is part of the conceptual model, the statistical power of a voxel in the attention network would be wrongly enhanced by a neighbouring voxel in the default mode network. Therefore, RSA should be used with caution when there is a "null" condition or condition group (e.g., the Relax in this study) in the experimental design to avoid possible misinterpretation of the results.

3.4.7 Limitations

One limitation of this study is the relatively small number of conditions in our experimental design compared to previous RSA studies [18], [42]. A greater number of conditions can effectively reduce the chance of having false discoveries. We tried to control the false discovery rate by adopting a non-parametric statistical method (i.e., TFCE), which does not make assumptions about noise distribution and makes statistical inferences on a cluster level. We also recruited a relatively large number of participants (n = 19) for this type of study, for a greater statistical power and a less biased estimation of population means. However, a greater number of conditions and trials per condition are always preferred in RSA studies. Our design was constrained by the use of an attention task that requires subject response, which is not necessary for most previous RSA studies, since they focused on decoding stimuli rather than internal processing states. Future studies should consider designing experiments in a way that data about multiple trials and/or conditions could be acquired per subject response. This will effectively reduce the time to acquire a dataset that is ideal in size (i.e., number of conditions and trials per condition) for an RSA study.

3.5 Conclusions

This paper is among the very few pioneering studies that adopted a representational similarity analysis framework to investigate the neural representation of attentional states. We designed a condition-rich experiment and recorded fMRI data while listeners performed in an auditory attention task. We identified an extended attention network, in which individual brain regions show different specialization in spatial or non-spatial attention. We also extracted representational dissimilarity features from each voxel, and compared these features with ideal conceptual models or behavioral performance. We identified the medial occipital lobe as the region actively encoding the spatial information about auditory attention; the right IFS is the sole region that encodes information about the gender / pitch of the attended talker. The neural representation of the parietal regions are correlated with the behavioral performance, demonstrating their important role in spatially demanding tasks.

Chapter 4

Information flow in auditory selective attention: a representational similarity analysis for EEG-fMRI fusion

4.1 Introduction

In **Chapter 2** and **Chapter 3**, we studied the neural representation of auditory attention in electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), respectively. In EEG, we extracted a representation feature called a representational dissimilarity matrix (RDM) from both the EEG time course and its alpha oscillation power at each time point. In fMRI, we extracted the same type of RDM feature at each voxel. These RDM features summarize the differences between each pair of experimental conditions, and thus can unveil what information is encoded at each time point (as in EEG) or at each brain location (as in fMRI). We studied the patterns in these EEG or fMRI RDMs through their correlation with conceptual model RDMs, and identified time intervals and brain regions that encode information pertinent to spatial or non-spatial auditory attentional controls.

This study of neural representations can be extended to a multimodal data fusion analysis, since we already collected both EEG and fMRI data separately, but with the same condition-rich experimental design. The RDMs of EEG and fMRI have the same dimension, scale and biological meaning (i.e., how information is encoded), and therefore can be used to estimate the correspondence of information between these two imaging modalities. One intuitive approach is to correlate the EEG RDM at each time point with the fMRI RDM at each voxel. This process yields a 4-dimensional array of correlation coefficients, of which each element represents the commonality between EEG and fMRI representations at a certain time and at a certain brain location. This method, known as representational similarity analysis (RSA), was proposed and formulated by Kriegeskorte et al. [11]. It is immensely useful for multimodal neuroimaging fusion, because signals like EEG and fMRI have different dynamics, scales, noise levels, etc. Thus, it is difficult, if not impossible, to find a direct correspondence between these two modalities. With RSA, however, we can collect EEG and fMRI data with the same experimental design, calculate dissimilarity features from EEG at each time point and from fMRI at each voxel, and conduct a time-by-location correlation analysis to search for when and where EEG and fMRI share common information. This technique has already been applied to study the cognitive functions underlying visual object recognition. Cichy et al. [18] designed an experiment in which they presented participants with images of objects. They extracted neural representations at each time point from magnetoencephalography (MEG), and at each voxel from functional magnetic resonance imaging (fMRI). Then, they searched for time instances and voxels that shared similar patterns in their neural representation, which depicted a spatiotemporally resolved information flow during visual object recognition.

In this study, we aim to adopt the RSA framework to explore the dynamics of cognitive control in specific brain regions. Previous works on spatial and non-spatial auditory attention revealed that superior precentral sulcus (sPreCS), a region along the “dorsal” pathway [6], is actively recruited in spatial attention [59], while the inferior frontal sulcus (IFS), a region along the “ventral” pathway [6], engages in attention to the non-spatial aspects (e.g., pitch, timing of events, etc.) of sound [59]. In the parietal lobe, the intraparietal sulcus (IPS) has been identified as critical in spatial mapping of sound, and is more active when the task is more spatially demanding [70], [77]. The information dynamics of these regions (i.e., iPreCS, IFS, IPS), together with regions in the auditory (superior temporal gyrus, STG) and visual cortex (calcarine sulcus, CAL), will be studied via the RSA approach.

4.2 Materials and methods

4.2.1 Participants

Data from the 19 adults (19 – 30 years old, 8 women) who participated in our fMRI study (**Chapter 3**) were used here. These subjects also belong to the cohort who participated in our EEG study (**Chapter 2**). Therefore, both EEG and fMRI data are available from these participants.

4.2.2 Stimulus and experiment

Details about the stimulus and the experimental design can be found in Section 2.2.2 and Section 2.2.3. In brief, short (500 ms for each), human-voiced (two female, two male, and one synthesized gender-neutral talker) syllables (/ba/, /da/, /ga/) were used as the stimuli. These syllables were spatialized (90° or 30° to the left or right) using a set of generic head-related transfer functions to create an acoustically and spatially demanding auditory scene. The listeners were cued to pay attention to a particular direction or to a particular talker when a 4-syllable mixture was played. Then, they were asked to identify what syllable came from the target.

We designed an experiment with 21 conditions. These conditions differ by the required type of attention (i.e., spatial, talker, or no), location of the target (left or right), gender of the target talker (female or male), and whether or not an extra difference in location or gender of talker exists between the target and a masker. Details about the experimental design is shown in Figure 2.1.

4.2.3 EEG preprocessing and analysis

We discussed how we extracted RDM features from EEG signals in Section 2.2. In brief, EEG signals were first preprocessed with bandpass filtering, downsampling, and independent component analysis (ICA). Then, we conducted time-frequency decomposition on the EEG signals using a continuous wavelet transform (CWT), and estimated the alpha oscillation power by averaging the magnitude squared of CWT coefficients from 8 to 14 Hz. Next, for each subject and each time point in data, we trained a linear support vector machine (SVM) for each pair of conditions using multivariate features comprised of EEG time course or alpha power information across all electrodes. The average decoding accuracy, estimated from leave-one-trial-out cross-validation, was used to denote the dissimilarity between each pair of conditions. These pair-wise dissimilarity values were stored in a matrix called the representational dissimilarity matrix (RDM), which is the neural representation feature we extracted from the EEG signals. From this step, we yielded two sets of EEG RDMs — one for EEG time course, and one for alpha power. Each can be expressed as a function of time, meaning that there is one RDM at each time point. Since we only observed weak, transient effects with EEG time course RDMs in Section 2.3.2, only alpha RDMs were used in this study.

4.2.4 fMRI preprocessing and analysis

We discussed how we extracted RDM features from fMRI signals in Section 3.2.4. In brief, we adopted the preprocessing pipeline recommended by SPM12 [64] — realignment, coregistration, segmentation and normalization. After preprocessing, we first conducted a general linear model (GLM) analysis to calculate the contrast between each condition and an explicitly modeled resting state. Then, at each voxel in each contrast map, we extracted a multivariate feature using a searchlight method, and used this feature to estimate the dissimilarity between each pair of conditions. Same as in the EEG analysis, we used an RDM to summarize the pair-wise differences. From this step, we yielded a set of fMRI RDMs with one RDM at each voxel. In this study, since we are interested in the dynamics of a few selected ROIs (sPreCS, IFS, IPS, laSTG and CAL), we averaged the voxel-level RDMs across all voxels within each of these ROIs (as defined in Destrieux et al. [68]) to yield ROI-level RDMs.

4.2.5 EEG-fMRI fusion

With the EEG alpha RDMs as a function of time, and the ROI-level fMRI RDMs of five ROIs (sPreCS, IFS, IPS, laSTG and CAL), we conducted an EEG-fMRI fusion analysis. For each ROI, we correlated its ROI-level RDM with the EEG RDMs at each time points 4.1. This yielded a time series of correlation coefficients (fusion correlations), which reveals how well EEG and fMRI correspond with each other at a specific time point and a specific brain region. A link between these two modalities can be established if their correlation is significantly above chance.

Statistical inferences were made with threshold-free cluster enhancement (TFCE), the same method as the one discussed in Section 3.2.5. In brief, the statistical power at one point can be enhanced by the other points in the same cluster. Thus, effects with more continuity in the feature space (e.g., time, frequency, space, etc.) are more likely to be significant. We randomly shuffled data label 10,000 times, and used the maximum TFCE score of each permutation to form a null distribution. We set the 95th percentile of this null distribution as the threshold to identify significant effects.

In this study, we applied TFCE to fusion correlations of a few selected ROIs to identify time intervals during which an fMRI RDM correlates with EEG RDMs significantly above chance. We also applied the same TFCE method to examine whether the correlation traces

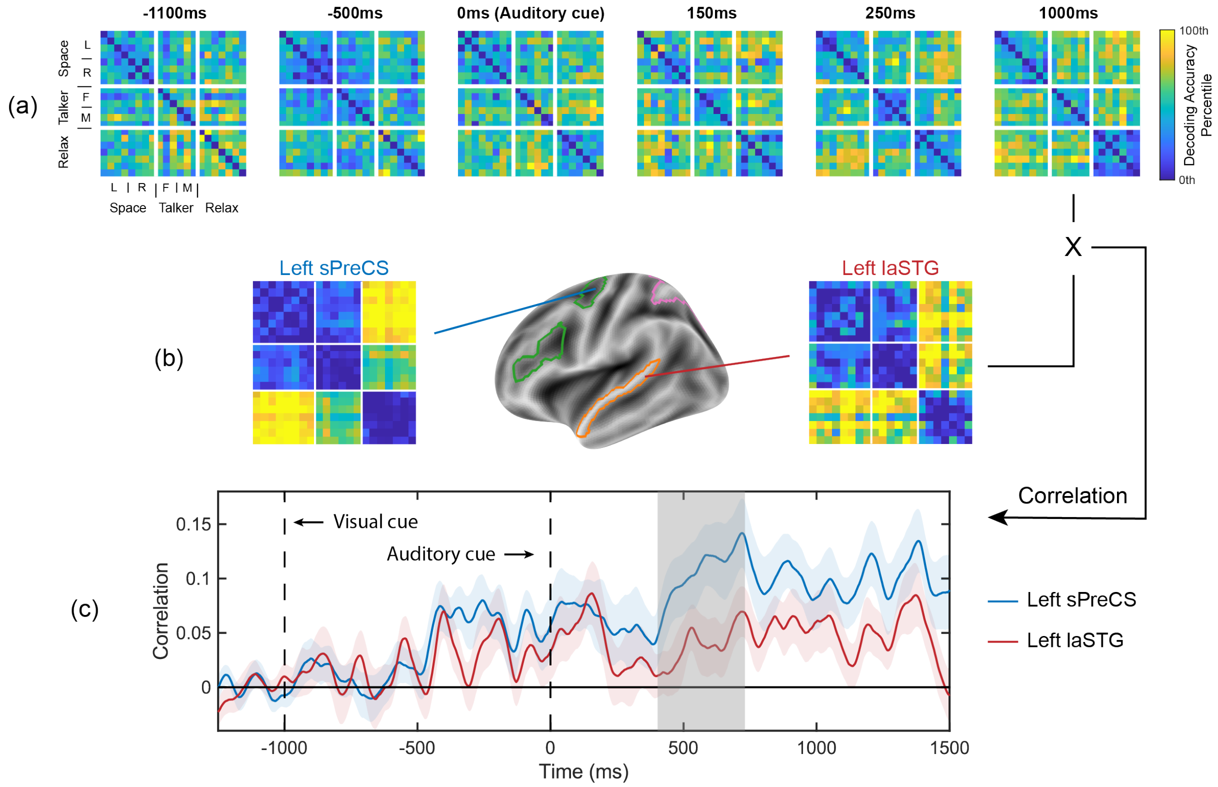


Figure 4.1: (a) EEG RDMs calculated from alpha power at several timepoints in the cue period (b) Average RDMs of left sPreCS and left laSTG. These two RDMs were correlated with EEG RDMs at each time point to yield the two time series of correlation coefficients in (c). The blue and red shaded area represent standard error. The grey shaded area shows the interval when the two traces are significantly different from each other (398 – 684 ms, p < 0.05). Notice the divergence of the two traces after the onset of auditory cue.
Acronyms: sPreCS, superior precentral sulcus; laSTG, the lateral aspect of superior temporal gyrus.

of selected pairs of regions are significantly different during some intervals. The correlation coefficients were passed through Fisher's z-transform before being examined by TFCE.

4.3 Results and discussion

In Figure 4.1c, we compare the fusion correlation results between left sPreCS and left laSTG. These two regions have similar correlation in the cue period before an auditory cue is presented. The correlation for sPreCS increases at around 400 ms after the onset of auditory cue, and stays significantly greater than that for laSTG for approximately 300 ms (398 – 684 ms, $p < 0.05$). This result may reflect the roles that these two regions play in different types of auditory attention. sPreCS has been reported to be actively engaged during covert visual attention to location [80], [81], and it can be recruited to complete auditory spatial attention tasks [59]. The GLM analysis results in Section 3.3.1 also show that sPreCS has greater activation in Space than in Talker attention conditions. Thus, sPreCS might be more specialized in spatial than in non-spatial attention, and the increase in its fusion correlation after the onset of auditory cue may indicate that the information about "spatial vs. non-spatial attention" is encoded in this region during this period of time. Unlike sPreCS, laSTG has been reported to play a hybrid role in handling spatial and non-spatial information; lesion studies have revealed its contribution to both spatial [82], [83] and temporal [84] perception. Therefore, it may not encode much information specific to attention types (i.e, spatial or non-spatial) in the cue period, which might lead to the observed relatively lower fusion correlation than sPreCS during this window.

Figure 4.2 compares the fusion correlation results between left IFS and left CAL in the cue period. From the GLM analysis results in Section 3.3.1, we learned that CAL does not show significant difference in activation between active attention and no attention conditions (Figure 3.6). However, it encodes strong information about the direction of attention (Figure 3.8). Thus, we hypothesized that CAL is recruited when there is cognitive need for spatial mapping of a sound object, but is not always active during the attention task. IFS, on the other hand, is strongly engaged in both spatial and talker attention (Figure 3.6). It may serve a more active role than CAL when spatial mapping of sound is not the primary cognitive task.

In Figure 4.2, we observe that the fusion correlation for left IFS is significantly above zero in a few short intervals before and right after the onset of auditory cue. It is also consistently above zero from around 800 ms till the end of the cue period. CAL shows no significance during this period. A direct comparison between these two traces clearly differentiates the role played by these two ROIs, in a way that matches our expectation.

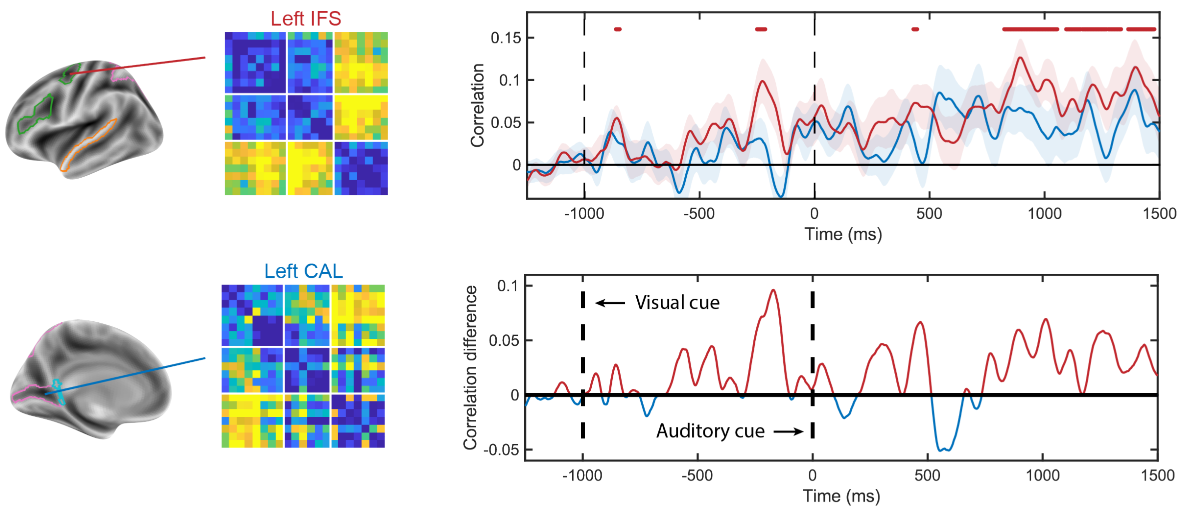


Figure 4.2: Comparison of fusion correlation between Left IFS and Left CAL in the cue period. Correlation is calculated using the full RDM. Top panel shows the correlation traces of two ROIs (red: IFS; blue: CAL). Shaded area represents standard error. Bars on top show intervals in which the corresponding correlation trace is significantly above zero (p<0.05). Bottom row shows the difference between these two traces.
Acronyms: IFS, inferior frontal sulcus; CAL, calcarine sulcus.

In Figure 4.3, we compared the fusion correlation results between left sPreCS and left IPS in the cue period. The fusion correlation traces for these two ROIs are mostly in line with each other. One interesting observation is that the neural process in sPreCS seems to happen earlier than that in IPS, indicated by the significance bars on the top. There is also a peak in the fusion correlation for sPreCS at around 700 ms, which is ahead of the peak for IPS at around 900 ms. These findings suggest a flow of information between these two regions.

In Figure 4.4, we compared the fusion correlation results between right CAL and right sPreCS in the cue period. Unlike in the previous two comparisons, here we used a portion of the RDM, instead of the full RDM, for this analysis — only Space and Talker conditions were counted for calculating the correlations. In this way, we can exclude the effects of Relax conditions, and focus on the question: where and when does the brain encode information

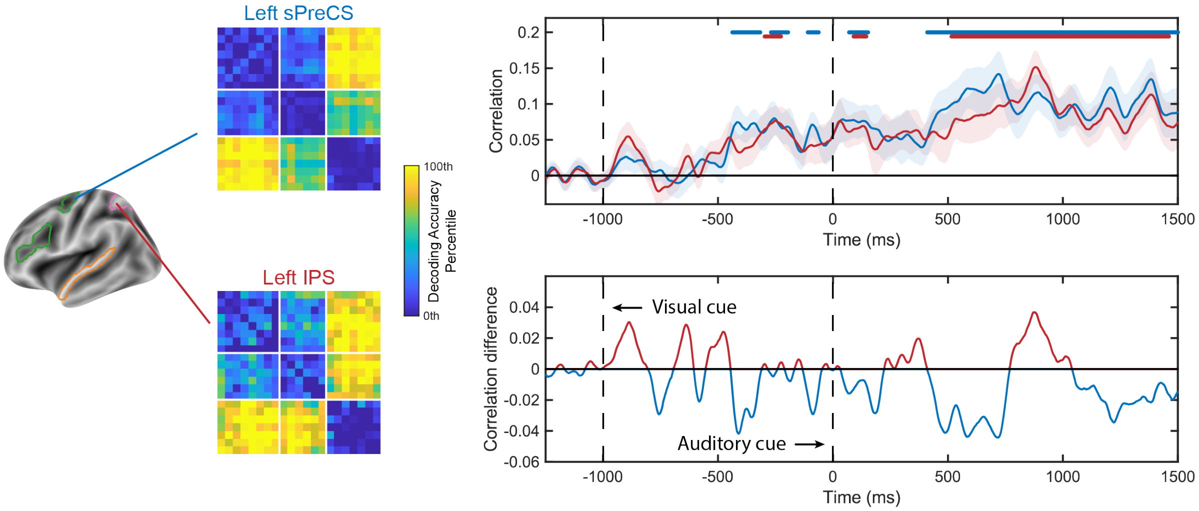


Figure 4.3: Comparison of fusion correlation between Left sPreCS and Left IPS in the cue period. Correlation is calculated using the full RDM. Top panel shows the correlation traces of two ROIs. Shaded area represents standard error. Bars on top shows intervals in which the corresponding correlation trace is significantly above zero ($p<0.05$). Bottom row show the difference between these two traces.
Acronyms: sPreCS, superior precentral sulcus; IPS, intraparietal sulcus.

about the difference between spatial and non-spatial attention. We observe that, even though sPreCS shows significantly stronger activation in Space than in Talker (Figure 3.6c), its encoding of the difference between Space and Talker does not happen a lot during the cue period. Its fusion correlation is significantly above zero only for a very brief period of time at around 520 ms. Different from sPreCS, the fusion correlation for CAL exhibits two salient peaks at around 520 ms and 800 ms. Given that CAL encodes information about the direction of attention, these two peaks may have indicated that CAL is recruited during these two windows for mapping the auditory cue in space.

