Auditory Prostheses: Hearing Loss, Attention, and Fatigue

By

BRITT WILLIAM YAZEL

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____

Lee M. Miller, Co-Chair

____

Sanjay S. Joshi, Co-Chair

____

Joy J. Geng

____

David P. Corina

____

Karen A. Moxon

Committee in Charge

2021

Table of Contents

Abstract

Understanding speech in noisy environments requires more than good hearing. Everyday acoustic scenes are complex and dynamic – far more than most laboratory or clinical settings – thus they extract a heavy cognitive toll. For instance, when following a conversation, we must exert sustained effort while continually switching attention among different talkers. However, despite the tremendous importance of these mechanisms for developing hearing interventions, we know relatively little about them. The following research has three primary aims: to design, build, and test a new class of attentional prosthetic platform as a potential aid to those with hearing loss; to characterize brain signals (EEG) that predict and track the locus of selective spatial attention during conversational turn taking; and to demonstrate the functional relationships between selective attention and listening effort that robustly signal listener fatigue.

Chapter 1

“Introduction”

Hearing and spoken language are of the utmost importance for communication, safety, and individual identity, yet there is much unknown as to how these aspects of human existence functionally work in the brain. Much in the same way hearing is a major part of the human experience, hearing loss is something that affects us all, either directly or indirectly. More than 500 million people worldwide are estimated to have hearing loss, with age related hearing loss being the largest percentage, affecting older adults across the globe (Li et al., 2018). While we have made great strides in treating many types of hearing loss, to continue advancing forward we must improve our understanding of how hearing, attentive listening, and speech processing functionally operate in the brain. Only then can we create assistive devices and technologies that truly treat the root elements of hearing loss without inflicting uncomfortable and even detrimental side effects in the process.

Auditory prosthetic devices for treating hearing loss date back to the 17th century with the invention of the "ear trumpet"; however, the first modern auditory prosthetic, the hearing aid, was not invented until 1898 by Miller Reese Hutchison (Mills, 2011). Over the subsequent 60 years, hearing aids underwent many iterations. Moving from vacuum tubes and then onto transistors, a turning point happened in the 1960s when Bell Telephone Laboratories developed the first digital hearing aid (Levitt, 2018). Since that point, hearing aid technology has exponentially moved forward increasing in both processing power and amplification range, all while decreasing dramatically in size and cost.

While modern hearing prosthetic technology can dramatically improve a listener’s ability to hear the sound frequencies that they otherwise would not be able to, they also carry with them issues relating to comfortability and can readily decrease a listener’s ability to detect the locations of sound sources. This can be particularly detrimental in complex auditory scenes where the localization of sounds is critical for a person’s ability to separate and comprehend overlapping voices, also known as the “cocktail party problem” (Zion Golumbic et al., 2013; Kidd, 2017).

Humans localize sound on the azimuth plane through two primary means, interaural level differences (ILDs) and interaural time delays (ITDs). Both techniques leverage the fact that, generally, humans are equipped with two independent ears, with each ear receiving a similar, albeit unique, representation of the external world. The differences between what each ear receives and processes are the source of the information used to determine sound locations (Tollin and Yin, 2009). ILDs form predominately due to the mass of the head impeding the direct path of the sound waves to the ear canal. A sound located more directly in the path of a given ear will appear louder to the brain, with the degrees of loudness differences “mapped” within the brain to an array of azimuth positions ranging from 0° center to 90° on either side of the head. ITDs form due to the spatial geometry of the head, and since the ears are separated in space, the distance from a source sound to either ear will be, by necessity, different for all angles other than 0° directly in front of or directly behind the head. Likewise, since the speed of sound is a constant in our normal Earth atmosphere, the time at which sound reaches each ear will be slightly different depending on the travel distance. These two means are critical for sound source localization, and, unfortunately, traditional hearing aids can have detrimental effects on them both (Denk et al., 2019).