In Figure 4.5, we compared the fusion correlation results between left IFS and left sPreCS in the stimulus period. Again, only Space and Talker attention are used for this analysis. These two ROIs are claimed to have different specializations: IFS for non-spatial attention, and sPreCS for spatial attention. The fusion results show that they both encode the difference between Space and Talker, and they seem to have different dynamics in terms of

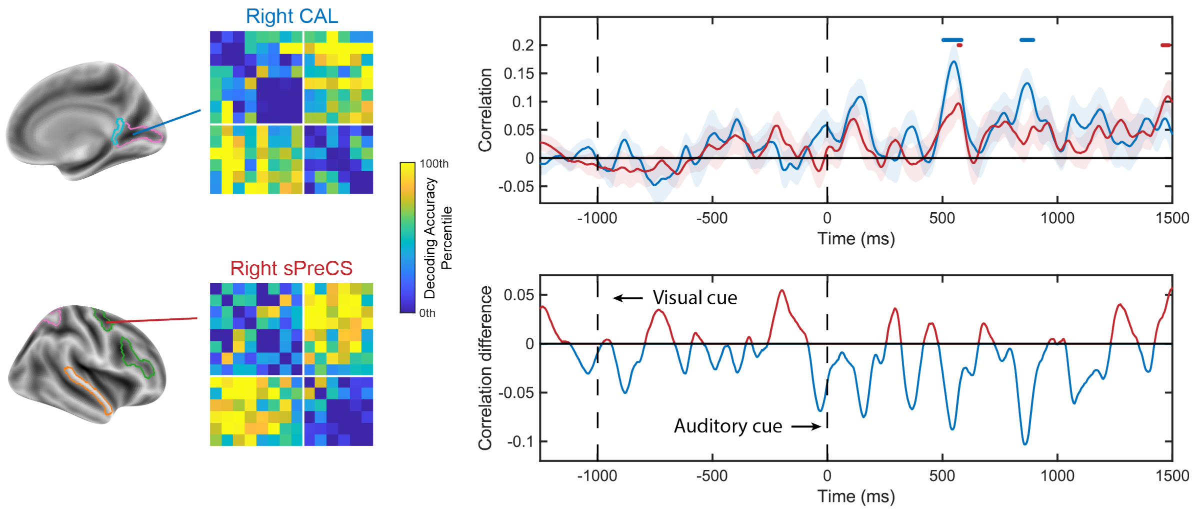


Figure 4.4: Comparison of fusion correlation between right CAL and right sPreCS in the cue period. Correlation is calculated using the Space and Talker conditions only (i.e., excluding the Relax conditions). Top panel shows the correlation traces of two ROIs. Shaded area represents standard error. Bars on top shows intervals in which the corresponding correlation trace is significantly above zero (p<0.05). Bottom row shows the difference between these two traces.
Acronyms: sPreCS, superior precentral sulcus; CAL, calcarine sulcus.

when this information is being encoded. In IFS, this process happens exactly around when the target is being played, while in sPreCS, this process seems to be on throughout the stimulus period, and becomes more significant 150 ms after the target onset. This may have suggested a difference in the neural mechanism behind spatial and non-spatial attention. In spatial attention, the listeners focus on one side of the auditory scene through suppressing the perception of the unattended side, and this suppression mechanism is reflected in alpha oscillations [85]. The talker attention tasks in this study, however, require the listeners to "let in" all the auditory input, and match the perceived sound with a template in their working memory. These two different attention strategies may have caused the observed difference in dynamics between IFS and sPreCS.

In Figure 4.6, we compared the fusion correlation results between left sPreCS and left CAL in the stimulus period. Only the top left portion of the RDM (i.e., the Space-Space cell of the RDM) was used to calculate correlation. Results show that the fusion correlation

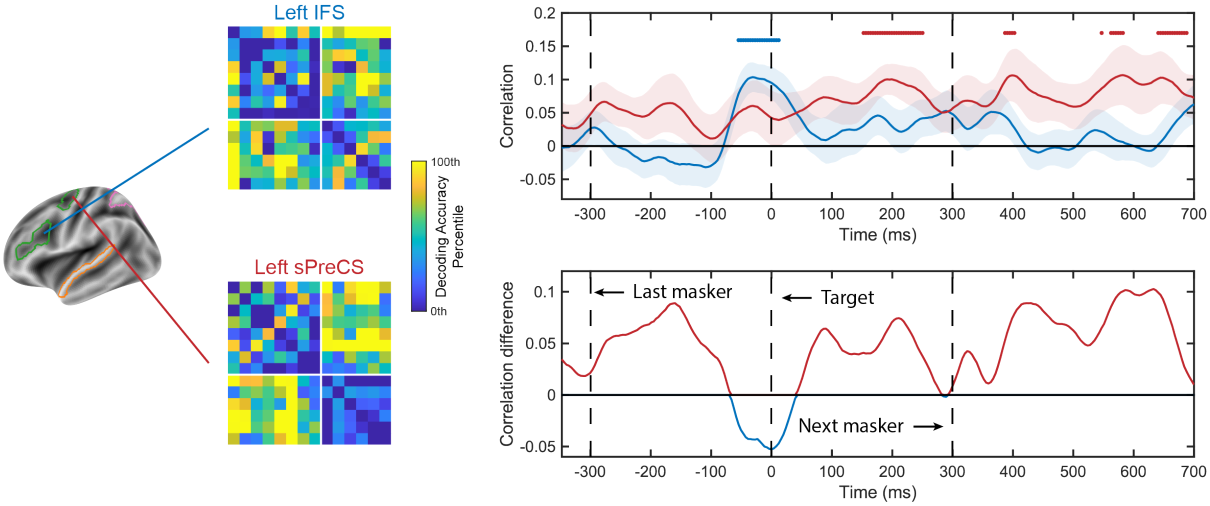


Figure 4.5: Comparison of fusion correlation between left IFS and left sPreCS in the stimulus period. Correlation is calculated using the Space and Talker conditions only (i.e., excluding the Relax conditions). Top panel shows the correlation traces of two ROIs. Shaded area represents standard error. Bars on top shows intervals in which the corresponding correlation trace is significantly above zero ($p<0.05$). Bottom row shows the difference between these two traces.
Acronyms: sPreCS, superior precentral sulcus; IFS, inferior frontal sulcus.

for CAL is significantly above chance from around 100 ms to 200 ms after the target onset. This confirms our previous claim that CAL is only recruited when spatial mapping of sound is needed, which is consistent with findings in previous studies on auditory spatial attention in congenitally blind humans [86], [87].

One limitation of this study is the relatively small number of conditions in our experimental design compared to previous RSA studies [18], [42]. A greater number of conditions can effectively reduce the chance of having false discoveries, because randomness in correlation decreases as the number of conditions increases. With a small number of conditions (for example, $n = 6$), even RDMs filled with random numbers may yield a correlation greater than 0.5 with around 5% of chance (Figure 4.7). In our study, as we used smaller and smaller portions of an RDM for fusion analysis, the variance (or noise) in the correlation results dramatically increased. We even tried to use only the Talker-Talker portion (i.e., the cell in the middle of an RDM, $n = 6$) to calculate correlation, but no effect survived the TFCE test

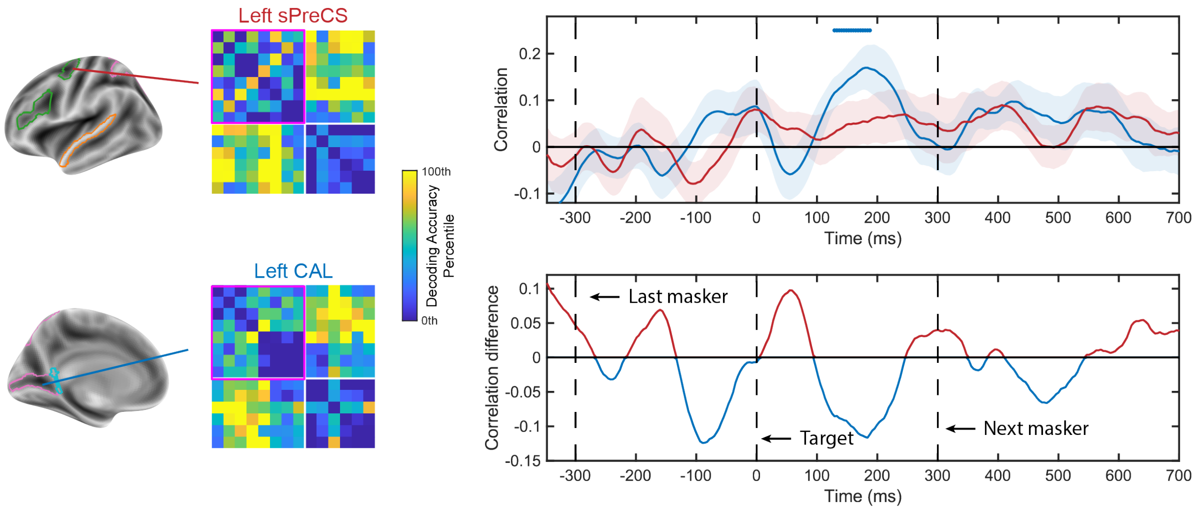


Figure 4.6: Comparison of fusion correlation between left sPreCS and left CAL in the stimulus period. Correlation is calculated using the top left portion of the RDM labelled with pink (i.e., the Space-Space cell of the RDM). Top panel shows the correlation traces of two ROIs. Shaded area represents standard error. Bars on top shows intervals in which the corresponding correlation trace is significantly above zero (p<0.05). Bottom row shows the difference between these two traces.
Acronyms: sPreCS, superior precentral sulcus; CAL, calcarine sulcus.

(which, is an effective tool for controlling false discovery rate, because it estimates the null distribution using a permuted version of the original data, and thus takes the randomness in correlation into account). We also recruited a relatively large number of participants (n = 19) for this type of study, for a greater statistical power and a less biased estimation of population means. However, a greater number of conditions and trials per condition are always preferred in RSA studies. Our design was constrained by the use of an attention task that requires subject response, which is not necessary for most previous RSA studies, since they focused on decoding stimulus rather than differences in the internal processing state across different listenting conditions. Future studies should consider designing experiments in a way that data of multiple trials and/or conditions could be acquired per subject response. This will effectively reduce the time to acquire a dataset that is ideal in size (i.e., number of conditions and trials per condition) for an RSA study.

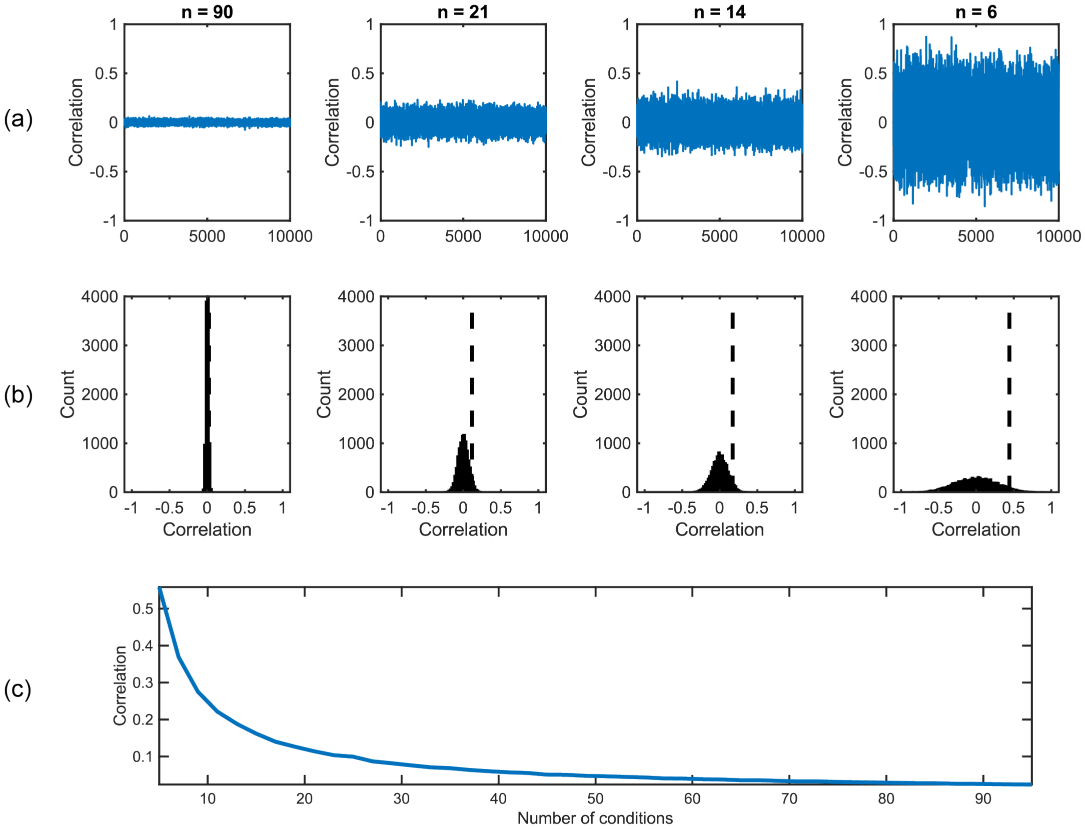


Figure 4.7: Randomness in correlation decreases as the number of conditions increases. (a) Simulation of correlation between random RDMs with different number of conditions. 10,000 pairs of RDMs filled with random numbers are generated for each condition number. The distributions of the result correlations are shown in (b). Dashed lines denote the 95th percentile of the respective distribution. (c) The 95th percentile of the distribution of random correlations as a function of condition number.

4.4 Conclusion

This work is among the first few studies that deploy a representational similarity analysis (RSA) for multimodal data fusion to study the dynamics of auditory attentional control. We designed a condition-rich experiment, which required spatial or non-spatial auditory attention from the listeners. We collected EEG and fMRI data separately with the same experimental design, and extracted neural representation features from both modalities. Then we correlated these representation features to search for significant information correspondence in time and space. The fusion analysis revealed that the calcarine sulcus is only active during the task when spatial mapping of sound is needed, suggesting its major role in processing spatialized auditory targets. We observed a difference in dynamics between the inferior frontal sulcus and the superior precentral sulcus during the stimulus period, which might have reflected a difference in the suppression mechanism between spatial and non-spatial attention.

Chapter 5

Attention-based auditory brain-computer interfaces

5.1 Overview of chapter

In **Chapter 2**, I studied the neural representation of auditory attention in EEG signals. I estimated the dissimilarity between each pair of conditions at each time point via a machine learning classification approach — multivariate feature vectors of each condition, derived from EEG time course or its alpha oscillation power, were used to train and test a binary linear support vector machine (SVM). The average classification accuracy of the SVM, estimated from a leave-one-trial-out cross-validation, was used to quantify the difference between each pair of conditions. Results in this study showed that the attentional state of the listener could be well decoded from EEG signals, even with data from only one single time point. This motivated me to think whether, with the help of more data and more sophisticated algorithms, we can decode the attentional state of a listener from single-trial EEG data, and achieve a classification accuracy that is sufficiently good for a practical brain-computer interface (BCI).

In this chapter, I present four different studies in which I explored the feasibility of decoding attention from EEG signals for the design of an attention-based BCI system. For the study in **Section 5.2**, I use the data collected in Chapter 2 and decode attentional efforts with a linear SVM from the EEG time course and its alpha oscillation power, an approach similar to the decoding analysis I discussed in Chapter 2. In **Section 5.3**, I use the same dataset, but decode attention with a convolutional neural network, which can automatically learn the feature to be used for classification [88]. Section 5.4 and Section 5.5 are two research projects I designed and conducted as a Research Intern in the Audio and Acoustics Research Group at Microsoft Research during the summers of 2019 and 2020. These works are included in this dissertation with permission from the leader of the research group, Dr. Ivan Tashev. In **Section 5.4**, I design an experiment with streams of auditory and tactile stimuli, and instructed the participants to direct their auditory, tactile or multi-sensory attention to one particular stream. Then, I extracted time-frequency features from their EEG signals, and decoded their attentional state during this task. In **Section 5.5**, I design an experiment with polyphonic music, ask the listeners to attend to one instrument, and decode attention from their envelop following response in EEG signals. Together, these studies demonstrate means to improve an auditory BCI design with better accuracy, efficiency and user-friendliness.

5.2 Decoding auditory attention from single-trial EEG for a high-efficiency brain-computer interface[1]

5.2.1 Abstract

Brain-computer interface (BCI) systems enable humans to communicate with a machine in a non-verbal and covert way. Many past BCI designs used visual stimuli, due to the robustness of neural signatures evoked by visual input. However, these BCI systems can only be used when visual attention is available. This study proposes a new BCI design using auditory stimuli, decoding spatial attention from electroencephalography (EEG). Results show that this new approach can decode attention with a high accuracy (>75%) and has a high information transfer rate (>10 bits/min) compared to other auditory BCI systems. It also has the potential to allow decoding that does not depend on subject-specific training.

5.2.2 Introduction

Electroencephalography (EEG) offers a noninvasive and portable method for monitoring brain activity, making it a popular technology for brain-computer interfaces (BCIs) [26]. Many successful BCI systems use visual stimuli as the sensory input, and decode a user's attention from neural signatures such as event-related potentials (ERPs). These visual paradigms efficiently transmit information to a computer, as quantified by their information transfer rate (ITR). For example, one recent study on visual ERP-based BCI reported an average ITR as high as 20.26 bits/min [33].

Though visual BCI systems are efficient, they cannot be used in scenarios where visual attention is already engaged by real world demands (e.g. walking or driving), or by users with visual impairment. Some previous studies developed auditory BCI systems to tackle

[1] This section is adapted with permission from paper: Winko W. An, Alexander Pei, Abigail Noyce, Barbara Shinn-Cunningham, "Decoding auditory attention from single-trial EEG for a high-efficiency brain-computer interface", in *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2020*

these problems. For example, Kim et al. [28] used multiple streams of spatialized modulated signal with a constant-frequency carrier as the stimuli, and decoded attention from auditory steady-state response (ASSR). An et al. [89] used more user-friendly stimuli, synthesized melodies, and developed a novel BCI paradigm. In another study to reduce user fatigue, Huang et al. [30] proposed using drip drop sounds as the input. However, the efficiency of these auditory systems is substantially lower than most visual-based BCIs. For example, ITRs of the three aforementioned studies were all below 3 bits/min, making them less useful in real applications.

The current study proposed an auditory BCI system with high efficiency. It used spatialized human-voiced syllables as the stimuli, and decoded selective attention from EEG. Inspired by previous studies on auditory attention [36], [38], [39], [90], both ERP and EEG spectrogram were used as features to train and test a series of linear classifiers for attention decoding.

5.2.3 Methods

5.2.3.1 Participants

Thirty adults (19 – 44 years old, 14 female) participated in this study. No participant reported hearing loss or any history of neurological disorders. The Institutional Review Board of Boston University approved this study. All participants gave written informed consent, and were paid for taking part in the study.

5.2.3.2 Experiment

Before the experiment started, participants were asked to sit comfortably in a soundproof booth in front of a computer monitor. The syllables /ba/, /da/ and /ga/, spoken by native English talkers, were used as stimuli. The syllables were spatialized by a set of generic head-related transfer functions (Media Lab, MIT), and played through a pair of insert earphones (ER1, Etymotic Research). The intensity of sound was adjusted to a comfortable listening level for each individual to ensure syllable intelligibility. The simulated locations of these syllables were 90° from the left (L90), center, or 90° from the right (R90, Figure 5.1a), in the horizontal plane.

At the beginning of each trial, a visual cue (VC) was shown on the screen for one second, which could be one of the two words: "Space" or "Relax" (Figure 2.1b). "Space" indicated that participants should direct spatial attention in the upcoming trial, while "Relax" represented a control trial where no attention would be required. An auditory cue (AC) — a spatialized /a/ sound — was given after the VC to direct the participant's attention. In "Space" attention trials, the AC specified the target location (either L90 or R90). In no-attention "Relax" trials, the AC always came from the center (i.e. a neutral value). After a 1000 ms silent period, a 4-syllable mixture was played. All syllables were 600 ms long, and their onsets were separated by 300 ms. In "Space" attention trials, the first and the last syllables were always distractors (D1) played from the center. Of the second and third syllables, one was the target (T), which came from the AC location. The other syllable was the second distractor (D2) that came from the opposite side. Syllables /ba/, /da/, and /ga/ were randomly permuted among T, D1 and D2. The task was to ignore D1 and D2, and to identify T using the keyboard ("1" for /ba/, "2" for /da/, and "3" for /ga/). Visual feedback was given after each response to show whether the answer was correct. In no-attention trials, the participants were asked to ignore all syllables, and give a random answer at the end. The inter-trial interval was set to be 2 seconds with jitter.

Each participant completed 756 trials in total. These trials differed in the locations and talkers of the spoken syllables, and in the type of attention (spatial or non-spatial) required. Data from only four conditions (72 trials in each condition) are presented in this analysis: 1) selective spatial attention trials in which the four syllables occur in location order Center-Left-Right-Center (Spa_LR); 2) selective spatial attention trials with syllables in order Center-Right-Left-Center (Spa_RL); 3) a control no-attention condition with syllables in order Center-Left-Right-Center (Ctr_LR); 4) a control no-attention condition with syllables

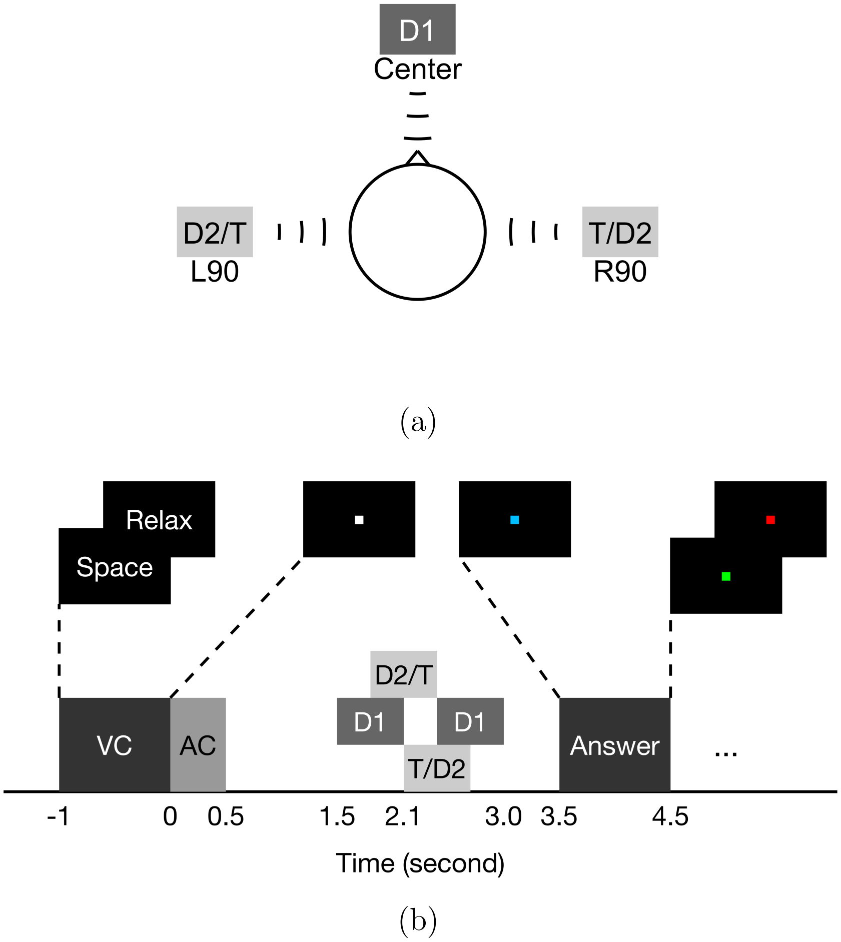


Figure 5.1: (a) Spoken syllables were spatialized to center and to 90° left (L90) or right (R90), always in the horizontal plane. The syllable from the center was always a distractor (D1). The target (T) might be at L90 or R90, with a second distractor (D2) at the opposite side. (b) Illustration of the events within a trial. A visual cue (VC) was followed by an auditory cue (AC). A 4-syllable mixture was played 1 second after the AC. The participants should respond when the fixation dot turned blue. A green or red dot was presented at the end of the trial, showing the correctness of the response.

in order Center-Right-Left-Center (CTR_RL). From these four conditions, we attempted two binary attention decoding problems: 1) Spa_LR vs Ctr_LR; and 2) Spa_RL vs Ctr_RL. Note that for each of these comparisons, the stimuli presented were exactly matched; only the instructions given to the participant differed. The following sections of this paper will explore the differences in neural signatures for each pair, and how effectively attention can be decoded.