Modern hearing aids function by recording the outside world with a small microphone (per ear), processing this recording to increase the sound level at the auditory frequencies for which the listener has hearing loss and playing this processed audio back to the listener with as little latency as possible while preserving the integrity of the binaural audio. While this may on the surface sound like a straightforward task, accomplishing these goals within the battery, power, and form factor limitations is incredibly complicated with very little margin for error. Likewise, failing these goals works against the acceptance of hearing prostheses in the hearing impaired community, with many finding them unsightly or embarrassing to wear, uncomfortable to listen to for long lengths of time, or outright detrimental to their day to day lives (Denk et al., 2019).

Hearing aids first begin to distort the spatial cues of the auditory scene at the microphones themselves, where non-optimal positioning can have an effect on the integrity of the sound as it would have been heard though the listener’s own ears (Denk et al., 2018). Further, the hearing aid’s innate modifications to the sound level and the processing delays occurring at many stages of the processing pipeline greatly distort the binaural spatial cues necessary for azimuth sound localization, leading many wearers to find themselves unable to localize sounds nearly as precisely as their healthy hearing counterparts (Denk et al., 2019). In complex scenes, this can manifest in a “spatial compression", where voices from many different angles now appear to overlap one another, which, beyond the binaural cues being distorted, can also mask many of the monaural spectral cues necessary to attend onto specific voices and their contents (Marrone et al., 2008).

A way to deal with the unfortunate deficits regarding the technology behind real-time audio processing and amplification found in hearing prostheses is the use of a technology known as "auditory beamforming," currently pioneered by Gerald Kidd, Jr. with his work on visually guided hearing aids (Kidd et al., 2013). Auditory beamforming is a technique whereby multiple microphones are precisely spatially arranged relative to one another, known as a microphone array, and through knowing these relative locations a device can amplify select angles of sound while suppressing all others. This creates a "spotlight" so-to-speak on the auditory scene that can be leveraged by hearing prosthetic technologies to better solve the "cocktail party problem". With this implementation, much of the work in separating the spectral and spatial properties of competing talkers is placed on the prosthesis itself, making the task of auditory attention much easier on part of the listener. In a way, this auditory beamforming implementation is more of an "attentional prosthesis" than it is simply an auditory prosthesis.

Our work seeks to advance that of Kidd et al. and the study of hearing loss and the "cocktail party problem" in three ways. First, we designed and implemented a prototype platform, written for Android and run-on mobile technology, that can do the processing necessary for complex auditory beamforming. We named this platform "Cochlearity"

(Anderson et al., 2018). Second, we aimed to better understand the attentional dynamics of the brain in multi-talker situations, and how the locus of attention can be decoded in the brain. We did this for both healthy hearing and hearing-impaired individuals by recording the neural signals using EEG, and by training predictive models based on where a listener is attending and where they are not attending. This research gives us a greater understanding of how these bio-signals can be leveraged for potentially even smarter attentional prostheses. And lastly, to explore the effects of prolonged, fatiguing auditory attentive situations on healthy hearing and hearing-impaired listeners, we designed and tested a novel paradigm requiring listeners to continuously switch their selective attention over an hour of stimulus presentation. The purpose is to fatigue a participant's selective attentional faculties, which we can then track to better understand how this prolonged attention is affected over time. Likewise, while not a part of these studies, the last step in this research would have been to use the prolonged, fatiguing stimulus on participants while wearing the 'Cochlearity' attentional prosthetic platform to see how and if it is able to modulate the fatiguing effects on participants of the time span by offloading that task onto the device itself.

Chapter 2

“Towards mobile gaze-directed beamforming: a novel neuro-technology for hearing loss”

This work was performed in partnership with Markham Anderson and was originally published in the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

I. Abstract

Contemporary hearing aids are markedly limited in their most important role: improving speech perception in dynamic "cocktail party" environments with multiple, competing talkers. Here we describe an open-source, mobile assistive hearing platform entitled "Cochlearity" which uses eye gaze to guide an acoustic beamformer, so a listener will hear best wherever they look. Cochlearity runs on Android and its eight-channel microphone array can be worn comfortably on the head, e.g., mounted on eyeglasses. In this preliminary report, we examine the efficacy of both a static (delay-and-sum) and an adaptive (MVDR) beamformer in the task of separating an "attended" voice from an "unattended" voice in a two-talker scenario. We show that the different beamformers can complement each other to improve target speech SNR (signal to noise ratio), across the range of speech power, with tolerably low latency.