5.2.3.3 EEG processing

EEG was collected using a 64-channel Biosemi system throughout the experiment, sampled at 2048 Hz. Raw EEG data first were bandpass filtered (0.1 - 50 Hz), and were then downsampled to 256 Hz. An independent component analysis was conducted subsequently using EEGLAB [41], [91]. Components that represented eye blinks, eye movements, and muscle artifacts were removed from further analysis.

5.2.3.4 ERP and time-frequency analysis

The continuous EEG data were segmented into epochs to study differences in ERPs and oscillation activity between conditions. In this study, ERP is defined as the condition-wise average EEG waveform time-locked to the onset of the first syllable. The spectro-temporal representation of EEG was studied using a continuous wavelet transform (CWT) implemented using a custom MATLAB script. The wavelet bases (Morlet wavelet with ω_0 = 6) were normalized to have unit total energy at all scales [44].

A group-level cluster-based permutation test [45], implemented with FieldTrip [92], was used to examine the difference in ERP and in CWT between each spatial attention condition (i.e. Spa_LR and Spa_RL) and its corresponding control condition (i.e. Ctr_LR and Ctr_RL, respectively). Both cluster-forming and cluster-significance thresholds were set at 0.05.

5.2.3.5 Feature extraction and classification

Subject-specific linear discriminant analysis (LDA) models were used to decode attention from single-trial EEG data for each of the two classification problems (i.e., Spa_LR vs Ctr_LR, and Spa_RL vs Ctr_RL). Inspired by the results in Section 5.2.4.1 and 5.2.4.2, the feature used for training and testing the model contained multi-channel EEG time-courses as well as the magnitude of the CWT, averaged within each 100 ms interval between 1500 ms and

2700 ms after the onset of the AC (i.e., from the onset of the first syllable to the offset of the third syllable). The CWT magnitudes were also averaged within five frequency bands: delta (2 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 14 Hz), beta (14 – 30 Hz) and gamma (30 – 40 Hz). The decoding accuracy of each binary classification was derived from a leave-one-trial-out cross-validation with 1000 repetitions.

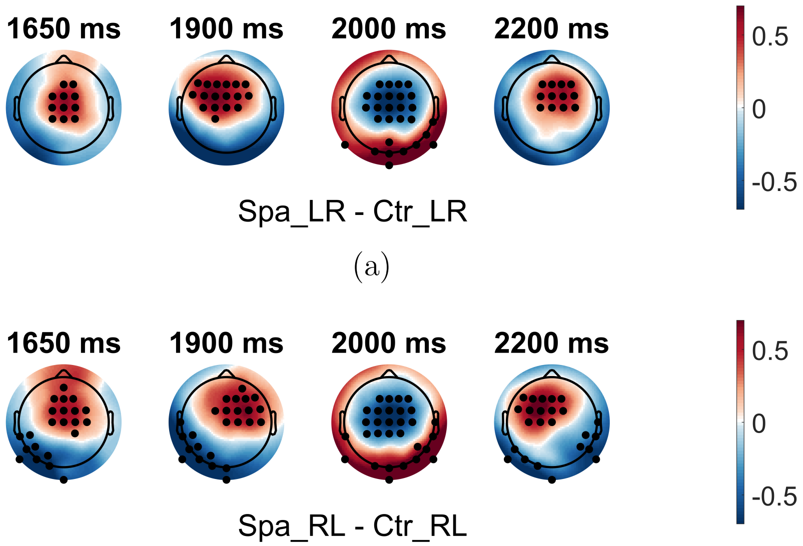
5.2.4 Results

5.2.4.1 ERP analysis

The differences in ERPs between spatial attention and control conditions are shown in Figure 5.2. Significant differences were observed in frontal and parietal channels at multiple time instances. The topographic pattern of the ERP difference was similar for the two contrasts, with a slight difference in the lateralization of the positivity at 1900 ms and 2200 ms. Such lateralization is likely affected by the spatialized location of the syllable being played at those moments.

5.2.4.2 Time-frequency analysis

Event-related synchronization (ERS) and desynchronization (ERD), defined as the percent change in value from one condition to a baseline (i.e., Ctr_LR and Ctr_RL in this study), were used to evaluate the signal change in the time-frequency domain when attention was engaged. Strong ERS was seen in the alpha band before the onset of the last distractor (2400 ms, Figure 5.3a). Higher values of alpha ERS were seen in the frontal and parieto-occipital sensors (Figure 5.3b). In addition, the ERS in the beta band, and the ERDs in the delta, theta and gamma band were also significant in at least one channel throughout the stimuli period.



(a)

(b)

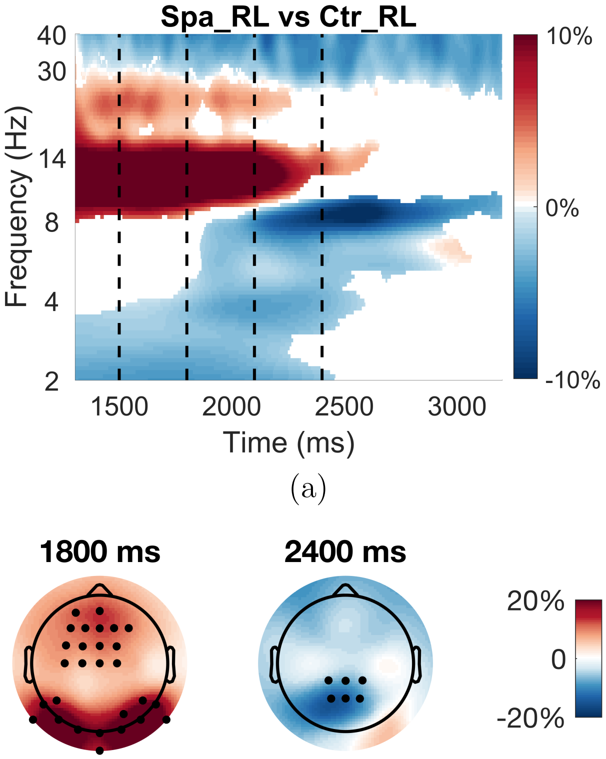
Figure 5.2: Topographic maps of ERP differences between spatial attention and control conditions. Time stamps are with respect to the onset of the auditory cue. Solid dots represent channels with significant effects ($p < 0.05$). Unit: μV

5.2.4.3 Decoding accuracy

Attention can be decoded accurately from EEG in most participants. All results were above 50%, the absolute chance level for a binary classification (Figure 5.4). However, since studies on brain signal classification are generally susceptible to a high false positive rate, Combrisson and Jerbi [93] proposed a method to correct the chance level based on sample size, number of classes, and the desired confidence interval. Even with the corrected chance level (56.94%, 95% confidence), only one classification fell below chance (Figure 5.4a). Table 5.1 shows the average decoding accuracy, their equivalent ITR, and the best ITR among all participants.

Table 5.1: Decoding results & information transfer rate (ITR)

Conditions	**Average accuracy**	**Average ITR (bits/min)**	**Best ITR (bits/min)**
Spa_LR vs Ctr_LR	74.15%	9.44	23.53
Spa_RL vs Ctr_RL	75.83%	10.25	31.70



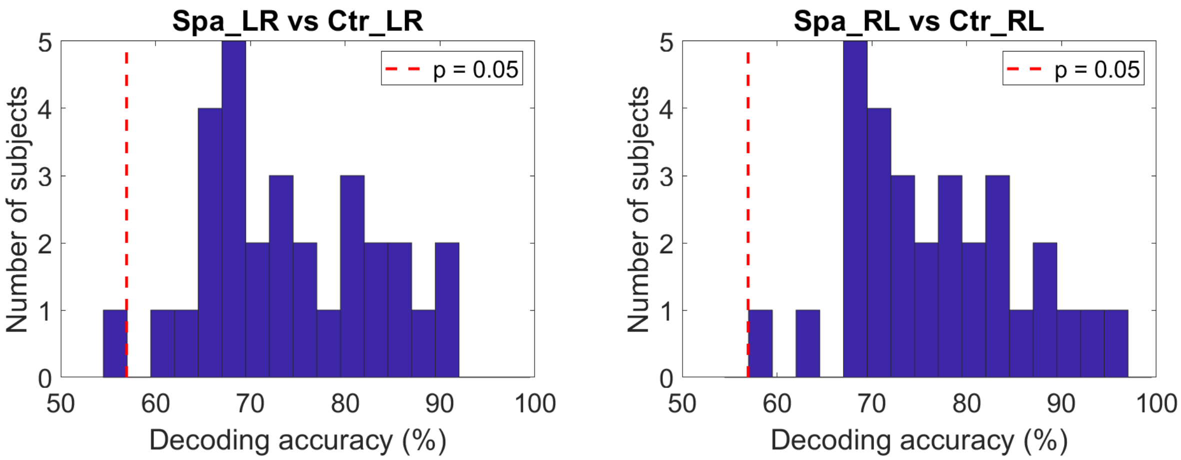
(a)

(b)

Figure 5.3: (a) Average event-related synchronization (positive values) or desynchronization (negative values) across all channels. The values are masked by their significance derived from a non-parametric statistical test ($p < 0.05$). Black dashed lines represent the onset of four syllables. (b) Topographic maps of the alpha power difference between Spa_RL and Ctr_RL. Time stamps are with respect to the onset of the auditory cue. Solid dots represent channels with significant effects for at least one frequency bin ($p < 0.05$).

5.2.4.4 Behavioral correlate

In order to explore the relationship between decoding and attentional effort, the decoding accuracy for each participant was correlated with behavioral performance. In this study, behavioral performance is defined as the percent correct of the syllable identification tasks in spatial attention trials (see Section 5.2.3.2), which represents a proxy for the participant's mental engagement during the task. The results showed significant correlation between behavior and decoding accuracy for both Spa_LR vs Ctr_LR ($\rho = 0.433$, p = 0.017) and Spa_RL vs Ctr_RL ($\rho = 0.567$, p = 0.001, Figure 5.5).



(a)

(b)

Figure 5.4: Histogram of decoding accuracy. The red dashed lines at 56.94% represent the corrected chance level.

5.2.5 Discussion

This study introduced a new auditory BCI system that can generate a binary output within 2 seconds. Human-voiced syllables were used as the stimuli, which are natural, user-friendly, and unlikely to cause fatigue even with extensive usage. Users can voluntarily attend or ignore these stimuli to control the value of the output (e.g., “yes” or “no”). The efficiency of the proposed system is substantially greater than that reported in previous studies that used modulated signals [28], [29], melodies [89], or drip drop sounds [30] as the stimuli. The best ITRs across participants even outperformed some visual BCI systems [33], [94]. The high ITR achieved in this study was due in part to the use of short trials — the classifications were run with only 1.2 seconds of EEG data. Such a brief delay between attentional control and a BCI output may even enable a conversation-level interaction with a computer. To achieve even higher efficiency, in the future, we will explore the feasibility of decoding the direction of spatial attention (left or right) from single-trial EEG. Together with the no-attention condition, we can build a 3-way classifier, which may have better value in real applications.

The current decoding method uses high-dimensional features for classification. Inspired by results in the ERP and the time-frequency analysis, these features contain information

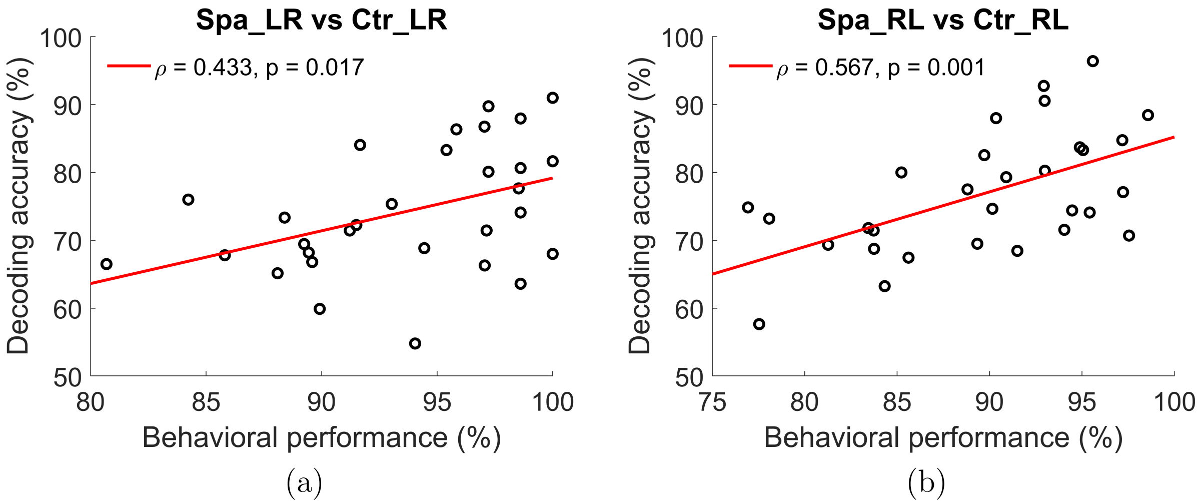


Figure 5.5: Scatter plots showing each subject's behavioral performance and decoding accuracy.

represented in either the time domain or spectro-temporal domain. However, these features may not contribute equally to classification. Including irrelevant features may even decrease the accuracy of the model. Similarly, some EEG channels may contribute more than the others. Shrinking the number of channels while maintaining a high decoding score, if possible, would be important to building unobtrusive BCI systems with few channels. In the future, we will conduct a feature selection analysis by estimating feature weights, and reduce the dimensionality of features used for classification.

It is nearly impossible for participants to sustain full attention throughout the whole experiment. At least some of their incorrect responses during spatial attention tasks are likely due to attention drifting. The strong correlation between decoding accuracy and behavioral performance suggests that some of the wrong classifications might simply originate from a lack of attentional effort during such trials. Therefore, the proposed BCI system has the potential to achieve even higher efficiency if the user is always fully engaged, which is usually the case during real-life applications.

Significant differences in ERP and CWT were shown in group-level statistics, suggesting that some of the contrasting features are common across subjects. Although user-specific classifiers were our main focus, this suggests that a general decoder might be feasible that is not trained on individual subjects. Such a decoder would largely reduce the amount of time and data required to implement a system for a new user. A future study on the feasibility of building a general classifier for all participants is warranted.

5.2.6 Conclusions

The current study proposed a new BCI system based on auditory attention. It not only yielded high efficiency compared with previously reported auditory BCI systems, but also presented pleasant, user-friendly stimuli that allow comfortable long-term use. The system also has the potential to allow decoding that does not depend on subject-specific training.

5.3 Decoding auditory attention from EEG using a convolutional neural network[1]

5.3.1 Abstract

Brain-computer interface (BCI) systems allow users to communicate directly with a device using their brain. BCI devices leveraging electroencephalography (EEG) signals as a means of communication typically use manual feature engineering on the data to perform decoding. This approach is time intensive, requires substantial domain kno wledge, and does not translate well, even to similar tasks. To combat this issue, we designed a convolutional neural network (CNN) model to perform decoding on EEG data collected from an auditory attention paradigm. Our CNN model not only bypasses the need for manual feature engineering, but additionally improves decoding accuracy (∼77%) and efficiency (∼11 bits/min) compared to a support vector machine (SVM) baseline. The results demonstrate the potential for the use of CNN in auditory BCI designs.

5.3.2 Introduction

Electroencephalography (EEG), a noninvasive, mobile, and low cost neuroimaging technique, has become a popular method for developing brain-computer interfaces (BCIs) [26]. Many successful BCI systems have been built around visual attention: when users are asked to focus on a particular visual object, their attentional state can be decoded from EEG signatures such as evoked responses and oscillations. Due to the strength and robustness of visual responses in EEG, these visual paradigms can achieve high decoding accuracy and transmission efficiency. For example, Lin et al. reported an average information transfer rate (ITR) of 20.26 bits/min in their BCI system built on visual event-related potentials

[1] This section is adapted with permission from paper: Winko W. An, Alexander Pei, Abigail Noyce, Barbara Shinn-Cunningham, "Decoding auditory attention from EEG using a convolutional neural network", in *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021*

(ERPs, [33]).

Visual BCIs require the deployment of visual attention, which is not always feasible in real-life scenarios like while walking or driving, or for users with visual impairment. An alternative solution using a different sensory modality is therefore desirable. Previous BCI studies have attempted to decode auditory attention from EEG signals. Kim et al. [28] used two streams of modulated signal with a constant-frequency carrier as the stimuli and decoded users' attention from their auditory steady-state response (ASSR). Kaongoen and Jo [95] developed a hybrid auditory BCI paradigm combining ASSR and ERP. Considering that these modulated signals are not particularly pleasant and can cause user fatigue, researchers have also explored using more user-friendly stimuli, such as drip-drop sounds [30], sequences of tones [89], and music [96], in their BCI design. However, the improved user-friendliness was achieved at the cost of system efficiency — none of these studies yielded an ITR over 3 bit/min. We recently reached a balance between these two goals [62]. We directed users' attention to spatialized human-voiced syllables, trained a support vector machine (SVM) with time-frequency measures of EEG, and achieved high decoding throughput ($\sim$10 bits/min).

One possible way to improve the results in [62] is to adopt a deep learning approach, such as a convolutional neural network (CNN). As opposed to conventional machine learning algorithms like SVM, CNNs do not depend on hand-crafted features for classification. Instead, it automatically learns kernel functions through training, which can help extract features that differentiate multiple classes. CNNs have been widely used in computer vision and more recently in general EEG studies [97], but have not been popularly used in auditory BCIs. In this study, we explored the efficacy of CNNs in decoding auditory attention by comparing with a SVM baseline. We also examined the correlation between CNN decoding and behavioral performance to find a possible cause of the observed individual differences in decoding results.

5.3.3 Methods

5.3.3.1 Participants

Thirty adults with normal hearing (19 – 44 years old, 14 female) were recruited for this study. The Institutional Review Board of Boston University approved this study. Participants were briefed and consented before partaking in this study, and were compensated for their time.

5.3.3.2 Experiment

Subjects sat in a soundproof booth while wearing a pair of insert earphones (ER1, Etymotic Research). The sound stimuli consisted of syllables /ba/, /da/ and /ga/ spoken by native English speakers with varying pitch. To spatialize the sound stimuli, the audio waveforms were convolved with head-related transfer functions provided by the Media Lab, MIT [40]. The simulated locations were center, 30° from the left (L30) or right (R30), or 90° from the left (L90) or right (R90, Figure 5.6a), in the horizontal plane.

The trial began with a one-second visual cue (VC). The VC "Space" indicated that the subject should perform spatial attention, while the VC "Relax" required no attention from the subject (a third condition, "Talker", is not reported here.) After the VC ended, a 500 ms auditory cue (AC) was played. For "Space" attention trials, the AC was a spatialized /a/ syllable, coming from either L90 or R90. In the "Relax" trials, the /a/ syllable came from the midline. 1000 ms after the AC, a 4-syllable mixture consisting of permuted syllables /ba/, /da/, and /ga/ was played. Each syllable was 600 ms in duration and had 300 ms delays between each subsequent syllable onset. The first and last syllables were distractors (D1), which came from the center. The second and third syllables were either another distractor (D2) or the target syllable (T). The target came from the same location as the AC, while D2 came from a location different than the target. For the "Space" trials, subjects were required to report the target using a key press ("1" for /ba/, "2" for /da/, and "3" for /ga/). During "Relax" trials, subjects were asked to passively listen and report a random syllable. Visual feedback was provided indicating correct responses.

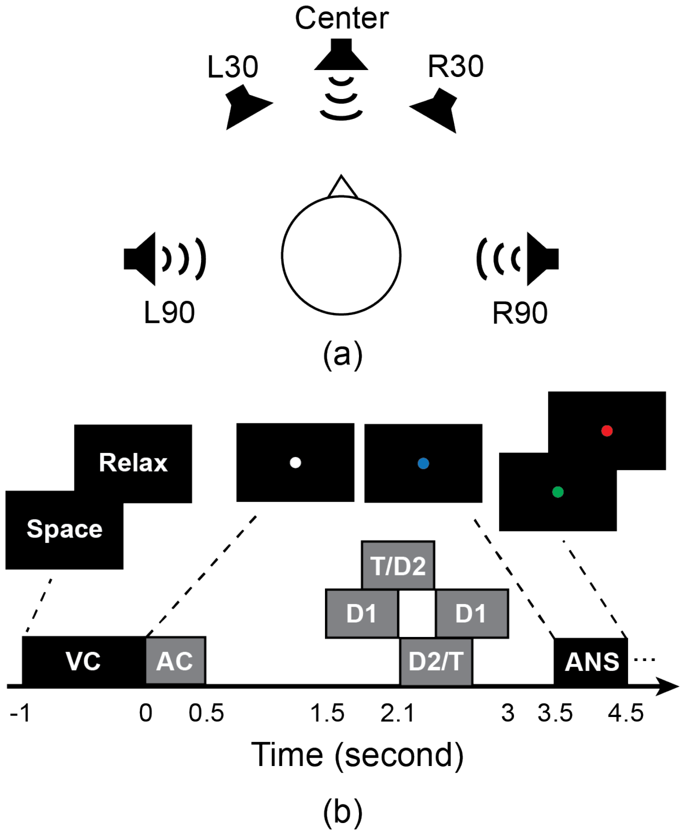


Figure 5.6: (a) (Adapted from [62] with the authors' permission) Spoken syllables were spatialized to center, 30° left (L30), or right (R30), and 90° left (L90), or right (R90), always in the horizontal plane. This figure shows one possible scenario where sounds come from L90, R90 and center. (b) Illustration of the events within a trial. A visual cue (VC) was followed by an auditory cue (AC). A 4-syllable mixture was played 1 second after the AC. Participants were asked to respond when the fixation dot turned blue. A green or red dot at the end of the trial provided feedback.

In total, subjects completed 756 trials for the entire experiment, which lasted for approximately 2 hours. Each trial had variations in location and pitch of the talkers, ordering of the syllables, and attention type; only a subset of the overall data (i.e., trials that required spatial or no attention) was used in this study. We collapsed all spatial attention trials into one condition (288 trials), and all no-attention trials into another condition (252 trials) to perform binary classification. In both conditions, the exact same stimuli were presented; the only difference between the conditions was the instruction for the task.

5.3.3.3 EEG processing

EEG was collected using a 64-channel Biosemi system sampled at 2048 Hz. Raw EEG data were bandpass filtered (0.1 – 50 Hz), and were then downsampled to 256 Hz. As opposed to the previous study, which used independent component analysis (ICA) for artifact removal, we used artifact subspace removal (ASR) to remove artifacts because ASR is more feasible during real-time BCI decoding [98].

5.3.3.4 Feature extraction for support vector machine

Based on prior knowledge about the neural signatures of spatial attention [38], we used both time and frequency representations of the EEG data for decoding. The data for each trial were cropped to contain only the time window from 1.5 to 2.7 seconds after the AC, which we expected to contain the critical neural signatures of interest while the subject is actively performing attention. Continuous wavelet transforms (CWT) were used to generate a spectro-temporal representation of the time-series data. A Morlet wavelet with $\omega_0 = 6$ was used as the wavelet base. Normalization was done to have unit total energy at all scales [44]. The CWT coefficients were then collapsed into five distinct frequency bands that are known to contain signatures of cognitive processes: delta (2 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 14 Hz), beta (14 – 30 Hz) and gamma (30 – 50 Hz). This process yielded a multidimensional time-series for each channel consisting of the channel voltage and the magnitude of the wavelet coefficients in each frequency band. To reduce the data dimensionality and computational demands, data were binned into 100 ms windows. The resulting time-series matrix across channels and features was flattened to produce a single vector. A support vector machine (SVM) with a linear kernel was used to decode this data vector for spatial attention conditions vs. no attention conditions. We performed 10-fold cross-validation to generate training and test sets. The decoding accuracy was averaged across the 10 folds. This process was repeated 20 times, for a total of 200 trained models. Each subject was trained and tested independently from other subjects.

5.3.3.5 Convolutional neural network

Instead of manual feature engineering, the preprocessed EEG time-series in the same 1.5 – 2.7 second time-window was input into a CNN. Our architecture consisted of only three convolutional layers to avoid overfitting, given that our data set is relatively small. Average pooling layers were interwoven between convolutional layers to further reduce the dimensionality of the input. The resulting output of the convolutional layers was flattened and fed through fully connected layers followed by rectified linear unit (ReLU) layers to get a single prediction of the binary class. We additionally used dropout layers to assist with overfitting. Details about the layers can be seen in Figure 5.7.

The training schemes for the CNN differed slightly from that used for the SVM to avoid overfitting due to overtraining. 10-fold cross validation was performed with 20 samples from each condition (40 samples in total) being held out as the validation set in each fold. The CNN was then trained for 40 epochs, and the model with the lowest validation loss across the 40 epochs was used as the final model to classify the testing set. The rationale is that if a model is overtrained in late epochs, the overfitting would lead to an increase in validation loss. By choosing the model with the lowest validation loss, we are technically stopping the training process before the model becomes too complicated to generalize properly, and thus avoiding overfitting. 20 iterations of random initialization were performed for this 10-fold cross validation. The testing accuracy was averaged across these iterations to estimate the classification performance of the model. We used a cross-entropy loss function, Adam optimizer with a learning rate of 0.0001, a lambda weight decay of 0.01, and a batch size of 50.

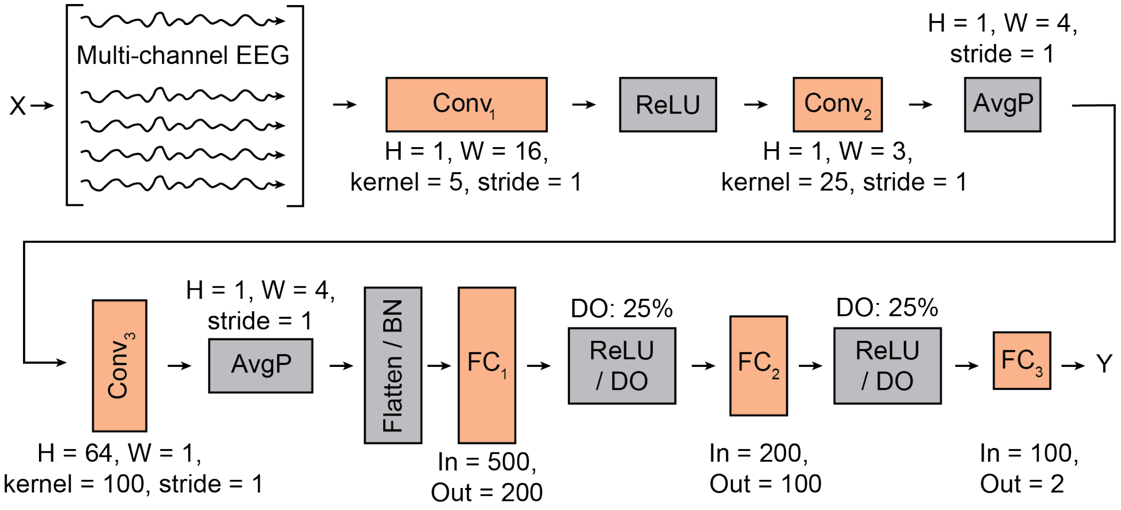


Figure 5.7: The CNN architecture used in this study.
Conv – convolutional layer; H – height of kernal; W – width of kernal; ReLU – rectified linear unit; AvgP – average pooling layer; BN – batch normalization layer; FC – fully connected layer; DO – dropout layer

5.3.4 Results and Discussion

5.3.4.1 Classification accuracy

The average classification accuracy of the SVM approach was 72.10% (Tab. 5.2), a slight drop from the result (∼75%) in [62], where ICA was adopted for artifact removal. Because ASR requires much less time to process than ICA and can be used in a real-time manner, it seems reasonable to replace ICA with ASR in a BCI system design for real-life applications.

The CNN method proposed in this study significantly improved the classification performance compared to the SVM approach (paired t-test, $p<0.001$, Figure 5.8). It achieved a 77.01% decoding accuracy; moreover, each individual subject showed a performance gain over SVM. Given that attention was decoded from only 1.2 seconds of data, the proposed BCI system is highly efficient. The average equivalent ITR [99] is 11.11 bits/min, and the best ITR among all participants is 32.03 bits/min, on par with some visual BCI paradigms. In the future, we will attempt other advanced machine learning methods, such as convolutional long short-term memory (ConvLSTM) and adaptive learning [100] to seek even better classification performance.

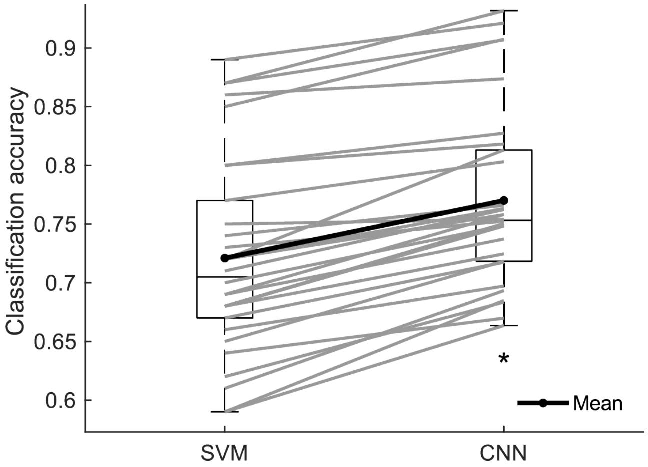


Figure 5.8: SVM and CNN classification results. Each gray line represents data from one subject. CNN yielded a significantly higher average classification accuracy than SVM. *p<0.001

Table 5.2: Classification accuracy & information transfer rate (ITR)

Classifier	Average accuracy	Average ITR (bits/min)	Best ITR (bits/min)
SVM	72.10%	7.30	25.00
CNN	77.01%	11.11	32.03

5.3.4.2 Performance gain with CNN

The gain in classification accuracy of CNN over SVM varied across participants. Figure 5.9 shows that this improvement is strongly and negatively correlated with the SVM classification results (ρ=-0.541, p=0.002). The CNN method seems to have benefited subjects with a lower decoding score more than those with a higher one, and thus reduced the variability in decoding accuracy across subjects. One possible reason is that the SVM accuracy is low in some subjects not only because there is less distinguishing information in their EEG signals, but because such information is not reliably extracted from EEG using the CWT method. The CNN approach does not rely on hand-crafted features, but rather learns through training what features to use. It may help preserve information that is present, but does not get represented in generic, hand engineered features. The CNN approach thus may be especially beneficial to participants with a low SVM score.

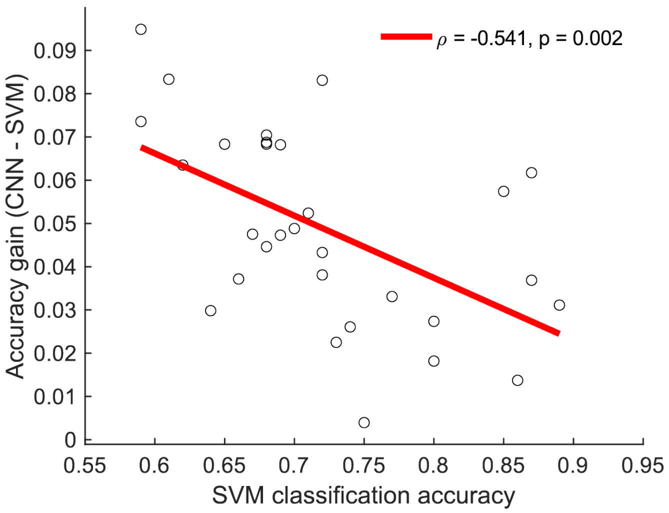


Figure 5.9: The gain in classification accuracy from using CNN over SVM is negatively correlated with the SVM classification results. Each circle represents data point of one subject.

5.3.4.3 Correlation with behavioral performance

The participants exhibit a wide range of behavioral performance in this study — some nearly achieved a perfect score in the attention task, while some others answered correctly in less than 70% of the trials. Interestingly, we observed a strong positive correlation between the participants' behavioral performance and their attention decoding accuracy using CNN (ρ=0.583, p<0.001, Figure 5.10). This suggests that the variance in individual CNN classification results shown in Figure 5.8 can be partially explained by how well a participant performed in the attention task. If the main reason for giving an incorrect response is that a subject's attentional focus drifted, the proposed BCI system has the potential to achieve even better accuracy and efficiency if the user is always fully engaged and motivated, which is more likely during real-life applications.

5.3.5 Conclusions

This study proposed a method to decode auditory attention from single-trial EEG for the purpose of building a BCI system. We adopted a subspace-based artifact removal pipeline,

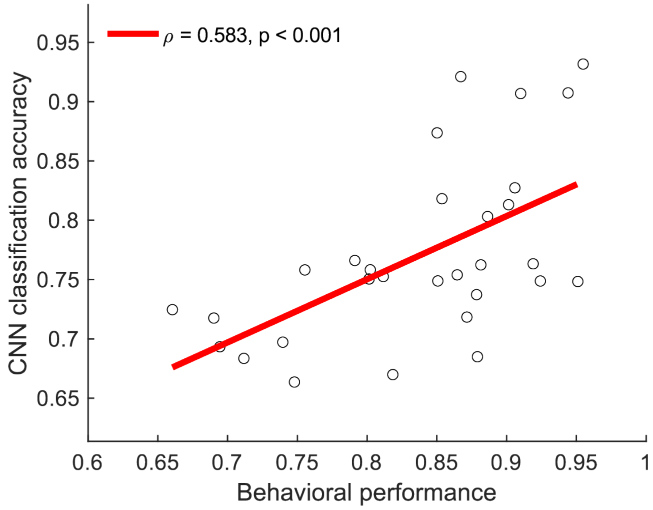


Figure 5.10: The CNN classification accuracy is positively correlated with the subject's behavioral performance in the attention task.

which can process signals in a real-time manner. The CNN approach yielded high classification accuracy and efficiency, outperforming a SVM baseline as well as previous studies. The CNN decoding results are strongly correlated with the participants' behavioral performance in the attention task, suggesting a possible improvement in decoding, when used in real-life applications, where users are highly motivated.

5.4 Decoding auditory and tactile attention for use in an EEG-based brain-computer interface[1]

5.4.1 Abstract

Brain-computer interface (BCI) systems offer a non-verbal and covert way for humans to interact with a machine. They are designed to interpret a user's brain state that can be translated into action or for other communication purposes. This study investigates the feasibility of developing a hands- and eyes-free BCI system based on auditory and tactile attention. Users were presented with multiple simultaneous streams of auditory or tactile stimuli, and were directed to detect a pattern in one particular stream. We applied a linear classifier to decode the stream-tracking attention from the EEG signal. The results showed that the proposed BCI system could capture attention from most study participants using multisensory inputs, and showed potential in transfer learning across multiple sessions.

5.4.2 Introduction

Brain-computer interface (BCI) systems offer a non-verbal and covert way for humans to communicate a control signal to a computer. Among the neuroimaging modalities that are currently available, electroencephalography (EEG) has become the most popular choice for BCI applications due to its noninvasiveness, mobility, and low cost [26]. EEG monitors brain activity through sampling the electrical potential along the scalp at a very high rate. The high temporal resolution of EEG oscillations allows capturing certain neural signatures of a brain state or mental efforts, which can be used to decode users' intention.

Many successful BCI systems rely on external stimulation, especially with visual stimuli.

[1] This section is adapted with permission from paper: Winko W. An, Hakim Si-Mohammed, Nicholas Huang, Hannes Gamper, Adrian KC Lee, Christian Holz, David Johnston, Mihai Jalobeanu, Dimitra Emmanouilidou, Edward Cutrell, Andrew Wilson, Ivan Tashev, "Decoding auditory and tactile attention for use in an EEG-based brain-computer interface", in *Proceedings of the 8th International Winter Conference on Brain-Computer Interface, 2020*

For example, visual P300 is a well-studied event-related potential (ERP) that is elicited as a response to infrequent events, or "oddballs." It usually happens around 300 ms after the onset of the event, and could be captured by sensors in the parietal area [26]. Another popular neural signature of visual stimuli is the steady-state visual evoked potential (SSVEP), which is the response in the visual cortex to a constant-frequency flickering stimulus. These vision-based paradigms have high efficiency for transmitting bits to a computer, which can be quantified by their information transfer rate (ITR). Previous studies yielded an ITR of 20.26 bits/min using P300 [33], or 30.10 bits/min using SSVEP [27].

A major disadvantage of using visual stimuli for BCI is the level of visual attention required to complete the task in the presence of competing stimuli and the fact that it could interfere with competing tasks (e.g., walking, driving) when vision is primarily involved. It also requires correctable vision and voluntary gaze control, making it inaccessible to users with severe visual impairment or locked-in-syndrome. In view of this, previous studies have focused on developing an attention-based BCI system using auditory or tactile stimuli. They used modulated signals with a constant carrier frequency as the input, usually in multiple streams of spatial sound [101] or vibration on fingers [102]. Attention was decoded from auditory steady-state response (ASSR) or steady-state somatosensory evoked potential (SSSEP), where the EEG signal is locked to the modulation frequency of the attended stream. However, these sinusoidal carriers with a constant frequency were perceived by users to be annoying or fatiguing [29]. There were also no behavioral metrics to verify the attentional state of the participants. Separately, another recent work on auditory selective attention revealed that lateralized alpha-band (8–12 Hz) power could be a more effective neural signature than ASSR for use in BCI [90]. This is in line with previous studies that have shown an important role of parietal alpha activity in attention to auditory stimuli [38], [39], even while walking [44].

The current study proposes a user-friendly, attention-based BCI paradigm using auditory and tactile stimuli. A task was embedded in the stimuli, which demanded attentional focus from the participants. The auditory stimuli were spatialized melodies, which sound more pleasant and are easier to attend to than monotones. The tactile stimuli were pulsed vibrations applied to the user's wrists, so that their hands were freed for other hypothetical concurrent work. Both the melodies and the vibrations were amplitude-modulated, which might induce steady-state responses. Since various neural signatures (e.g., ASSR, SSSEP, lateralized alpha and gamma activity) could be expected from the EEG signal, the multisensory attention was decoded by an individualized linear model with full spectral information (from alpha to gamma band). The model's ability in transfer learning was also evaluated through recording a subset of participants across multiple sessions.

5.4.3 Method

5.4.3.1 Participants

Twelve adults (32.2 ± 7.4 years old, 3 female) volunteered to participate in this study. Eleven participants were novel to BCI upon recruitment, among which seven had no experience with EEG experiments. No participants reported known history of neurological disorder or hearing loss.

5.4.3.2 Experiment

Before the experiment started, the participants were asked to sit comfortably in front of a computer, read the instructions from the screen, and familiarized themselves with the stimuli. The experiment consisted of 3 blocks using auditory stimuli, 3 blocks using tactile stimuli, and 3 blocks using both auditory and tactile stimuli simultaneously (mixed). At the beginning of each block, a text message ("audio", "tactile" or "mixed") was presented in the center of the screen, indicating the sensory modality about to be stimulated. The order of the blocks was randomized for each participant.

The experiment for audio blocks was adapted from a previous study on auditory selective attention [36]. Two streams of modulated signals were used as stimuli, with one presented to the left ear, and the other to the right ear. The audio signals were delivered through a pair of MaximalPower RHF 617-1N earpieces which transmit sound through acoustic tubes, thus reducing possible electromagnetic interference with the EEG signal. Each stream had multiple standard (S), high-pitched (H) and low-pitched notes (L). Each note contained six harmonics of the fundamental frequency (f0), making it sound more natural than a single sinusoid. The configurations of the two streams are summarized in Table 5.3. The left stream was formed by 9 repetitions of 400-ms notes, among which the first five were always standard (Figure 5.11a). The f0 of the last four notes determined whether the melodic pattern of the stream was "rising" (···-S-H-H-H-H), "falling" (···-S-L-L-L-L) or "zig-zag" (···-S-H-H-S-S). Similarly, the right stream was formed by 12 repetitions of 300-ms notes, with the first five always being standard. The melodic pattern of this stream could be "rising" (···-S-H-H-H-H-···), "falling" (···-S-L-L-L-L-···), or "zig-zag" (···-S-H-H-H-S-···), depending on the f0 of the last seven notes. The two streams always started at the same time and were played back simultaneously.

There were 24 trials in each audio block. At the beginning of each trial, a visual cue (VC) was shown on the screen to direct the participant's attention to the left stream, right stream or neither of the two (Figure 5.11c). The cue was replaced by a white fixation dot after 1 second, and two streams of melodies started to play 0.5 second later. The participants were asked to identify the melodic pattern of the attended stream and answer with the keyboard after the fixation dot turned blue. Visual feedback at the end of each trial indicated whether they identified the melodic pattern correctly (green dot) or incorrectly (red dot). The average behavioral performance was shown at the end of each block. The inter-trial interval was set to 2 seconds.

The design of the tactile experiment was analogous to that of the auditory one. The tactile stimuli consisted of two streams of vibration which were applied separately to the left and right wrist of the participant. The streams were rendered through two coin-type loudspeakers (DAEX19CT-4, Dayton Audio) taped to the participant's wrists (Figure 5.11d). Similar to

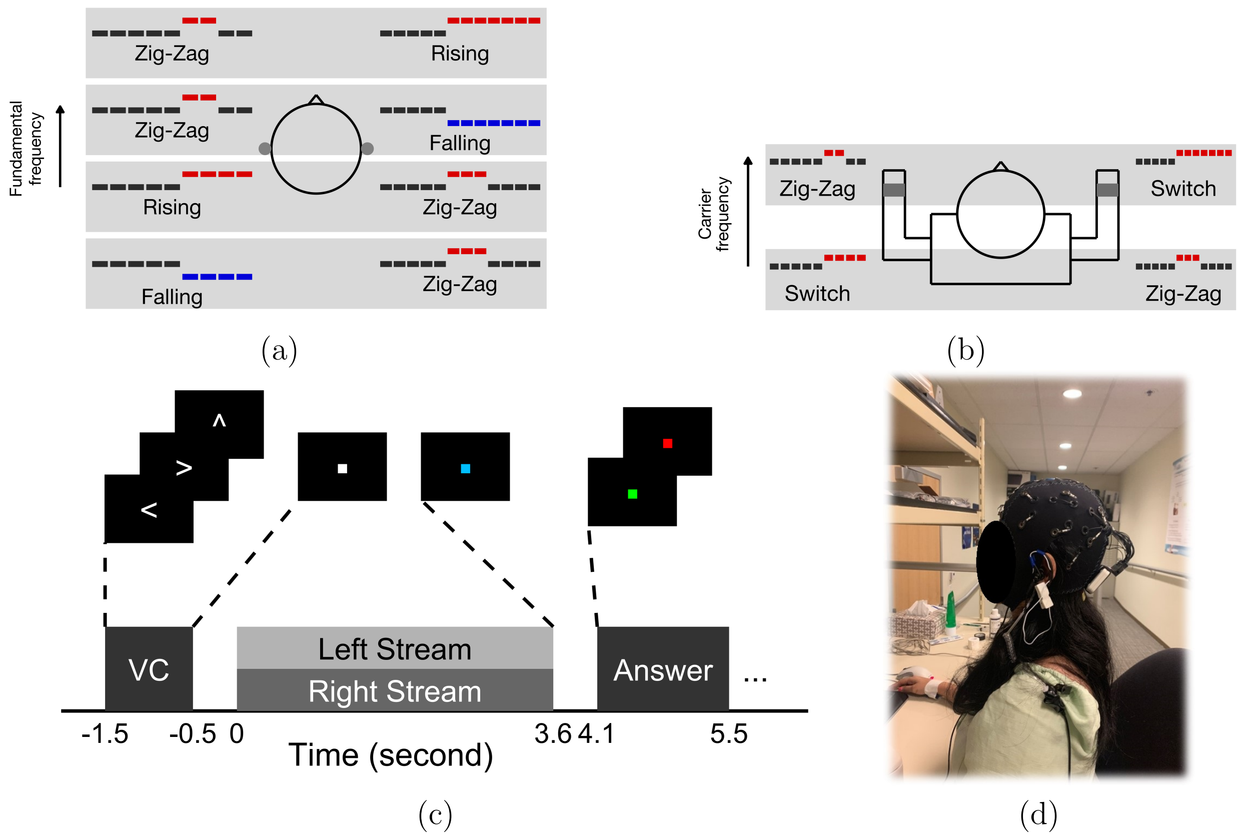


Figure 5.11: Illustration of (a) four scenarios of left and right sound streams in an audio block; (b) two scenarios of left and right vibration streams in a tactile block; (c) event sequence in one trial; (d) photograph illustrating the experimental setup.

the audio blocks, modulated signals in the form of pulse trains were used for both streams. Their configurations are shown in Table 5.4. The modulation and the carrier frequencies were carefully selected through piloting, so that the participants could feel, but not hear the vibration. Unlike in the audio blocks, where there were three types of notes (S, H and L), there were only two types of vibration pulses in tactile blocks, standard (S) and oddball (O). The reason behind this difference is that though most participants could perceive a change in the tactile carrier frequency, they were unable to identify whether it was increasing or decreasing relative to S. Hence, a single oddball condition was the only choice for the tactile experiment. The design of the two spatially separated vibration streams was very similar to that of the sound streams (Figure 5.11b). The first five pulses in the left stream were always standard, and the last four could form a "switch" pattern (· · ·-S-O-O-O-O) or a "zig-zag"

Table 5.3: Configurations of sound streams

Stream	**Length (ms)**	**Modulation frequency (Hz)**	**Fundamental frequency (Hz)**		
			Low	Standard	High
Left	400	37	703	740	777
Right	300	44	396	440	484

pattern (· · ·-S-O-O-S-S). The first five pulses in the right stream were always standard, and the last seven could form a "switch" (· · ·-S-O-O-O-O-· · ·) or a "zig-zag" (· · ·-S-O-O-O-S-· · ·) pattern. There were 24 trials in each tactile block. In analogy to the audio condition, the participants were asked to identify the vibration pattern in the attended stream and respond with the keyboard.