II. Introduction

Everyday auditory environments are cluttered, noisy, and distracting. This presents a complex perceptual and computational challenge known as the "cocktail party": how to extract relevant acoustic information while filtering out the background noise. Individuals with healthy hearing tend to perform well in typical multi-talker environments, as our brains are adept at discriminating sound source locations and identities . However, while hearing aids can significantly boost the detection and comprehension of sounds for those with hearing loss, particularly in quiet backgrounds, and can even improve "downstream" effects on auditory cognitive function (Acar et al., 2011), they do not adequately address the issue of understanding speech in noise.

Modern digital hearing aids often use multiple features to improve perception in loud, crowded environments, such as on-board directional microphone systems and adaptive speech enhancement or noise reduction algorithms. But even with these sophisticated features, aids cannot effectively “listen” to what the user wants; they often fail in real situations, amplifying noise as much as the desired information. This shortcoming leads to listening confusion, poor real-world speech comprehension, and low rates of use for assistive devices (Blazer et al., 2016).

We sought to address this issue by creating a system that can be automatically guided by a user’s intentions — in this case their eye gaze direction — and thereby serve as the basis for an intelligent hearing aid (Favre-Felix et al., 2017).

Our approach builds on the seminal work of Gerald Kidd et al., Hart, and Marzetta (Marzetta, 2008; Hart et al., 2009; Kidd et al., 2013). Like Kidd et al., we use gaze-directed beamforming, a method of highly directional sound amplification and attenuation, to isolate sounds within a “beam” of auditory space. And like Kidd et al., the direction of the beam will be steered through real-time gaze tracking, as an analog to listening intention. The primary differing factors between Kidd’s system and our “Cochlearity” platform are three-fold. First, Cochlearity is implemented on widely available mobile device hardware using Android as opposed to on workstation-class desktop hardware. Second, this first version of Cochlearity will be entirely open-source and available under a standard, permissive license to encourage broader adoption and further improvements. And finally, we make use of both passive and adaptive beamforming algorithms, as opposed to just passive. In this report, we evaluate whether combining passive and active beamforming algorithms in parallel might improve performance substantially with little computational cost.

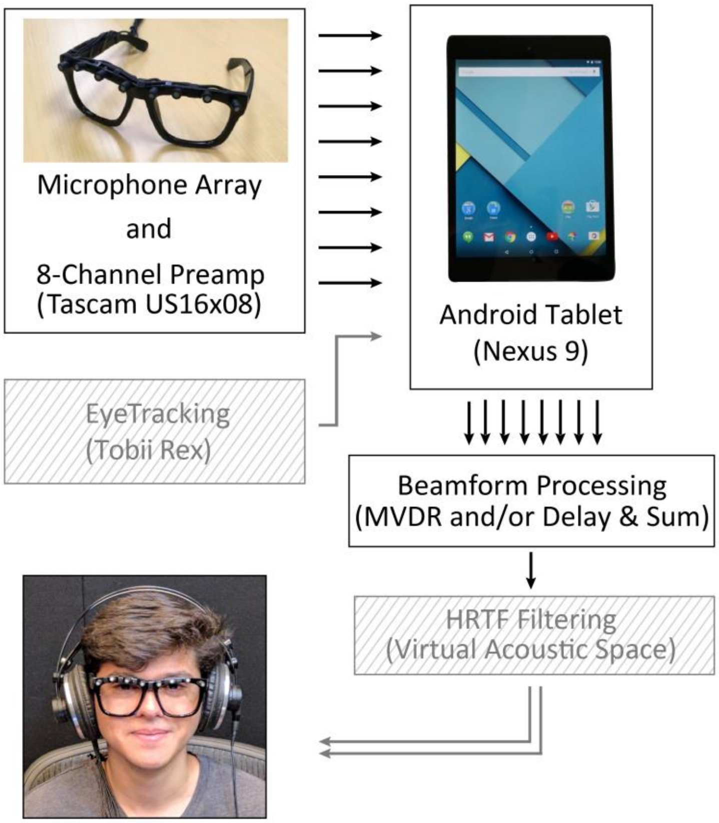


Figure 2.1. “Cochlearity”: a mobile, gaze-directed beamforming platform for assistive listening.

III. Design

Unlike a more traditional PC-based implementation of acoustic beamforming, we power Cochlearity (the software application) with a Nexus 9 tablet running Android (6.x). We use an array of eight microphones, but Android OS has lacked support for input of more than two audio channels, and compatible tablets tend to furnish only a single (occasionally dual) microphone. Therefore, we implemented our own software for I/O and augmented our hardware with a Tascam US-16x08 audio recorder, which serves as our multi-channel Analog to Digital Converter (ADC), connected to the tablet via USB. Similarly, for gaze input we use a Tobii Rex eye tracker, connected to the tablet via USB and running the Tobii Gaze Android Software and driver (Figure 2.1).

IV. Beamforming

The premise of acoustic beamforming is to combine signals from an array of multiple (in our case eight) precisely spaced microphones to emphasize sound energy from a certain direction and suppress it from all others. A passive beamformer uses only the array geometry and speed of sound to combine the signals mathematically; as a result, it will tend to be simple with low latency. An adaptive beamformer additionally uses statistical learning about noise sources in the environment, which can improve performance but may be practically limited in real-time applications by the additional computational cost. In both cases, the beamformer outputs a single, spatially sensitive audio signal that contains proportionally more information from one region of space than from any others (Kidd et al., 2015). Cochlearity currently implements two distinct beamforming algorithms, ‘delay-and-sum’ and ‘Minimum Variance Distortionless Response’ (MVDR).

Cochlearity first buffers the 8-channel USB audio inputs into 1024 sample (21 millisecond) frames, which are then passed on to the filtering or beamforming operations.

Shorter frames would have enabled lower latency but would diminish granularity for the discrete Fourier transform used by our adaptive beamformer and would eventually have presented an I/O bottleneck. Thus, the samples being processed are always necessarily (at-least) -21 milliseconds relative to real-time, though with processing and I/O operations this latency is considerably longer, detailed later.

A. Delay-and-sum Beamforming (D&S)

Delay-and-sum beamforming is a passive algorithm that leverages the propagation time of sound, which manifests as signal delays from one microphone to the next. This delay varies with the angle of the target audio relative to the microphone array. By offsetting the signal in each channel by the delay for a given “steering” angle and then summing the resulting signals across channels, it delivers a signal that contains a constructively reinforced component coming from the desired angle, with all other angles destructively attenuated (Vu et al., 2010).

B. Minimum Variance Distortionless Response (MVDR) Beamforming

The MVDR beamformer is an adaptive algorithm, as opposed to the delay-and-sum beamformer. In addition to compensating for the time delays due to steering angle, it uses an adaptive filter to null interference from other angles (Capon, 1969). Time and frequency are divided into bins of fixed size, and for each time-frequency bin, an NxN noise correlation matrix is computed (N being the number of microphones). From the noise correlation matrices, a linear transformation is computed to minimize the noise (anything other than the desired signal) on current and future inputs, with the constraint being that the signal originating on the target angle be preserved (Habets et al., 2010).

To reduce computational expense, our implementation of the MVDR beamformer used only the two end microphones, bringing the noise correlation matrices down to a 2x2 dimension instead of 8x8. The delay-and-sum beamformer, however, used all eight microphones to improve the resolution of its constructive/destructive interference operation, with scarcely higher cost than a 2-microphone delay-and-sum beamformer. When used in isolation, each beamforming algorithm is given the full audio bandwidth as input.