Table 5.4: Configurations of vibration streams

Stream	**Length (ms)**	**Modulation frequency (Hz)**	**Carrier frequency (Hz)**	
			Standard	Oddball
Left	400	27	120	210
Right	300	17	120	210

In multisensory ("mixed") blocks, the streams of sound and vibration that were used in the audio and the tactile blocks were played concurrently. The melodic pattern and the vibration pattern of the streams on the same side of the participant were matched. For example, a "rising" or "falling" left sound stream was matched to a "switch" left vibration stream; a "zig-zag" right sound stream was matched to a "zig-zag" right vibration stream. Since the notes and pulses on the same side had the same length, when the two streams were played simultaneously, the onset of the frequency change for the two sensory modalities on the same side was synchronized. The task was the same as the one for the audio blocks, where the participants were asked to identify the melodic pattern of the attended stream.

The user interface for all tasks was created in MATLAB. During the experiment, EEG signals were collected using a wireless, gel-based 24-channel system (mBrainTrain Smarting), at a sampling of 500 Hz.

5.4.3.3 EEG processing

The EEG signals were processed using EEGLAB [41] functions and custom MATLAB scripts. The signals were first band-pass filtered by a finite-impulse-response bandpass filter with cut-off frequencies at 0.1 Hz and 50 Hz. After re-referencing to the common average, an adaptive mixture independent component analysis (AMICA) [91] method was used to separate noise and artifact components from the signals. An automatic EEG artifact detector, ADJUST [103], was then used to select and remove components representing eye blinks and movement. On average, 3.08 ± 1.67 components were removed from each participant.

The continuous EEG data were then segmented into epochs for further analysis. Each epoch contained data within 500 ms before and 3600 ms after the stimulus onset of each trial. The 216 epochs (9 blocks x 24 trials/block) were then divided into 9 conditions depending on their sensory modality (audio/tactile/mixed) and attention type (attend left/attend right/no attention). The EEG data were further cleaned by removing trials with extreme values, which might represent random noise or strong motion artifacts. Trials with peak values beyond three standard deviations from their conditional average were removed from the pool. On average, 21.54 ± 1.62 trials per condition remained for each participant.

5.4.3.4 Feature extraction and classification

A participant-specific linear discriminant analysis (LDA) model was used to decode attention type (left, right or no attention) from single-trial EEG data within each one of the three sensory modalities (audio, tactile or mixed). Spectral information of each epoch, in the form of the magnitude of its Fast Fourier Transform (FFT) coefficients (8 - 50 Hz), was used as the feature to train and test the model. The FFT was calculated using a 3-second sliding window with 90% overlap. Since an epoch (4.1 seconds) was longer than the FFT window length, multiple samples were drawn from each trial, which served well for the purpose of data augmentation. Features of multiple channels were concatenated into a single vector. Its dimensionality was then reduced by principal component analysis (PCA) retaining 99% of the variance.

The accuracy of each 3-way classification was derived from a 10-fold cross-validation (1000 repetition). To prevent information leakage, trials were divided into training and testing sets before data augmentation. The classification of one trial was done by averaging the sample-level posterior probabilities of all testing samples that belonged to that trial, and choosing the one with the highest probability as the decoding output.

5.4.3.5 Feature weight estimation

In order to verify whether the decoding was based on neurologically relevant factors, a post-hoc feature weight estimation was run using neighbourhood components analysis (NCA). It estimated the weight of each feature dimension. The feature before PCA was used for this analysis, and each estimated weight represented the importance of one frequency bin at one particular channel.

5.4.3.6 Cross-session validation

In order to evaluate the transfer learning ability of the proposed BCI system, three participants were invited back to repeat the exact experiment one week after their first attempt. They were chosen based on their decoding score from Session 1 — one with the highest score (participant 72, >70%), one around the average (participant 45, ∼60%), and one around the chance level (participant 78, ∼40%). Classification was first done using the within-session decoding method as described above, i.e., training and testing an LDA model using the data only from Session 2. A subsequent cross-session decoding was conducted by training a model with data from one session, and having it tested on another. The cross-session decoding result was compared to that of the within-session decoding to evaluate the model's ability to generalize.

5.4.4 Results

5.4.4.1 Attention decoding

The absolute chance level of a 3-way decoding is 33.33%. However, studies on brain signal classification, like the current one, are generally susceptible to a high chance of false positives due to small sample size. To tackle this problem, Combrisson and Jerbi [93] suggested calculating the chance level as a function of sample size, number of classes, and the desired confidence interval based on a binomial cumulative distribution. Using this method, the significant chance level in this study is corrected to 43.06% (p=0.05).

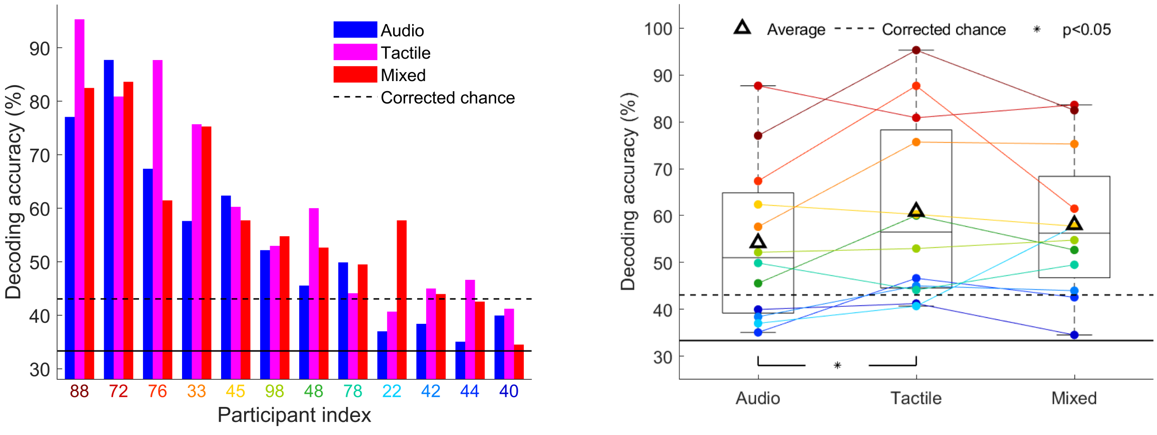
The decoding accuracy of most participants exceeded the corrected chance level, despite the existence of substantial individual differences (Figure 5.12a). EEG of 4 participants were not significantly classifiable in at least one sensory modality, while 3 participants' decoding was over 70% in at least two sensory modalities. The highest decoding accuracy for audio, tactile and mixed conditions were 87.72%, 95.32% and 83.63%, respectively. Their equivalent ITRs are shown in Table 5.5. Within the three types of sensory modalities, tactile conditions had the highest average decoding accuracy (Table 5.5), which is significantly higher than that of audio conditions (p=0.0348).

Table 5.5: Decoding accuracy and information transfer rate (ITR)

Modality	Average decoding	Average ITR (bits/min)	Best ITR (bits/min)
Audio	54.18%	1.90	13.69
Tactile	60.87%	3.37	18.27
Mixed	58.02%	2.69	11.57

5.4.4.2 Behavioral performance

Most participants could identify the melodic patterns with high accuracy for the audio (92.53%±9.10%) and the mixed (93.92%±7.91%) conditions (Figure 5.13). Out of the 12



(a) (b)

Figure 5.12: (a) The decoding accuracy of each participant in the order of high to low average score. (b) The decoding accuracy grouped by sensory modality. Each line represents a participant. The lines and labels are color-coded, where a warmer color denotes a higher average decoding score.

participants, 6 completed the audio or the mixed task with a perfect score. The tactile task appeared to be the most difficult, with a behavioral performance (65.28%±16.89%) significantly lower than that of the other two sensory modalities ($p<0.001$). Only one participant completed the tactile task with more than 90% correct. One interesting observation is that the behavioral performance in tactile blocks seemed to divide the participants into two subgroups — one with a performance score above 75% (n=5), and one below (n=7). The behavioral results of these subgroups in other sensory modalities were not separable.

5.4.4.3 Behavioral correlate

Despite the fact that the highest decoding score and the worst behavioral performance both happened in tactile blocks, there was no significant correlation between these factors across subjects (Fig 5.14). Neither a linear nor a quadratic function could fit all the data in tactile blocks with high confidence. However, a subgroup analysis revealed that the participants with a high behavioral score in tactile task (>70%) had a linear behavioral correlate with their decoding score in all sensory conditions. This relationship was not seen in the other subgroup with low behavioral scores in tactile task.

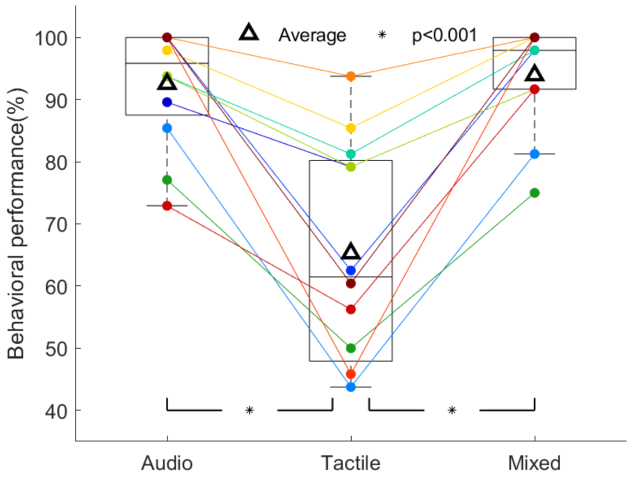


Figure 5.13: The behavioral results of all participants performing the pattern identification task in each sensory modality. Each line represents a participant. The lines are color-coded by the participant's average decoding accuracy (same as in Figure 5.12).

5.4.4.4 Feature weight

The estimated feature weights of 4 participants with good decoding scores are shown in Figure 5.15. Surprisingly, high weights are not associated with any of the modulation frequencies as in an ASSR/SSSEP feature (dashed lines). Instead, they appear mostly in alpha (8 – 12 Hz) and gamma bands (>30 Hz) in these participants (shaded regions). Topographic maps of maximum feature weights in alpha band reveal dominant patterns in the parietal and occipital channels, especially in decoding tactile attention (Figure 5.16). For gamma band feature weights, the dominant patterns reside in the frontal and temporal channels. (Figure 5.17).

5.4.4.5 Cross-session validation

Three participants were invited to repeat the experiment for cross-session validation. The decoding scores of participant 45 and 78 from Session 2 were in line with their results from Session 1 (Figure 5.18). The scores of participant 88 dropped from the first session, but still remained high in all three modalities. Notably, the cross-session validation of all three

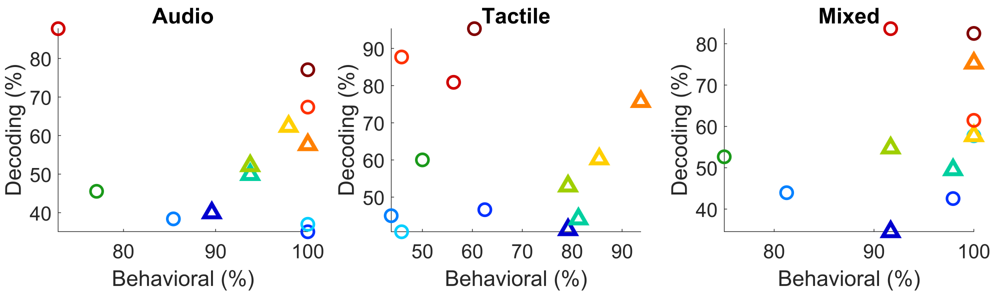


Figure 5.14: Correlation between behavioral performance and attention decoding. Each circle or triangle represents a participant whose behavioral performance in tactile blocks was blow or above 70%, respectively. The circles and triangles are color-coded by the participant's average decoding accuracy (same as in Figure 5.12).

participants stays in the same range as their within-session decoding.

5.4.5 Discussion

5.4.5.1 Transmission efficiency

The efficiency of the proposed BCI system is on par with previous works. The average ITR derived from audio blocks (1.90 bits/min) was comparable to previous studies on ASSR-based BCIs [29], [104]. The average ITR in tactile blocks outperformed the results of previous BCI designs with vibrotactile actuators attached to the user's thumb (∼1.19 bits/min, calculated from the reported accuracy) [102], or five fingers (1.2 bits/min) [105]. A recent study reported a higher ITR (4.9 bits/min), but electrical stimulation on four fingers was needed [106]. Promisingly, the highest ITR achieved in the tactile condition was comparable to that of some vision-based BCI systems previously reported [94]. There is great potential in improving the decoding results with the current dataset. For example, the pattern identification task used in this experiment is not dissimilar to detecting oddball events. One possible direction is to combine a P300 feature into the current one to improve decoding accuracy.

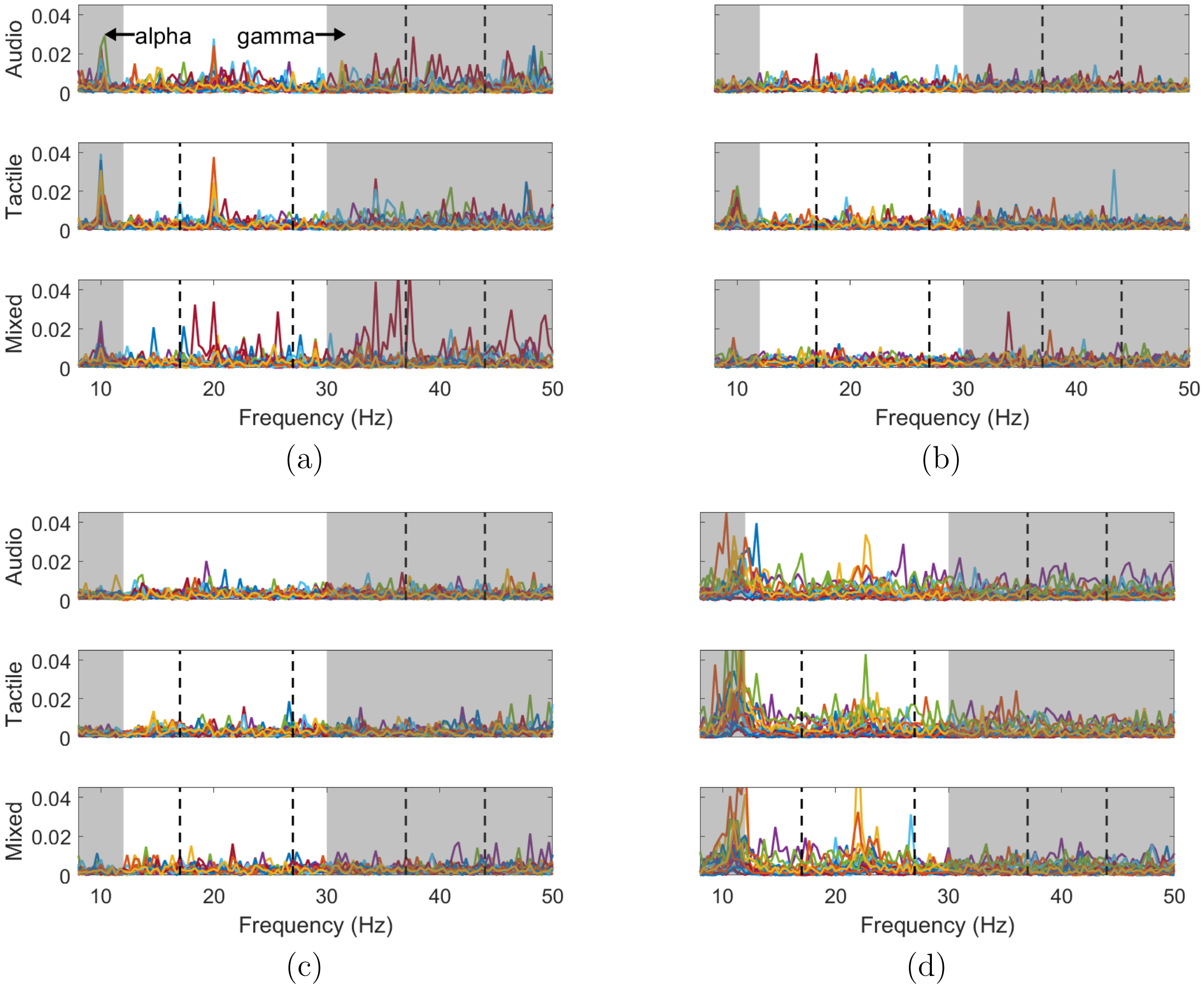


Figure 5.15: Estimated feature weights for participant (a) 72 (b) 76 (c) 33 (d) 45. Each trace represents a channel. Y-axis has arbitrary unit. The dashed lines in each panel represent the modulation frequencies of the stimuli used for that particular sensory modality. The shaded regions denote the alpha and the gamma bands

5.4.5.2 Decoding based on spatial attention

This study used modulated signals as the stimuli. We expected that attention would enhance the neural representation of the steady-state response of the modulation. Instead, the high feature weights in parietal alpha and temporal gamma indicated that spatial attention was the dominating factor of attention decoding in this study.

Steady-state responses played a less important role in this BCI design than in a typical ASSR/SSSEP-based BCI system. One explanation for such difference is the use of discrete

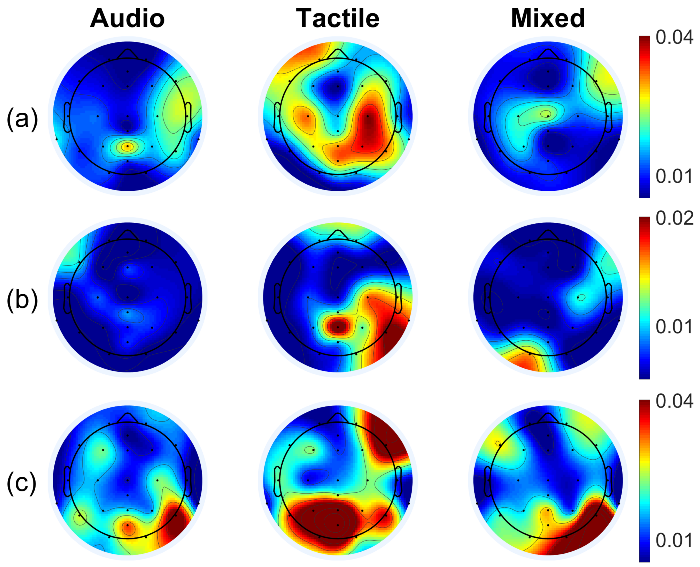


Figure 5.16: Topographic maps of maximum feature weights in alpha band for participant (a) 72 (b) 76 (c) 45.

stimuli. The pulses might have been processed by the brain as individual events instead of a continuous stream, so the brain never truly entered a steady state during the experiment. Using continuous stimuli, such as natural speech, might help enhance the representation of steady-state response in EEG.

5.4.5.3 Transfer learning

The cross-session validation results were comparable to their corresponding within-session results. It indicates that the participants might have adopted a similar strategy to focus even on different days. The modest drop in some participants might be due to a slightly different EEG cap placement between sessions. This result demonstrates some transfer learning ability in the proposed BCI system, showing potential in improving the model through multiple training sessions [107].

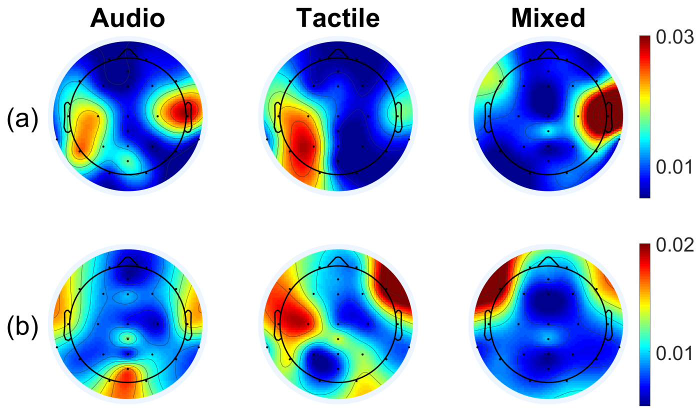


Figure 5.17: Topographic maps of maximum feature weights in gamma band for participant (a) 72 (b) 33.

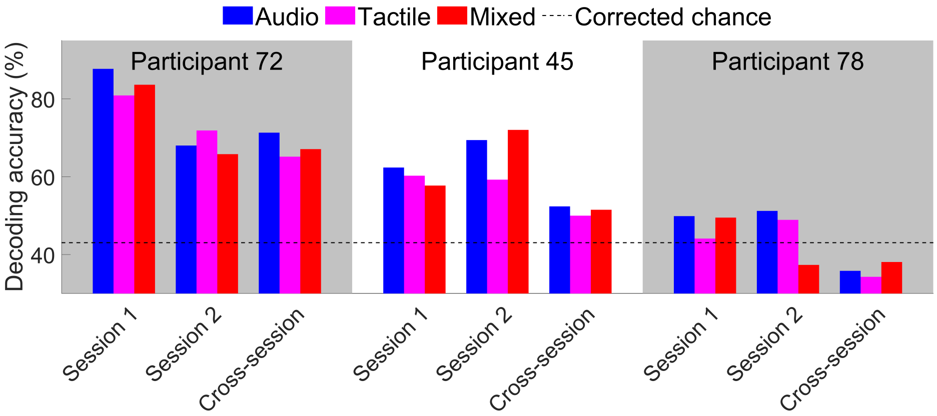


Figure 5.18: Decoding result of the cross-session validation analysis

5.4.6 Conclusions

The current study proposed a new BCI system based on auditory and tactile attention. It yielded an efficiency comparable to or even higher than the existing BCI paradigms, without engaging the user's hands or eyes. The highest efficiency achieved in the tactile condition was close to a visual-based BCI. The system also demonstrated certain transfer learning ability.

5.5 Decoding music attention from EEG headphones: a user-friendly auditory brain-computer interface[1]

5.5.1 Abstract

People enjoy listening to music as part of their life. This makes music an excellent choice for designing a user-friendly brain-computer interface (BCI) for long-term use. We propose a novel BCI system using music stimuli that relies on brain signals collected via Smartfones, an EEG recording device integrated into a pair of headphones. In a user study of the proposed system, participants were asked to pay attention to one of three musical instruments playing simultaneously from separate spatial directions. We used a stimulus reconstruction method to decode attention from EEG signals. Results show that the proposed system can achieve good decoding accuracy (>70%) while providing superior user-friendliness compared to a traditional EEG setup.