V. Methods

All testing was conducted in a sound treated room with dimensions of 3.5x2.5 meters. Eight Audio-Technica AT8537 phantom powered microphones with an 80Hz high-pass pre-amp filter were mounted linearly upon a set of eyeglasses with 1.86cm spacing and a total length of 13cm. The glasses were set upon on an anatomically accurate dummy head positioned facing forward (defined as 0^{o}) on a table, with two Tannoy Precision 6 speakers positioned 140cm away at -50^{o} and $+50^{o}$ pointing directly at the array. Speaker outputs were balanced using a digital sound level meter to within 1 dB SPL using Gaussian white noise.

During each performance test, the beamformer steering direction was manually set by the researcher. Two audio tracks were played simultaneously, each through one of the speakers at a comfortably loud listening level. The speaker positioned at -50^{o} played “20,000 Leagues Under the Sea” while the speaker at +50^{o} played “Journey to the Center of the Earth”, both by author Jules Verne. Both stories were read by the same male reader at a constant pacing and were equalized for power, but the voice at -50^{o} was pitch-shifted up by 7%, and the voice at +50^{o} was pitch shifted-down by 7% using Adobe Audition.

For all recordings, beamformer output was captured directly from the tablet through the headphone jack, which was then passed into a Sound Devices X-3 headphone amplifier, amplifying the audio before it was sent to the USB audio capture card, an Edirol (Roland) UA-25, and then onto the PC using Audacity.

The recordings were made in sets of two for each beamforming paradigm: beam steered to -50° azimuth (left), and +50° azimuth (right). Lastly, two reference recordings were made with all beamformer processing turned off, and the speech was played separately out of the -50° speaker or the +50°. These references were necessary as a point of comparison for the recordings, as they captured the same signal filtering imposed by the room, speaker placement, and microphones, as well as the generic I/O overhead in Android. Thus, any differences between them and the beamformed recordings should be due entirely to the processing imposed by Cochlearity’s beamforming.

To compare our two different beamformers, we use spectral coherence as a measure of the relatedness between our beamformed audio recordings and the reference recordings. Each recording was compared against the left or right talker reference for spectral coherence. Specifically, in each beamforming paradigm:

"Attended Voice" is the congruent coherence between the left speaker reference recording and the beamformed recording when steered to the left, and coherence for the right speaker reference recording with the beamformed recording when steered to the right, averaged together

"Unattended Voice" is the incongruent coherence between the left speaker reference recording and the right-steered beamformed recording, and coherence for the right speaker reference recording to the left-steered, beamformed recording, averaged together.

Likewise, whereas coherence shows beamformer performance as a function of frequency, overall performance can be summarized as SNR (dB) between attended and unattended voices. We calculated SNR as $10*\log_{10}$ of the average ratio in coherence between the two conditions, weighted by the speech power across frequencies.

A. Coherence Difference Index (CDI) and Latency

To quantify the effectiveness of a beamformer, we computed a "Coherence Difference Index (CDI)" for each paradigm (Table 2.1). We calculated this by taking the difference between the "attended" and "unattended" coherence for a given beamformer, and then performed a weighted average between 0 and 5000Hz (which captures most of the speech power), weighted by the spectral density estimate of both voices combined (using MATLAB's pwelch function). We then multiplied this number by a factor of 100. This provided a rough global approximation of how well each paradigm emphasized the voice from the steering direction *and* suppressed the interfering voice.

A “CDI Efficiency” was determined as the CDI per millisecond of latency (CDI ÷ Latency) (Table 2.1).

Latency was computed using a series of clicks played through the speakers and recorded both in a reference microphone (not connected to Cochlearity) and through Cochlearity for each beamforming paradigm (Table 2.1).

B. Spatial Analysis

To characterize the spatial effectiveness of our two beamformers, we performed an analysis in which we placed a speaker at 0 ^{o} and kept the beamformer steered to this angle. The story played by this speaker, “20,000 Leagues Under the Sea”, is referred to as the attended voice. Next, we moved a second speaker playing a masking voice, “Journey to the Center of the Earth”, in 10^{o} increments from -50^{o} to +50^{o}, recording 1 minute of speech at each location. We then assessed the performance of the beamformer at each masking angle relative to the 0^{o} fixation using the CDI as our metric, illustrating the effectiveness of the beamformer in extracting the attended voice from the masking voice.