5.5.2 Introduction

A brain-computer interface (BCI) offers a covert and non-verbal way to communicate with a computer. BCIs have great potential in applications including assistive technology and emotion monitoring [108]. Electroencephalography (EEG), due to its mobility, low cost, and proven relevance to cognitive functions [38], [44], has become a popular choice for BCI design. Previous studies have demonstrated great success in building EEG-based BCI systems using visual or auditory stimuli. Chen et al. [27] designed a high-throughput visual BCI system using flickering objects. When the user focuses on one of them, a neural signature known as the steady-state visual evoked potential (SSVEP) appears in EEG signals. However, SSVEP

[1] This section is adapted with permission from paper: Winko W. An, Barbara Shinn-Cunningham, Hannes Gamper, Dimitra Emmanouilidou, David Johnston, Mihai Jalobeanu, Edward Cutrell, Andrew Wilson, Kuan-Jung Chiang, Ivan Tashev, "Decoding music attention from "EEG headphones": a user-friendly auditory brain-computer interface", in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021*

requires a stable line of sight, which may not be available due to permanent or situational impairment (e.g., while driving). As an alternative solution, researchers applied a similar idea to designing auditory BCI systems, where the users were presented with multiple streams of pure tones modulated at different frequencies. The modulation frequency of the attended stream may result in a strong EEG component known as the auditory steady-state response (ASSR) [28].

One major disadvantage of SSVEP or ASSR paradigms is the use of flickering objects or modulated pure tones, which can cause fatigue in users. Recent studies endeavored to use more naturalistic and pleasant stimuli to improve the user-friendliness of BCI systems. Huang et al. [30] used drip-drop sounds in their BCI design, creating a relaxing auditory scene for the users. An et al. [62] designed an attention task with human-voiced syllables, and achieved a high accuracy in detecting whether the user is paying attention. In another study, An et al. [89] built an auditory BCI system with a sequence of tones forming melodic patterns.

Here we explore the feasibility of using music stimuli for BCI design. We decode a user's attention to a particular musical instrument while listening to polyphonic music. This idea was previously attempted by Treder et al. [32], who embedded oddballs in music streams and used the oddball-evoked response for attention decoding. Despite achieving a high accuracy, their system averages 40 seconds of data to generate one output, which may be slow for real-time applications. Here, we adopt a different decoding method called auditory attention decoding (AAD) [109] and decode attention within a time window of just 8 seconds. The AAD method linearly combines multi-channel EEG signals to reconstruct a stimulus envelope, which tracks the envelope of the attended stimulus more strongly than the unattended one. This method has been successfully applied in decoding attention to continuous speech for BCI purposes [110], [111]. To further improve the user-friendliness of the design, we used Smartfones (mBrainTrain, Serbia) as the form factor, which is a compact EEG recording device integrated into a pair of headphones. It is a saline-based system with three sensors on top of the head and four on each side around the ear, for a total of 11 sensors. It has less coverage than a traditional EEG cap, but is a good option for this study for its all-in-one design.

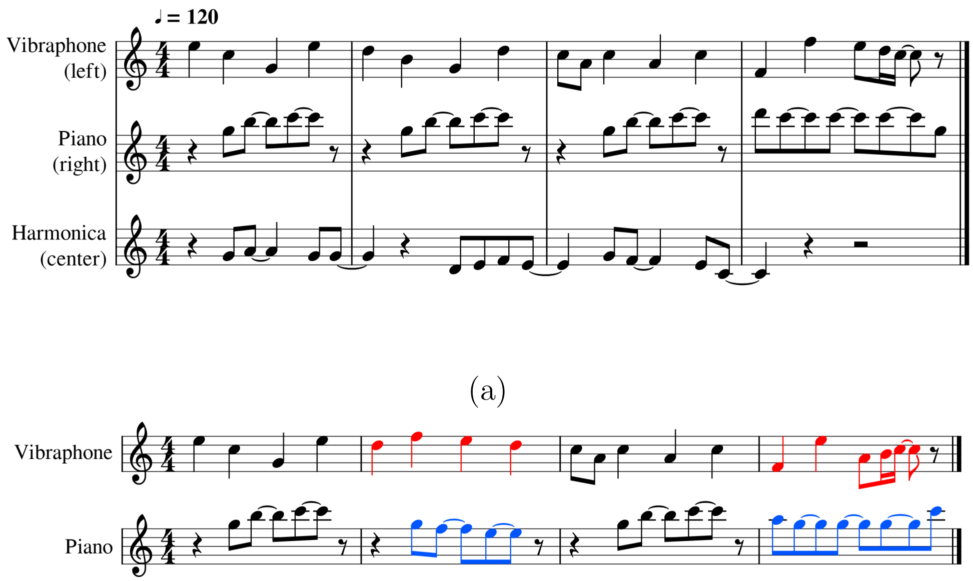
5.5.3 Materials and Methods

5.5.3.1 Participants and Stimuli

Nine adults (34.0 ± 3.1 years old, 4 female) volunteered to participate in this study. No participants reported a known history of neurological disorder or hearing loss. The study was reviewed and approved by the Institutional Review Board of Microsoft Research. A written consent was obtained upon participation.

The stimulus used in this study was a four-bar polyphonic piece composed of short melodic excerpts adapted from three popular songs (see Figure 5.19a). Each excerpt was assigned to a separate voice and instrument using MuseScore 3: vibraphone for "I'm yours" by Jason Mraz, piano for "Wherever you will go" by The Calling, and harmonica for "Forever young" by Alphaville. We hypothesized that using melodic excerpts from different songs for the three voices and assigning a different instrument to each voice would help listeners to pay attention to one voice at a time. The excerpts chosen followed the same chord progression (C major - G major - A minor - F major), which would ensure an overall pleasant listening experience.

Each excerpt consisted of four bars, for a total duration of 8 seconds. Besides the original excerpts (Standards), we generated oddball excerpts (Oddballs) by altering the second or the fourth bar of the Standards (Figure 5.19b). We created an oddball recognition task (see Sec. 5.5.3.2) using these stimuli to motivate participants to listen attentively. The excerpts were spatialized using a set of generic head-related transfer functions [40] to form three streams, where the perceived positions of vibraphone, harmonica, and piano were left, center, and right, respectively. The loudness of these streams was normalized using A-weighting, after which the streams were combined into polyphonic mixtures.



(a)

(b)

Figure 5.19: (a) Score sheet of the standard stimuli. (b) The 2nd or 4th bars of the standard stimuli were modified to create oddball bars, colored in red and blue.

5.5.3.2 Experiment

At the start of the experiment, the participants were asked to sit comfortably in front of a computer, read the instructions from the screen, and familiarize themselves with the stimuli. The experiment consisted of 28 trials for attention to vibraphone, 28 trials for attention to piano, and 14 trials for attention to harmonica. For this study, we only focused on generating binary outputs, i.e., distinguishing attention to vibraphone from attention to piano. The data from the attention to harmonica condition were only used for calculating the decoder (see Sec. 5.5.3.3) and as a sanity check. All trials were divided into 5 blocks with 14 trials per block, and their order was randomized for each participant. In the beginning of a trial, a left, right or up arrow, was presented on the screen as a visual cue (VC) to direct attention to the instrument on the left, right or center, respectively (see Figure 5.20). After a 1-second delay we played two repetitions of the music mixtures through Smartfones. In the stream to be attended, the first repetition was always a Standard, while the second repetition could be either a Standard or an Oddball. The task for participants was to identify whether the two repetitions were the same or different in the attended stream, and answer with a mouse click. Visual feedback (FB) was provided by a green dot displayed for a correct answer, or a red dot for an incorrect answer.

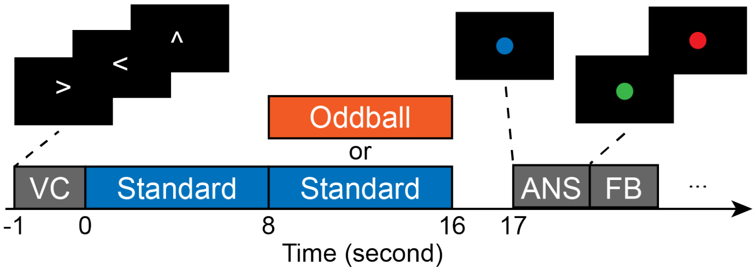


Figure 5.20: A trial started with a visual cue (VC) directing attention to the instrument on the left, right or center. It was followed by two music stimuli. Visual feedback (FB) was provided after an answer (ANS) was received.

5.5.3.3 Auditory attention decoding

EEG signals, sampled at 500 Hz, were passed through a Hamming windowed sinc FIR bandpass filter (2–8 Hz), and were split into epochs starting from the onset of each stimulus. Attention was decoded using AAD [109]. The envelopes of the individual voices in the stimuli (Figure 5.19) were extracted using the Hilbert Transform, and then lowpass filtered at 8 Hz and downsampled to 64 Hz to derive the stimulus feature $s(t)$ (Figure 5.21). The response feature, $r(t)$, was derived by downsampling the bandpass filtered EEG signals to 64 Hz. The AAD algorithm sought to find a decoder $g(\tau, n)$ that could linearly map $r(t)$ back to $s(t)$ [112] as:

$$\hat{s}(t) = \sum_{n} \sum_{\tau} r(t+\tau, n) g(\tau, n), \tag{5.1}$$

where $\hat{s}(t)$ is the reconstructed stimulus feature, n denotes the EEG channel index, and $0 \leq \tau \leq 600$ ms specifies a range of time-lags relative to the instantaneous occurrence of the stimulus feature, which is used to model the latency between a stimulus envelope and its corresponding envelope-following response in EEG signals. The decoder $g(\tau, n)$ is essentially a spatial-temporal filter that linearly transforms the EEG signals at time-lags τ from 0 to 600 ms post-stimulus to predict the corresponding auditory input. We can estimate $g(\tau, n)$ by minimizing the mean-square-error between the actual stimulus envelope $s(t)$ and the reconstructed envelope $\hat{s}(t)$ plus a regularization term:

$$\min \sum_t [s(t) - \hat{s}(t)]^2 + \lambda \sum_n \sum_\tau g(\tau, n)^2, \tag{5.2}$$

where λ is the regularization parameter set to avoid over-fitting. The optimal λ can be determined through cross-validation [113]. The decoder g can be computed using the following equation:

$$\mathbf{g} = (\mathbf{R^T R} + \lambda \mathbf{I})^{\mathbf{-1}} \mathbf{R^T s}, \tag{5.3}$$

where $\mathbf{R}$ is the matrix of response features $r(t)$ delayed by all possible values in τ with zero padding [112].

Auditory attention was decoded from each epoch. For each participant, we first pooled all epochs of EEG signals except for the one to be decoded to form the response feature $r(t)$. The envelopes of corresponding target voices were concatenated to form the stimulus feature $s(t)$. With the decoder g calculated via (5.3), we reconstructed a stimulus envelope $\hat{s}(t)$ using (5.1). We then correlated $\hat{s}(t)$ with the envelopes of each of the vibraphone, piano and harmonica voices in that epoch to generate three correlation coefficients using Pearson's correlation: ρ_{vibr}, ρ_{pian}, ρ_{harm}, respectively. We hypothesize that the correlation between the reconstruction and the envelope of the target instrument to be higher than the ones with the unattended instrument. To verify this hypothesis, we examined the difference between ρ_{vibr} and ρ_{pian} using a paired t-test. The Benjamini-Hochberg method was used to control the false discovery rate (FDR) in multiple comparisons (alpha = 0.05).

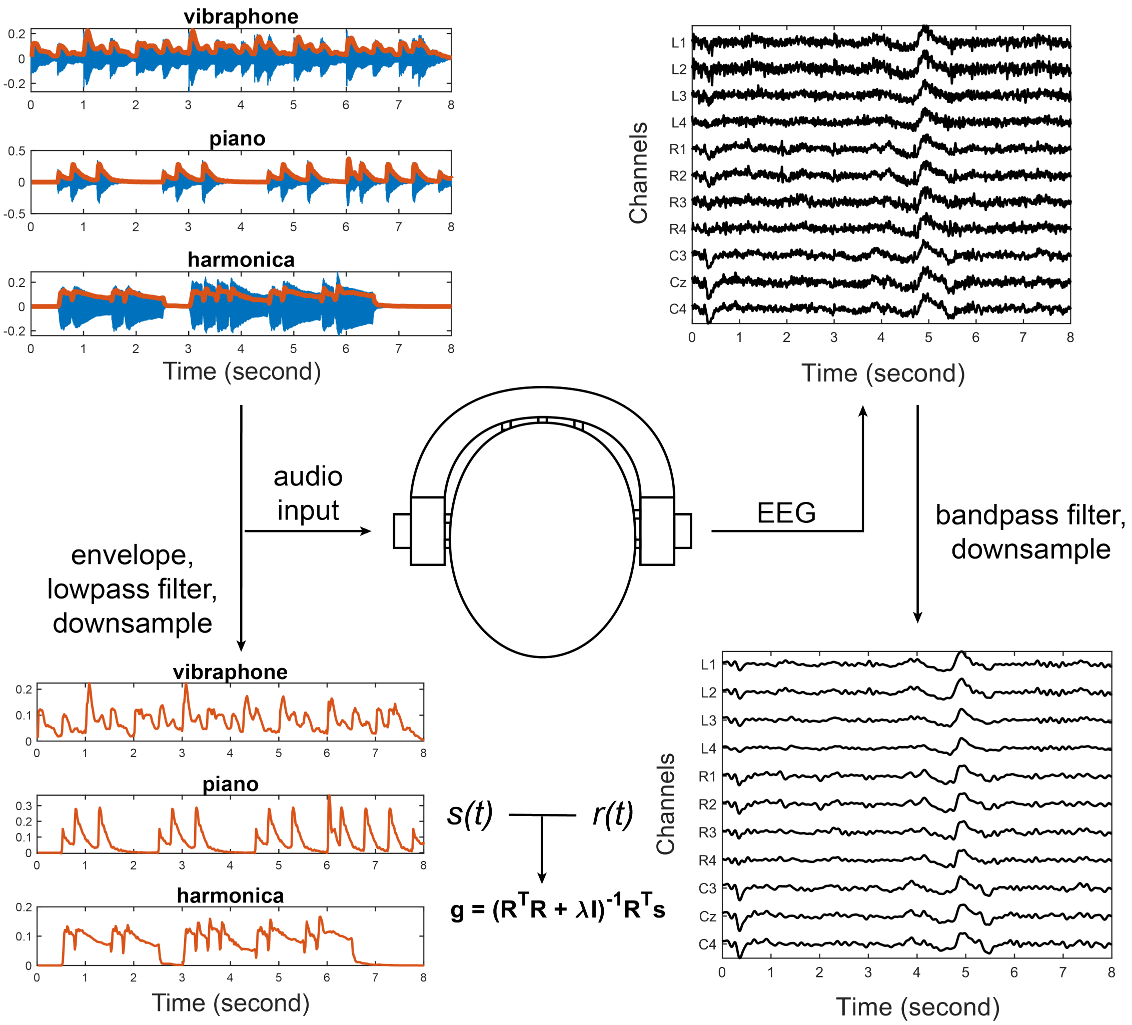


Figure 5.21: Illustration of the auditory attention decoding algorithm

5.5.3.4 Segment-based feature selection

The AAD method introduced in Sec. 5.5.3.3 uses an entire 8-second epoch to decode auditory attention. However, participants may not sustain their attention throughout the whole epoch, for example due to interference from a distracting stream, or due to the way they scheduled their attention to perform the task. During periods of reduced attention to the target instrument, the neural representation of the masking stimuli might interfere with or mask the target stimulus. We hypothesize that excluding data from periods of reduced attention may reduce noise and improve the overall decoding performance. We added a segment-based feature selection step to exclude irrelevant time segments from decoding.

After an epoch-specific decoder g was calculated, we applied it on multiple segments of EEG signals instead of the whole epoch. These segments were 2 s in duration with an overlap of 80%, resulting in a total of 18 segments per epoch. The strength of attention during each segment was estimated by comparing the strength of the correlation of the EEG with the vibraphone and the piano envelopes. Specifically, we calculated the absolute value of the segment-wise correlation difference (|SCD|, Figure 5.22a), defined as:

$$|\mathrm{SCD}_k| = |\rho_{\mathrm{vibr},k} - \rho_{\mathrm{pian},k}|, \tag{5.4}$$

where $k = \{1, \cdots, 18\}$ is the segment index, and $\rho_{\mathrm{vibr},k}$ or $\rho_{\mathrm{pian},k}$ represent the correlation between a segment-wise reconstruction with its corresponding segment-wise vibraphone envelope or piano envelope, respectively. If attention to either the vibraphone or piano is strong during a particular segment, the neural response for that segment should resemble the attended instrument voice more than the unattended one, i.e., |SCD| should be non-zero. During segments with reduced attention, |SCD| should approach zero.

Segments with small |SCD| values were excluded from analysis. The threshold was determined by the distribution of all |SCD| values in the training data (see Figure 5.22c). Values above the median of the distribution were retained (see Figure 5.22d). The correlations $\rho_{\mathrm{vibr},k}$ and $\rho_{\mathrm{pian},k}$ of surviving segments in each epoch were averaged to calculate ρ_{vibr} and ρ_{pian} after feature selection, respectively. A paired t-test was conducted to reveal any statistically significant change in these correlation measures with and without feature selection (alpha = 0.05, FDR corrected).

5.5.3.5 Classification

We used ρ_{vibr} and ρ_{pian} as the features to decode attention to vibraphone and attention to piano. We trained and tested a subject-specific linear support vector machine using leave-one-out cross-validation with 1000 repetitions. Classification was run on data with and without feature selection separately.

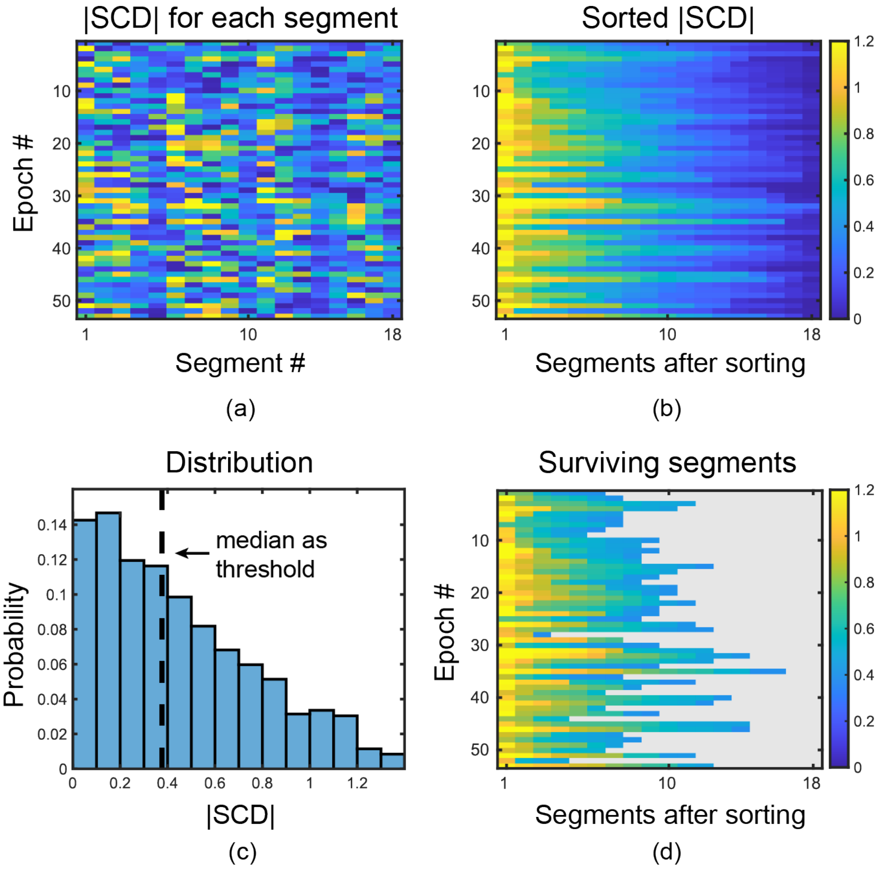


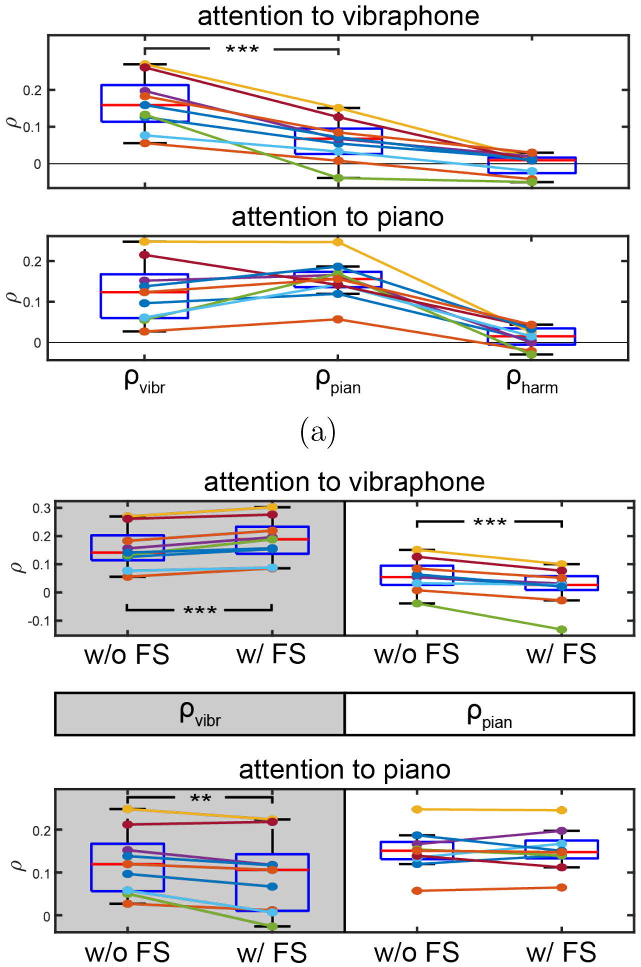
Figure 5.22: Illustration of the process of segment-based feature selection. (a) The correlation difference (SCD) was calculated for each segment in each epoch. (b) |SCD| sorted for each epoch (for visualization only). (c) The median of the distribution of all |SCD| values was used as the threshold. (d) The same figure as in (b), but with sub-threshold values masked by grey.

5.5.4 Results and Discussion

5.5.4.1 Correlation with envelopes

The correlation between the reconstructed envelope and the attended stimulus envelope is strongly modulated by attention, even without feature selection. When the participants were paying attention to the vibraphone, their average ρ_{vibr} was significantly higher than ρ_{pian} ($p<0.001$, Figure 5.23a). When attention was on the piano, ρ_{pian} was greater than ρ_{vibr}. However, the difference was not found to be statistically significant ($p=0.063$). In both conditions, ρ_{harm} was around 0 for all participants.

With the segment-based feature selection, the differences between ρ_{vibr} and ρ_{pian} were magnified. For the attention to vibraphone condition, feature selection significantly boosted ρ_{vibr} (p<0.001, Figure 5.23b) and suppressed ρ_{pian} (p<0.001). Similarly, ρ_{vibr} was suppressed by feature selection when attention was on piano (p=0.007), with ρ_{pian} statistically unchanged (p=0.906). We conclude that the proposed feature selection method identified segments relevant for the classifier to determine which instrument the participant paid attention to.



(a)

(b)

Figure 5.23: (a) Correlation between the reconstruction and the envelope of vibraphone (ρ_{vibr}), piano (ρ_{pian}) or harmonica (ρ_{harm}) without feature selection. Each line represents a subject. (b) Comparison of correlations with feature selection (w/ FS) and without (w/o FS). **, p< 0.01; ***, p< 0.001; FDR corrected for multiple comparisons.

5.5.4.2 Decoding accuracy

Experimental results indicate that the proposed method allows decoding of attention to music. Without feature selection, the average decoding accuracy was 63.77%, which is above the chance level (60.71%) with 95% confidence [93] (see Figure 5.24). We also observed great individual variability in the results, a known observation in many auditory BCI studies [62], [95].

The positive effect of feature selection on correlation measures (see Sec. 5.5.4.1) resulted in a boost in decoding accuracy. With segment-based feature selection, the average decoding accuracy improved to 71.23% (Figure 5.24), with a performance gain observed for all participants. Notably, this gain was more remarkable for subjects with a low decoding score before feature selection was implemented — the subjects with a decoding accuracy below 65% (Subject 4, 8, 2 and 1, see Figure 5.24) benefited an average of 11.0% from feature selection, which led to much smaller individual variability in the results. The decoding performance achieved in this study is comparable to previous works on auditory BCI using the same linear decoding method [110], [111], despite the use of a user-friendly EEG recording device with fewer sensors, less spatial coverage and lower signal-to-noise ratio compared to a conventional EEG cap. In addition, since we decoded attention with short data (8 seconds), the overall efficiency of the BCI system, evaluated by its information transfer rate (ITR) [99], is higher than similar studies with longer decoding windows (1.01 bit/min compared to ≤0.50 bits/min) [109], [111] (see Table 5.6).[1] One limitation of this study, however, is the small number of participants recruited (nine), which will be improved in follow-up studies in the future.

[1] Only results obtained from linear AAD were compared with results in this study. ITR was calculated based on the number of classes, sample length and decoding accuracy reported in these studies.

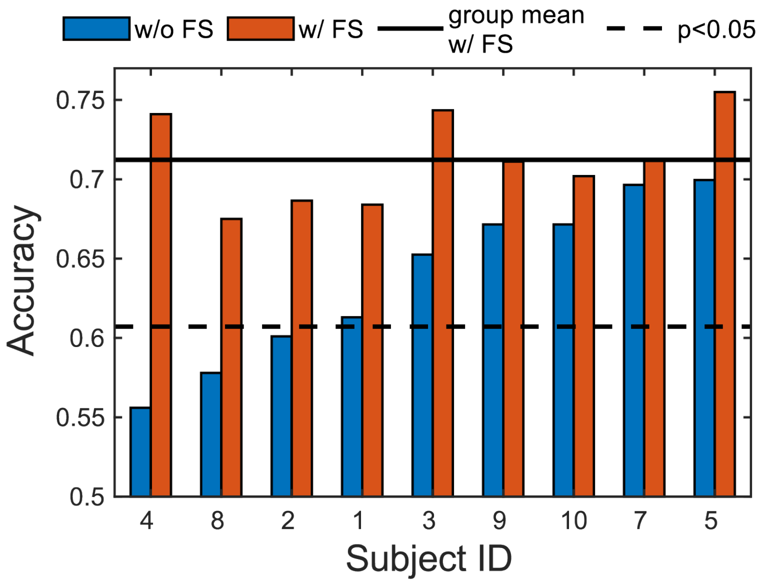


Figure 5.24: Decoding accuracy with feature selection (w/ FS) and without (w/o FS). The average for w/ FS is 71.23%. The subjects are sorted by their decoding accuracy w/o FS in ascending order.

Table 5.6: Comparison with previous studies using AAD

Study	**Sensors (#, type)**	**Sample length (s)**	**Accu. (%)**	**ITR (bits/min)**
O'Sullivan et al. [109]	128, gel	∼60	89.0	0.50
Ciccarelli et al. [111]	64, gel	10	66.0	0.45
	18, dry	10	59.0	0.14
here	11, saline	8	71.2	1.01

5.5.5 Conclusions

This study investigated the feasibility of building a user-friendly BCI system by decoding auditory attention. The proposed system relies on short musical stimuli with three voices. Due to its harmonic nature, this stimulus type may be more pleasant to listen to than previously proposed auditory stimuli like modulated pure tones or tone sequences and thus better suited for long-term use in a BCI system. Furthermore, the proposed system uses a compact headphone-based form factor with fewer sensors and requires much less effort in system setup than a traditional EEG system, which may be an appealing feature for novel users.

5.6 Summary of chapter

In this chapter, I showed four different studies that demonstrated means to improve various aspects of an auditory BCI design. The study in Section 5.2 explored the feasibility of using short, human-voiced syllables as stimuli, and decoding attention using data as short as 1.5 seconds. This method effectively boosted the throughput of the system, which is substantially greater than most auditory BCI systems previously reported. In a follow-up study, an even better result was achieved in Section 5.3, which adopted the same stimuli and experimental design, but employed a CNN over SVM for classification. The use of CNN offloads the burden of feature engineering, which is usually time-consuming and difficult. Moreover, the features that a CNN automatically learns to distinguish different classes are oftentimes more optimal than a hand-crafted one. In Section 5.4, I attempted the idea of designing an attention task with a sequence of tones, aiming at improving the user-friendliness of the BCI system. I furthered my exploration in this dimension in Section 5.5, where I played polyphonic music to the participants, and decoded their attention to instruments from EEG signals that were recorded from a headphone-like EEG recording device. The BCI system proposed in this study is unobtrusive and easy to setup, and the users can actually enjoy the stimuli while using the system. Collectively, these studies show that attention-based BCI is a concept with great potential and practical values in real-life.

Chapter 6

Summary and Conclusions

6.1 Summary of dissertation

In this dissertation, I aimed to achieve two major goals. **First**, I sought to decode auditory attention from EEG and fMRI signals, and study the neural representation of attentional control across space (i.e., brain regions) and time. I accomplished this goal by designing a condition-rich experiment that requires spatial or non-spatial auditory attention from listeners, and adopting a representational similarity analysis (RSA) framework to investigate the neural representation of auditory attention in EEG and fMRI. Then, I used neural representation features to fuse EEG and fMRI, which unveiled the information flow during the attention task with fine spatial and temporal resolution. **Second**, I sought to decode auditory attention from single-trial EEG signals for the design of an auditory brain-computer interface (BCI) system. I proposed several methods to improve the communication efficiency and user-friendliness of an auditory BCI design, and achieved promising results.

In **Chapter 2**, I designed a experiment with multiple conditions and recorded EEG signals while the listeners participated in an auditory attention task. I extracted representational dissimilarity features from the EEG time course and alpha oscillation power, and compared these features with ideal conceptual models or behavioral performance. I identified time intervals in which particular contrasts in attentional state, such as the difference between attention types or between attention to different locations, have strong representation in the EEG time course or in its alpha power. I also revealed that the listener's behavioral performance in the attention task is significantly and positively correlated with the P2 amplitude evoked by the target syllable.

In **Chapter 3**, I used the same experimental design and subjects as in **Chapter 2**, and recorded fMRI data while the listeners participated in the auditory attention task. I identified an extended attention network, in which individual brain regions show different specialization for spatial or non-spatial attention. I also extracted representational dissimilarity features from each voxel, and compared these features with ideal conceptual models or behavioral performance. I identified the medial occipital lobe as the region actively encoding the spatial information of auditory attention; the right IFS is the sole region that encodes information about the gender / pitch of the attended talker. The neural representations within the parietal regions are correlated with behavioral performance, demonstrating their important role in spatially demanding tasks.

In **Chapter 4**, I deployed a representational similarity analysis (RSA) for multimodal data fusion to study the dynamics of auditory attentional control. I correlated the representation features acquired in **Chapter 2** and **Chapter 3** to search for significant information correspondence in time and space. The fusion analysis revealed that the calcarine sulcus is only active during the task when spatial mapping of sound is needed, suggesting a major role in processing spatialized auditory targets. We observed a difference in dynamics between the inferior frontal sulcus and the superior precentral sulcus during the stimulus period, which might reflect a difference in the suppression mechanism between spatial and non-spatial attention.

In **Chapter 5**, I shifted my focus to real-life applications and reported four different studies demonstrating possible ways to improve various aspects of an auditory BCI design. The study in Section 5.2 explored the feasibility of using short, human-voiced syllables as stimuli, and decoding attention using data as short as 1.5 seconds. This method effectively boosted the throughput of the system, which is substantially greater than most auditory BCI systems previously reported. In a follow-up study, an even better result was achieved in Section 5.3, which adopted the same stimuli and experimental design, but employed a CNN over SVM for classification. The use of CNN offloads the burden of feature engineering, which is usually time-consuming and difficult. Moreover, the features that a CNN automatically learns to distinguish different classes are often more optimal than a hand-crafted one. In Section 5.4, I designed an attention task with a sequence of tones, aiming at improving the user-friendliness of the BCI system. I furthered my exploration in this dimension in Section 5.5, where I played polyphonic music to the participants, and decoded their attention to instruments from EEG signals that were recorded from a headphone-like EEG recording device. The BCI system proposed in this study is unobtrusive and easy to setup, and the users can actually enjoy the stimuli while using the system. Collectively, these studies show that attention-based BCI is a concept with great potential and practical values in real-life.

6.2 Significance and future directions

This dissertation presents one of the first few pioneering works to adopt RSA in the study of attentional states. It also proposes new ways to improve the design of an attention-based auditory BCI system. It enriches our understanding of how the nervous system functions to form auditory attention, and sheds light on how we can leverage this knowledge to develop real-life assistive applications.

6.2.1 Auditory selective attention

This dissertation demonstrates one possible way to study the neural representation of auditory attention. Previous studies mainly focused on learning the outcome or impact of attentional control (i.e., the “process”), through a direct comparison of neural signals between conditions [9]. The “representation” of this cognitive function, which is how information is robustly encoded in a brain location or time, however, is less explored. “Process” and “representation” are both important neural constructs that can unveil how cognitive control arises in mind and brain, and they can be studied through either classical analyses (e.g., ERP analysis, time-frequency analysis, general linear model) or RSA, respectively. Here, I conducted RSA alongside classical analyses. Through the comparison of their results, I showed how these two approaches may converge or diverge in different scenarios, which could be an important reference for future studies in this field.

One possible way to extend this study is to include information of more induced EEG oscillations in addition to the alpha band power. Previous studies suggest that EEG signals in the gamma (> 30 Hz) [114]–[116], beta ($14 - 30$ Hz) [117] and theta ($4 - 8$ Hz) [118], [119] band are modulated by auditory selective attention in different manners. Including features from all these bands into analysis would expand the search space for important information encoding.

6.2.2 RSA and multimodal data fusion

In this dissertation, I explored if / how we can fuse EEG with fMRI through the use of neural representation features. To date, only a handful of studies have attempted to use RSA for multimodal data fusion [18], [43], and none of them investigated auditory attentional state. Results here, for the first time, suggest that induced oscillations are as important as evoked responses in studying the internal state with a multimodal data fusion approach. EEG time courses and oscillations reflect brain activities of different sources (i.e., stimulus and non-stimulus driven), and should both be considered for the study of neural representation. Unsurprisingly, this study also demonstrates that the number of conditions in the experimental design plays a crucial role in controlling noise and the false positive rate in RSA discoveries. Randomness in correlation decreases as the dimension of an RDM increases, which is why a greater number of conditions is always preferred in an RSA study. Examining the statistical power of an effect with a non-parametric, cluster-based permutation test could be an option to control the family-wise error rate, but it cannot help if the effect of interest has similar or even less power than the noise.

6.2.3 Brain-computer interface

This dissertation demonstrates several ways to improve the throughput and user-friendliness of an attention-based auditory BCI system. The results suggest that using short stimulus for input and an artificial neural network for decoding is undoubtedly the combination that future studies should consider, if communication efficiency is the major concern. A more important question that this dissertation attempts to answer is what value an auditory BCI can offer to its users. In my opinion, a BCI would be better accepted if it requires short preparation time, little training and minimal extra hardware. One such example is introduced in Section 5.5, where the EEG electrodes are integrated into a pair of headphones. Users can simply put on the device and enjoy some music while using this BCI system. Such design brings an intuitive application to this device — a mind-steered music controller, with which the listener can choose to play next / last song, or adjust the volume without any movement. Therefore, it is offering convenience to the user without requesting much extra effort. Another idea is to embed a BCI system into a virtual reality (VR) or augmented reality (AR) device. Similar to the previous example, the users need to wear a headset when they are using the VR/AR system, which could be a convenient site for recording EEG signals. Future studies should consider how we can integrate a BCI system with existing wearable devices, and provide practical value to their users.

Bibliography

[1] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, Apr. 2008. DOI: 10.1016/j.tics.2008.02.003.

[2] A. Mihali, A. G. Young, L. A. Adler, M. M. Halassa, and W. J. Ma, "A Low-Level Perceptual Correlate of Behavioral and Clinical Deficits in ADHD," *Computational Psychiatry*, vol. 2, pp. 141–163, Oct. 2018. DOI: 10.1162/cpsy_a_00018.

[3] J. R. Booth, D. D. Burman, J. R. Meyer, Z. Lei, B. L. Trommer, N. D. Davenport, W. Li, T. B. Parrish, D. R. Gitelman, and M. M. Mesulam, "Larger deficits in brain networks for response inhibition than for visual selective attention in attention deficit hyperactivity disorder (ADHD)," *Journal of Child Psychology and Psychiatry*, vol. 46, no. 1, pp. 94–111, Jan. 2005. DOI: 10.1111/J.1469-7610.2004.00337.X.

[4] E. V. Orekhova and T. A. Stroganova, "Arousal and attention re-orienting in autism spectrum disorders: evidence from auditory event-related potentials," *Frontiers in Human Neuroscience*, vol. 8, no. 34, pp. 1–17, Feb. 2014. DOI: 10.3389/FNHUM.2014.00034.

[5] S. J. Lalani, T. C. Duffield, H. G. Trontel, E. D. Bigler, T. J. Abildskov, A. Froehlich, M. B. Prigge, B. G. Travers, J. S. Anderson, B. A. Zielinski, A. Alexander, N. Lange, and J. E. Lainhart, "Auditory Attention in Autism Spectrum Disorder: An Exploration of Volumetric MRI Findings," *Journal of clinical and experimental neuropsy-*

chology, vol. 40, no. 5, pp. 502–517, May 2018. DOI: 10.1080/13803395.2017.1373746.

[6] J. K. Bizley and Y. E. Cohen, "The what, where and how of auditory-object perception," *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, Oct. 2013. DOI: 10.1038/nrn3565.

[7] T. Kirschstein and R. Köhling, "What is the Source of the EEG?" *Clinical EEG and Neuroscience*, vol. 40, no. 3, pp. 146–149, Jul. 2009. DOI: 10.1177/155005940904000305.

[8] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, no. 6843, pp. 150–157, Jul. 2001. DOI: 10.1038/35084005.

[9] M. C. Freund, J. A. Etzel, and T. S. Braver, "Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach," *Trends in Cognitive Sciences*, vol. 25, no. 7, pp. 622–638, Jul. 2021. DOI: 10.1016/J.TICS.2021.03.011.

[10] C. R. Decharms and A. Zador, "Neural Representation and the Cortical Code," *Annual Review of Neuroscience*, vol. 23, pp. 613–647, Mar. 2000. DOI: doi.org/10.1146/annurev.neuro.23.1.613.

[11] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis - connecting the branches of systems neuroscience.," *Frontiers in systems neuroscience*, vol. 2, no. 4, pp. 1–28, Nov. 2008. DOI: 10.3389/neuro.06.004.2008.

[12] R. Cecere, J. Gross, A. Willis, and G. Thut, "Being First Matters: Topographical Representational Similarity Analysis of ERP Signals Reveals Separate Networks for Audiovisual Temporal Binding Depending on the Leading Sense," *The Journal of Neuroscience*, vol. 37, no. 21, pp. 5274–5287, 2017. DOI: 10.1523/JNEUROSCI.2926-16.2017.

[13] B. Kaneshiro, M. P. Guimaraes, H.-S. Kim, A. M. Norcia, and P. Suppes, “A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification,” *PLOS ONE*, vol. 10, no. 8, e0135697, Aug. 2015. DOI: 10.1371/JOURNAL.PONE.0135697.

[14] V. R. Sommer, Y. Fandakova, T. H. Grandy, Y. L. Shing, M. Werkle-Bergner, and M. C. Sander, “Neural Pattern Similarity Differentially Relates to Memory Performance in Younger and Older Adults,” *The Journal of Neuroscience*, vol. 39, no. 41, pp. 8089–8099, Oct. 2019. DOI: 10.1523/JNEUROSCI.0197-19.2019.

[15] B. J. Devereux, A. Clarke, A. Marouchos, and L. K. Tyler, “Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects,” *Journal of Neuroscience*, vol. 33, no. 48, pp. 18 906–18 916, Nov. 2013. DOI: 10.1523/JNEUROSCI.3809-13.2013.

[16] F. Carota, N. Kriegeskorte, H. Nili, and F. Pulvermüller, “Representational Similarity Mapping of Distributional Semantics in Left Inferior Frontal, Middle Temporal, and Motor Cortex,” *Cerebral Cortex*, vol. 27, no. 1, pp. 294–309, Jan. 2017. DOI: 10.1093/cercor/bhw379.

[17] L. Zhao, C. Chen, L. Shao, Y. Wang, X. Xiao, C. Chen, J. Yang, J. Zevin, and G. Xue, “Orthographic and phonological representations in the fusiform cortex,” *Cerebral Cortex*, vol. 27, no. 11, pp. 5197–5210, Nov. 2016. DOI: 10.1093/cercor/bhw300.

[18] R. M. Cichy, D. Pantazis, and A. Oliva, “Resolving human object recognition in space and time.,” *Nature Neuroscience*, vol. 17, no. 3, pp. 455–62, Jan. 2014. DOI: 10.1038/nn.3635.

[19] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, “Brain-computer interfaces for communication and rehabilitation,” *Nature Reviews Neurology*, vol. 12, no. 9, pp. 513–525, Aug. 2016. DOI: 10.1038/nrneurol.2016.113.

[20] E. W. Sellers, D. B. Ryan, and C. K. Hauser, “Noninvasive brain-computer interface enables communication after brainstem stroke,” *Science Translational Medicine*, vol. 6, no. 257, 257re7, Oct. 2014. DOI: 10.1126/SCITRANSLMED.3007801.

[21] F. Piccione, F. Giorgi, P. Tonin, K. Priftis, S. Giove, S. Silvoni, G. Palmas, and F. Beverina, “P300-based brain computer interface: Reliability and performance in healthy and paralysed participants,” *Clinical Neurophysiology*, vol. 117, no. 3, pp. 531–537, Mar. 2006. DOI: 10.1016/J.CLINPH.2005.07.024.

[22] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, “High-performance neuroprosthetic control by an individual with tetraplegia,” *The Lancet*, vol. 381, no. 9866, pp. 557–564, Feb. 2013. DOI: 10.1016/S0140-6736(12)61816-9.

[23] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, “Reach and grasp by people with tetraplegia using a neurally controlled robotic arm,” *Nature*, vol. 485, no. 7398, pp. 372–375, May 2012. DOI: 10.1038/nature11076.

[24] E. R. Buch, A. Modir Shanechi, A. D. Fourkas, C. Weber, N. Birbaumer, and L. G. Cohen, “Parietofrontal integrity determines neural modulation associated with grasping imagery after stroke,” *Brain*, vol. 135, no. 2, pp. 596–614, Feb. 2012. DOI: 10.1093/BRAIN/AWR331.

[25] S. L. Wolf, C. J. Winstein, J. P. Miller, E. Taub, G. Uswatte, D. Morris, C. Giuliani, K. E. Light, D. Nichols-Larsen, and E. Investigators, “Effect of Constraint-Induced Movement Therapy on Upper Extremity Function 3 to 9 Months After Stroke: The EXCITE Randomized Clinical Trial,” *The Journal of the American Medical Association*, vol. 296, no. 17, pp. 2095–2104, Nov. 2006. DOI: 10.1001/JAMA.296.17.2095.

[26] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain-computer interface paradigms," *Journal of Neural Engineering*, vol. 16, no. 1, p. 011 001, Feb. 2019. DOI: 10.1088/1741-2552/aaf12e.

[27] J. Chen, D. Zhang, A. K. Engel, Q. Gong, and A. Maye, "Application of a single-flicker online SSVEP BCI for spatial navigation," *PLoS ONE*, vol. 12, no. 5, e0178385, May 2017. DOI: 10.1371/journal.pone.0178385.

[28] D. W. Kim, J. H. Cho, H. J. Hwang, J. H. Lim, and C. H. Im, "A vision-free brain-computer interface (BCI) paradigm based on auditory selective attention," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 3684–3687, 2011. DOI: 10.1109/IEMBS.2011.6090623.

[29] H. J. Baek, M. H. Chang, J. Heo, and K. S. Park, "Enhancing the usability of brain-computer interface systems," *Computational Intelligence and Neuroscience*, vol. 2019, p. 5 427 154, 2019. DOI: 10.1155/2019/5427154.

[30] M. Huang, J. Jin, Y. Zhang, D. Hu, and X. Wang, "Usage of drip drops as stimuli in an auditory P300 BCI paradigm," *Cognitive Neurodynamics*, vol. 12, no. 1, pp. 85–94, Apr. 2018. DOI: 10.1007/s11571-017-9456-y.

[31] M. Arvaneh, I. H. Robertson, and T. E. Ward, "A P300-Based Brain-Computer Interface for Improving Attention," *Frontiers in Human Neuroscience*, vol. 12, no. 524, pp. 1–14, Jan. 2019. DOI: 10.3389/FNHUM.2018.00524.

[32] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification," *Journal of Neural Engineering*, vol. 11, no. 2, p. 026 009, Apr. 2014. DOI: 10.1088/1741-2560/11/2/026009.

[33] Z. Lin, C. Zhang, Y. Zeng, L. Tong, and B. Yan, "A novel P300 BCI speller based on the Triple RSVP paradigm," *Scientific Reports*, vol. 8, no. 3350, pp. 1–9, Dec. 2018. DOI: 10.1038/s41598-018-21717-y.

[34] C. S. Herrmann and R. T. Knight, “Mechanisms of human attention: event-related potentials and oscillations,” *Neuroscience & Biobehavioral Reviews*, vol. 25, no. 6, pp. 465–476, Aug. 2001. DOI: `10.1016/S0149-7634(01)00027-6`.

[35] M. Siegel, T. H. Donner, and A. K. Engel, “Spectral fingerprints of large-scale neuronal interactions,” *Nature Reviews Neuroscience*, vol. 13, no. 2, pp. 121–134, Jan. 2012. DOI: `10.1038/nrn3137`.