Lastly, while real-time gaze tracking and virtual 3-d audio rendering using head-related transfer functions or HRTFs are integral and fully realized parts of Cochlearity, the data reported in this study are only meant to characterize the effectiveness of Cochlearity’s beamforming implementation, and as such there is no gaze tracking component to the tests. A future study will explore the effects of how Cochlearity performs with human subjects.

VI. Results

The delay-and-sum beamformer did not perform well at frequencies <300Hz, with little difference between the attended voice and the unattended voice coherence. However, at frequencies above 300Hz, and most notably >1000Hz the beamformer was able to effectively separate the attended from the unattended voices. The low-frequency performance reflects, in part, the relatively small array size and confirms the literature that delay-and-sum beamforming works best at moderate to higher frequencies (Kidd et al., 2013) (Figure 2.2).

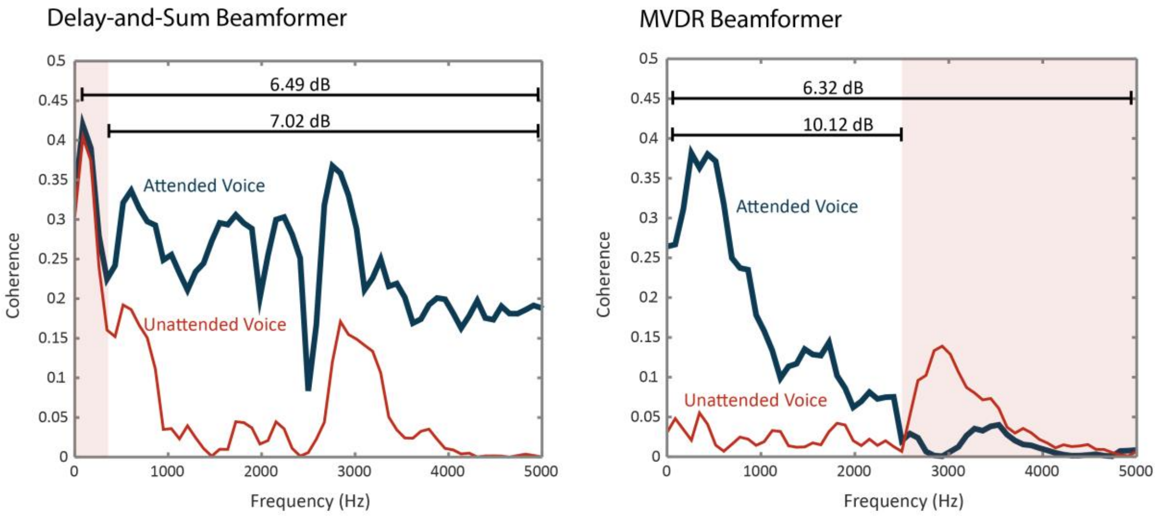


Figure 2.2. Delay-and-Sum and MVDR beamformer performance (red shaded frequencies indicate poor performance). Decibel (dB) labels indicate average SNR between attended and unattended voices.

The MVDR beamformer performed well, most notably at frequencies <2500Hz. Conversely, considering the performance of delay-and-sum, the MVDR beamformer performed the worst at middle to higher frequencies, leading to little or no improvement in signal coherence between the attended and unattended voices (Figure 2.2). Thus, the two beamformers complement one another in performing across the crucial frequency range where speech has high power.

	D&S	**MVDR**
Latency (ms)	127.60ms	145.24ms
Coherence Difference Index (CDI)	13.71	13.64
CDI Efficiency (CDI/ms)	**0.11**	**0.09**

Table 2.1. Latency, CDI, and CDI Efficiency at 100o separation of attended voice and masking voice

Spatially, both beamformers performed well, showing a clear trend in increasing CDI values the farther away the masking voice was from 0°. This is expected given that maximal spatial overlap between attended and unattended voices occurs at 0o. Regarding the magnitude of the CDI, at 0° both beamformers were equivalent, each with a CDI of ~0, but within +/-10-20° the MVDR beamformer outpaced the delay-and-sum, ending at -50° and +50° with more than double the CDI of the delay-and-sum (Figure 2.3). This indicates that the MVDR has better performance at much smaller masking angles than that of the delay-and-sum. Nevertheless, Table 2.1 shows that at 100° of separation between the attended and masking voice there is near equal CDI.