[36] I. Choi, L. Wang, H. Bharadwaj, and B. Shinn-Cunningham, “Individual differences in attentional modulation of cortical responses correlate with selective attention performance,” *Hearing Research*, vol. 314, pp. 10–19, Aug. 2014. DOI: `10.1016/j.heares.2014.04.008`.

[37] R. J. Giuliano, C. M. Karns, H. J. Neville, and S. A. Hillyard, “Early auditory evoked potential is modulated by selective attention and related to individual differences in visual working memory capacity,” *Journal of Cognitive Neuroscience*, vol. 26, no. 12, pp. 2682–2690, Dec. 2014. DOI: `10.1162/jocn_a_00684`.

[38] Y. Deng, I. Choi, and B. Shinn-Cunningham, “Topographic specificity of alpha power during auditory spatial attention,” *NeuroImage*, vol. 207, p. 116 360, Feb. 2020. DOI: `10.1016/j.neuroimage.2019.116360`.

[39] Y. Deng, R. M. Reinhart, I. Choi, and B. G. Shinn-Cunningham, “Causal links between parietal alpha activity and spatial auditory attention,” *eLife*, vol. 8, e51184, Nov. 2019. DOI: `10.7554/eLife.51184`.

[40] B. Gardner and K. Martin, “HRTF Measurements of a KEMAR Dummy-Head Microphone MIT Media Lab Perceptual Computing-Technical Report #280,” Media Lab, MIT, Tech. Rep., 1994.

[41] A. Delorme and S. Makeig, “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” en, *Journal of Neu-*

roscience Methods, vol. 134, no. 1, pp. 9–21, Mar. 2004. DOI: 10.1016/j.jneumeth.2003.10.009.

[42] R. M. Cichy, A. Khosla, D. Pantazis, and A. Torralba, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Scientific reports*, vol. 6, no. 27755, pp. 1–35, Jun. 2016. DOI: doi.org/10.1038/srep27755.

[43] V. Salmela, E. Salo, J. Salmi, and K. Alho, "Spatiotemporal Dynamics of Attention Networks Revealed by Representational Similarity Analysis of EEG and fMRI," *Cerebral Cortex*, vol. 28, no. 2, pp. 549–560, 2016. DOI: 10.1093/cercor/bhw389.

[44] W. W. An, K. H. Ting, I. P. Au, J. H. Zhang, Z. Y. Chan, I. S. Davis, W. K. So, R. H. Chan, and R. T. Cheung, "Neurophysiological Correlates of Gait Retraining with Real-Time Visual and Auditory Feedback," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1341–1349, Jun. 2019. DOI: 10.1109/TNSRE.2019.2914187.

[45] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177–190, Aug. 2007. DOI: 10.1016/j.jneumeth.2007.03.024.

[46] G. A. Rousselet, J. S. Husk, P. J. Bennett, and A. B. Sekuler, "Time course and robustness of ERP object and face differences," *Journal of Vision*, vol. 8, no. 12, pp. 1–18, Sep. 2008. DOI: 10.1167/8.12.3.

[47] A. C. Connolly, J. S. Guntupalli, J. Gors, M. Hanke, Y. O. Halchenko, Y.-C. Wu, H. Abdi, and J. V. Haxby, "The Representation of Biological Classes in the Human Brain," *Journal of Neuroscience*, vol. 32, no. 8, pp. 2608–2618, Feb. 2012. DOI: 10.1523/JNEUROSCI.5547-11.2012.

[48] R. M. Cichy and S. Teng, "Resolving the neural dynamics of visual and auditory scene processing in the human brain: a methodological approach," *Philosophical Transac-*

tions of the Royal Society B: Biological Sciences, vol. 372, no. 1714, p. 20 160 108, Feb. 2017. DOI: 10.1098/rstb.2016.0108.

[49] M. X. Lowe, Y. Mohsenzadeh, B. Lahner, I. Charest, A. Oliva, and S. Teng, "Spatiotemporal Dynamics of Sound Representations Reveal a Hierarchical Progression of Category Selectivity," *bioRxiv*, p. 2020.06.12.149120, 2020. DOI: 10.1101/2020.06.12.149120.

[50] A. Walther, H. Nili, N. Ejaz, A. Alink, N. Kriegeskorte, and J. Diedrichsen, "Reliability of dissimilarity measures for multi-voxel pattern analysis," *NeuroImage*, vol. 137, pp. 188–200, Aug. 2016. DOI: 10.1016/J.NEUROIMAGE.2015.12.012.

[51] H. Popal, Y. Wang, and I. R. Olson, "A Guide to Representational Similarity Analysis for Social Neuroscience," *Social Cognitive and Affective Neuroscience*, vol. 14, no. 11, pp. 1243–1253, Nov. 2019. DOI: 10.1093/SCAN/NSZ099.

[52] L. M. Bonacci, S. Bressler, J. A. Kwasa, A. L. Noyce, and B. G. Shinn-Cunningham, "Effects of Visual Scene Complexity on Neural Signatures of Spatial Attention," *Frontiers in Human Neuroscience*, vol. 14, no. 91, pp. 1–15, Mar. 2020. DOI: 10.3389/fnhum.2020.00091.

[53] G. F. Woodman, "A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention," *Attention, perception psychophysics*, vol. 72, no. 8, pp. 2031–2046, Nov. 2010. DOI: 10.3758/APP.72.8.2031.

[54] M. Seppänen, J. Hämäläinen, A.-K. Pesonen, and M. Tervaniemi, "Music Training Enhances Rapid Neural Plasticity of N1 and P2 Source Activation for Unattended Sounds," *Frontiers in Human Neuroscience*, vol. 6, no. 43, pp. 1–13, Mar. 2012. DOI: 10.3389/FNHUM.2012.00043.

[55] K. Tremblay, B. Ross, K. Inoue, K. McClannahan, and G. Collet, "Is the auditory evoked P2 response a biomarker of learning?" *Frontiers in Systems Neuroscience*, vol. 8, no. 28, pp. 1–13, Feb. 2014. DOI: 10.3389/FNSYS.2014.00028.

[56] Y. Liu, D. Zhang, J. Ma, D. Li, H. Yin, and Y. Luo, "The Attention Modulation on Timing: An Event-Related Potential Study," *PLOS ONE*, vol. 8, no. 6, e66190, Jun. 2013. DOI: 10.1371/JOURNAL.PONE.0066190.

[57] J. De Boer and K. Krumbholz, "Auditory Attention Causes Gain Enhancement and Frequency Sharpening at Successive Stages of Cortical Processing-Evidence from Human Electroencephalography," *Journal of Cognitive Neuroscience*, vol. 30, no. 6, pp. 785–798, Jun. 2018. DOI: 10.1162/jocn_a_01245.

[58] K. T. Hill and L. M. Miller, "Auditory attentional control and selection during cocktail party listening," *Cerebral Cortex*, vol. 20, no. 3, pp. 583–590, 2010. DOI: 10.1093/cercor/bhp124.

[59] S. W. Michalka, L. Kong, M. L. Rosen, B. G. Shinn-Cunningham, and D. C. Somers, "Short-Term Memory for Space and Time Flexibly Recruit Complementary Sensory-Biased Frontal Lobe Attention Networks," *Neuron*, vol. 87, no. 4, pp. 882–892, Aug. 2015. DOI: 10.1016/j.neuron.2015.07.028.

[60] D. H. Brainard, "The Psychophysics Toolbox," *Spatial Vision*, vol. 10, no. 4, pp. 433–436, Jan. 1997. DOI: 10.1163/156856897X00357.

[61] D. G. Pelli, "The VideoToolbox software for visual psychophysics: transforming numbers into movies," *Spatial Vision*, vol. 10, no. 4, pp. 437–442, Jan. 1997. DOI: 10.1163/156856897X00366.

[62] W. W. An, A. Pei, A. L. Noyce, and B. Shinn-cunningham, "Decoding auditory attention from single-trial EEG for a high-efficiency brain-computer interface," *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3456–3459, 2020.

[63] W. Penny, K. Friston, J. Ashburner, S. Kiebel, and T. Nichols, "Statistical Parametric Mapping: The Analysis of Functional Brain Images," *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, 2007.

[64] J. Ashburner, G. Barnes, C.-C. Chen, J. Daunizeau, G. Flandin, K. Friston, A. Jafarian, S. Kiebel, J. Kilner, V. Litvak, R. Moran, W. P. Adeel, R. Klaas, S. Sungho, T. P. Zeidman, D. Gitelman, R. Henson, C. Hutton, V. Glauche, J. Mattout, and C. Phillips, *SPM12 Manual*. 2020.

[65] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the Human Connectome Project," *NeuroImage*, vol. 80, pp. 105–124, Oct. 2013. DOI: 10.1016/j.neuroimage.2013.04.127.

[66] S. M. Smith and T. E. Nichols, "Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference," *NeuroImage*, vol. 44, no. 1, pp. 83–98, Jan. 2009. DOI: 10.1016/j.neuroimage.2008.03.061.

[67] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, vol. 8, no. 14, pp. 1–10, Feb. 2014. DOI: 10.3389/FNINF.2014.00014.

[68] C. Destrieux, B. Fischl, A. Dale, and E. Halgren, "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature," *NeuroImage*, vol. 53, no. 1, pp. 1–15, Oct. 2010. DOI: 10.1016/j.neuroimage.2010.06.010.

[69] K. Alho, J. Salmi, S. Koistinen, O. Salonen, and T. Rinne, "Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks," *Brain Research*, vol. 1626, pp. 136–145, Nov. 2015. DOI: 10.1016/j.brainres.2014.12.050.

[70] L. Kong, S. W. Michalka, M. L. Rosen, S. L. Sheremata, J. D. Swisher, B. G. Shinn-Cunningham, and D. C. Somers, "Auditory spatial attention representations in the

human cerebral cortex," *Cerebral Cortex*, vol. 24, no. 3, pp. 773–784, Mar. 2014. DOI: 10.1093/cercor/bhs359.

[71] J. P. Rauschecker and S. K. Scott, "Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing," *Nature Neuroscience 2009 12:6*, vol. 12, no. 6, pp. 718–724, May 2009. DOI: 10.1038/nn.2331.

[72] A. Degerman, T. Rinne, J. Salmi, O. Salonen, and K. Alho, "Selective attention to sound location or pitch studied with fMRI," *Brain Research*, vol. 1077, no. 1, pp. 123–134, Mar. 2006. DOI: 10.1016/J.BRAINRES.2006.01.025.

[73] M. Assem, S. Shashidhara, M. F. Glasser, and J. Duncan, "Precise topology of adjacent domain-general and sensory-biased regions in the human brain," *bioRxiv*, 2021. DOI: 10.1101/2021.02.21.431622.

[74] M. A. Silver and S. Kastner, "Topographic maps in human frontal and parietal cortex," *Trends in Cognitive Sciences*, vol. 13, no. 11, pp. 488–495, Nov. 2009. DOI: 10.1016/J.TICS.2009.08.005.

[75] J. D. Swisher, M. A. Halko, L. B. Merabet, S. A. McMains, and D. C. Somers, "Visual Topography of Human Intraparietal Sulcus," *Journal of Neuroscience*, vol. 27, no. 20, pp. 5326–5337, May 2007. DOI: 10.1523/JNEUROSCI.0991-07.2007.

[76] B. A. Wandell, S. O. Dumoulin, and A. A. Brewer, "Visual Field Maps in Human Cortex," *Neuron*, vol. 56, no. 2, pp. 366–383, Oct. 2007. DOI: 10.1016/J.NEURON.2007.10.012.

[77] S. W. Michalka, M. L. Rosen, L. Kong, B. G. Shinn-Cunningham, and D. C. Somers, "Auditory Spatial Coding Flexibly Recruits Anterior, but Not Posterior, Visuotopic Parietal Cortex," *Cerebral Cortex*, vol. 26, no. 3, pp. 1302–1308, Mar. 2016. DOI: 10.1093/cercor/bhv303.

[78] R. J. Zatorre, M. Bouffard, P. Ahad, and P. Belin, “Where is ’where’ in the human auditory cortex?” *Nature Neuroscience*, vol. 5, no. 9, pp. 905–909, Aug. 2002. DOI: 10.1038/nn904.

[79] C. Alain, D. Shen, H. Yu, and C. Grady, “Dissociable Memory- and Response-Related Activity in Parietal Cortex During Auditory Spatial Working Memory,” *Frontiers in Psychology*, vol. 1, no. 202, pp. 1–11, Dec. 2010. DOI: 10.3389/FPSYG.2010.00202.

[80] M. Corbetta, “Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neuralsystems?” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 3, p. 831, Feb. 1998. DOI: 10.1073/PNAS.95.3.831.

[81] M. Corbetta, E. Akbudak, T. E. Conturo, A. Z. Snyder, J. M. Ollinger, H. A. Drury, M. R. Linenweber, S. E. Petersen, M. E. Raichle, D. C. V. Essen, and G. L. Shulman, “A Common Network of Functional Areas for Attention and Eye Movements,” *Neuron*, vol. 21, no. 4, pp. 761–773, Oct. 1998. DOI: 10.1016/S0896-6273(00)80593-0.

[82] H.-O. Karnath, S. Ferber, and M. Himmelbach, “Spatial awareness is a function of the temporal not the posterior parietal lobe,” *Nature*, vol. 411, no. 6840, pp. 950–953, Jun. 2001. DOI: 10.1038/35082075.

[83] G. Vallar, “Extrapersonal Visual Unilateral Spatial Neglect and Its Neuroanatomy,” *NeuroImage*, vol. 14, no. 1, S52–S58, Jul. 2001. DOI: 10.1006/NIMG.2001.0822.

[84] K. Shapiro and A. P. Hillstrom, “Control of Visuotemporal Attention by Inferior Parietal and Superior Temporal Cortex,” *Current Biology*, vol. 12, no. 15, pp. 1320–1325, Aug. 2002. DOI: 10.1016/S0960-9822(02)01040-0.

[85] J. J. Foxe and A. C. Snyder, “The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention,” *Frontiers in Psychology*, vol. 2, no. 154, pp. 1–13, Jul. 2011. DOI: 10.3389/FPSYG.2011.00154.

[86] A. Garg, D. Schwartz, and A. A. Stevens, "Orienting Auditory Spatial Attention Engages Frontal Eye Fields and Medial Occipital Cortex in Congenitally Blind Humans," *Neuropsychologia*, vol. 45, no. 10, pp. 2307–2321, Jun. 2007. DOI: 10.1016/J.NEUROPSYCHOLOGIA.2007.02.015.

[87] O. Collignon, G. Vandewalle, P. Voss, G. Albouy, G. Charbonneau, M. Lassonde, and F. Lepore, "Functional specialization for auditory–spatial processing in the occipital cortex of congenitally blind humans," *Proceedings of the National Academy of Sciences*, vol. 108, no. 11, pp. 4435–4440, Mar. 2011. DOI: 10.1073/PNAS.1013928108.

[88] W. W. An, A. Pei, A. L. Noyce, and B. Shinn-Cunningham, "Decoding auditory attention from EEG using a convolutional neural network," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.

[89] W. W. An, H. Si-Mohammed, N. Huang, H. Gamper, A. K. Lee, C. Holz, D. Johnston, M. Jalobeanu, D. Emmanouilidou, E. Cutrell, A. Wilson, and I. Tashev, "Decoding auditory and tactile attention for use in an EEG-based brain-computer interface," in *2020 8th International Winter Conference on Brain-Computer Interface (BCI)*, IEEE, Feb. 2020, pp. 1–6. DOI: 10.1109/BCI48061.2020.9061623.

[90] M. Spüler and S. Kurek, "Alpha-band lateralization during auditory selective attention for brain–computer interface control," *Brain-Computer Interfaces*, vol. 5, no. 1, pp. 23–29, Jan. 2018. DOI: 10.1080/2326263X.2017.1415496.

[91] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Super-Gaussian Mixture Source Model for ICA," Lecture Notes in Computer Science, J. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, Eds., pp. 854–861, Mar. 2006. DOI: 10.1007/11679363_106.

[92] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen, "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data,"

Computational Intelligence and Neuroscience, vol. 2011, p. 156 869, Dec. 2011. DOI: 10.1155/2011/156869.

[93] E. Combrisson and K. Jerbi, "The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods*, vol. 250, pp. 126–136, Jul. 2015. DOI: 10.1016/j.jneumeth.2015.01.010.

[94] G. Townsend, B. K. LaPallo, C. B. Boulay, D. J. Krusienski, G. E. Frye, C. K. Hauser, N. E. Schwartz, T. M. Vaughan, J. R. Wolpaw, and E. W. Sellers, "A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns.," *Clinical neurophysiology*, vol. 121, no. 7, pp. 1109–20, Jul. 2010. DOI: 10.1016/j.clinph.2010.01.030.

[95] N. Kaongoen and S. Jo, "A novel hybrid auditory BCI paradigm combining ASSR and P300," *Journal of Neuroscience Methods*, vol. 279, pp. 44–51, Mar. 2017. DOI: 10.1016/j.jneumeth.2017.01.011.

[96] W. W. An, B. Shinn-cunningham, H. Gamper, D. Emmanouilidou, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, K.-j. Chiang, and I. Tashev, "DECODING MUSIC ATTENTION FROM " EEG HEADPHONES ": A USER-FRIENDLY AUDITORY BRAIN-COMPUTER INTERFACE," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Jun. 2021. DOI: 10.1109/ICASSP39728.2021.9414492.

[97] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers," *Journal of Neural Engineering*, vol. 18, no. 3, p. 031 002, Mar. 2021. DOI: 10.1088/1741-2552/ABC902.

[98] C.-Y. Chang, S.-H. Hsu, L. Pion-Tonachini, and T.-P. Jung, "Evaluation of Artifact Subspace Reconstruction for Automatic Artifact Components Removal in Multi-

channel EEG Recordings," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 1114–1121, Jul. 2020. DOI: `10.1109/tbme.2019.2930186`.

[99] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, Jun. 2002. DOI: `10.1016/S1388-2457(02)00057-3`.

[100] K.-J. Chiang, D. Emmanouilidou, H. Gamper, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, W. W. An, and I. Tashev, "A Closed-loop Adaptive Brain-computer Interface Framework: Improving the Classifier with the Use of Error-related Potentials," in *10th International IEEE EMBS Conference on Neural Engineering*, 2021.

[101] F. Ferracuti, A. Freddi, S. Iarlori, S. Longhi, and P. Peretti, "Auditory paradigm for a P300 BCI system using spatial hearing," in *IEEE International Conference on Intelligent Robots and Systems*, 2013, pp. 871–876. DOI: `10.1109/IROS.2013.6696453`.

[102] S. Ahn, M. Ahn, H. Cho, and S. Chan Jun, "Achieving a hybrid brain-computer interface with tactile selective attention and motor imagery," *Journal of Neural Engineering*, vol. 11, no. 6, p. 066 004, Dec. 2014. DOI: `10.1088/1741-2560/11/6/066004`.

[103] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, Feb. 2011. DOI: `10.1111/j.1469-8986.2010.01061.x`.

[104] H. J. Baek, H. S. Kim, J. Heo, Y. G. Lim, and K. S. Park, "Brain-computer interfaces using capacitive measurement of visual or auditory steady-state responses.," *Journal of Neural Engineering*, vol. 10, no. 2, p. 024 001, Apr. 2013. DOI: `10.1088/1741-2560/10/2/024001`.

[105] M. Severens, J. Farquhar, J. Duysens, and P. Desain, "A multi-signature brain-computer interface: Use of transient and steady-state responses," *Journal of Neural Engineering*, vol. 10, no. 2, p. 026 005, 2013. DOI: 10.1088/1741-2560/10/2/026005.

[106] J. Li, J. Pu, H. Cui, X. Xie, S. Xu, T. Li, and Y. Hu, "An Online P300 Brain–Computer Interface Based on Tactile Selective Attention of Somatosensory Electrical Stimulation," *Journal of Medical and Biological Engineering*, vol. 39, no. 5, pp. 732–738, Oct. 2019. DOI: 10.1007/s40846-018-0459-x.

[107] M. Wronkiewicz, E. Larson, and A. K. Lee, "Leveraging anatomical information to improve transfer learning in brain-computer interfaces," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046 027, Aug. 2015. DOI: 10.1088/1741-2560/12/4/046027.

[108] J. H. Lee, H. Gamper, I. Tashev, S. Dong, S. Ma, J. Remaley, J. D. Holbery, and S. H. Yoon, "Stress monitoring using multimodal bio-sensing headset," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–7. DOI: 10.1145/3334480.3382891.

[109] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015. DOI: 10.1093/cercor/bht355.

[110] A. Aroudi, T. De Taillez, and S. Doclo, "Improving Auditory Attention Decoding Performance of Linear and Non-Linear Methods using State-Space Model," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2020, pp. 8703–8707. DOI: 10.1109/ICASSP40776.2020.9053149.

[111] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019. DOI: 10.1038/s41598-019-47795-0.

[112] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, pp. 1–14, Nov. 2016. DOI: `10.3389/fnhum.2016.00604`.

[113] S. V. David and J. L. Gallant, "Predicting neuronal responses during natural vision," *Network: Computation in Neural Systems*, vol. 16, no. 2-3, pp. 239–260, 2005. DOI: `10.1080/09548980500464030`.

[114] S. R. Synigal, E. S. Teoh, and E. C. Lalor, "Including Measures of High Gamma Power Can Improve the Decoding of Natural Speech From EEG," *Frontiers in Human Neuroscience*, vol. 0, p. 130, Apr. 2020. DOI: `10.3389/FNHUM.2020.00130`.

[115] V. Viswanathan, H. M. Bharadwaj, and B. G. Shinn-Cunningham, "Electroencephalographic Signatures of the Neural Representation of Speech during Selective Attention," *eNeuro*, vol. 6, no. 5, Oct. 2019. DOI: `10.1523/ENEURO.0057-19.2019`.

[116] M. C. Cervenka, S. Nagle, and D. Boatman-Reich, "Cortical High-Gamma Responses in Auditory Processing," *American journal of audiology*, vol. 20, no. 2, pp. 171–180, Dec. 2011. DOI: `10.1044/1059-0889(2011/10-0036)`.

[117] Y. Gao, Q. Wang, Y. Ding, C. Wang, H. Li, X. Wu, T. Qu, and L. Li, "Selective Attention Enhances Beta-Band Cortical Oscillation to Speech under "Cocktail-Party" Listening Conditions," *Frontiers in Human Neuroscience*, vol. 11, no. 34, pp. 1–10, Feb. 2017. DOI: `10.3389/FNHUM.2017.00034`.

[118] A. S. Keller, L. Payne, and R. Sekuler, "Characterizing the roles of alpha and theta oscillations in multisensory attention," *Neuropsychologia*, vol. 99, pp. 48–63, May 2017. DOI: `10.1016/J.NEUROPSYCHOLOGIA.2017.02.021`.

[119] C. Kubetschek and C. Kayser, "Delta/Theta band EEG activity shapes the rhythmic perceptual sampling of auditory scenes," *Scientific Reports*, vol. 11, no. 2370, pp. 1–15, Jan. 2021. DOI: `10.1038/s41598-021-82008-7`.