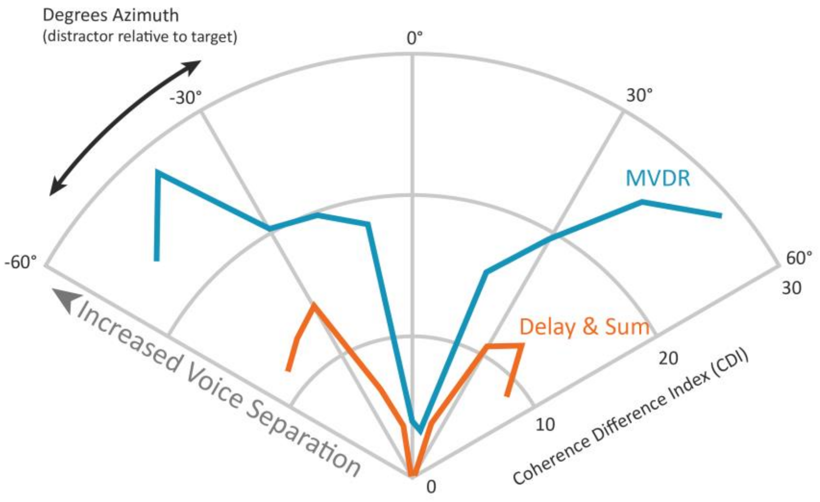


Figure 2.3. Spatial release from masking: how well each beamformer can reject interference.

The latency of our entire system (the time from sound production to playback) when running the two beamformers differed by approximately 17.6 milliseconds, with the delay-and-sum taking a total of 127.60 and the MVDR taking a total of 145.24 milliseconds. We should note that part of this latency is due to the necessary framing or buffering of the real-time audio (presently 21ms frame size); however, part of it is due to basic device I/O (input and output) operations on Android. This is encouraging given that great strides have been made in the time since Android 6.x was released to decrease audio pass-through latency. Even in this preliminary form, the overall output latency of our system still falls into a range that would allow sound to be naturally combined with visual cues such as mouth movements as audiovisual integration supports a synchrony window up to ~200ms (van Wassenhove et al., 2007). Since delay-and-sum had somewhat lower latency, as expected, but similar CDI performance as compared with MVDR (13.71 v 13.64 respectively), delay-and-sum beamformer had a CDI Efficiency marginally better than the MVDR. Therefore, by restricting the MVDR to two channels, performance is preserved – in a complementary frequency range – and latency is reduced, to be comparable to the simpler delay-and-sum algorithm.

VII. Conclusion

The results from this project show the potential for wearable, gaze-directed beamforming to improve speech perception in realistic environments. To our knowledge this is the first time multiple beamformers have been implemented successfully on a mobile, assistive listening platform, to capture the range of important speech frequencies with reasonable latency and computational cost.

Given that the MVDR beamformer worked most effectively on lower frequencies and the Delay-and-sum worked best on high frequencies, our current work aims to filter the input to restrict each algorithm to its best range and combine them to yield improved results. Future work will also demonstrate how Cochlearity performs with its real-time eye tracking and virtual 3-D rendered audio, with both hearing impaired and healthy listeners in a laboratory setting as well as in real-life, social scenarios.

Chapter 3

“Attentional modulation of neural speech-envelope tracking in hearing impaired listeners”

This work was completed at the Starkey Hearing Research Center in Berkley, California, and was funded and facilitated by Starkey Laboratories, Inc.

I. Introduction

As far as we know, humans are unique among mammals in that we have the capacity for complex, language-based communication. The neurophysiological processes by which we internally generate speech output and process incoming speech has been a huge topic of intrigue for centuries, and, most recently, one question is how we cope and thrive in environments with many competing sound sources. This phenomenon is known colloquially as the "cocktail party problem" (Cherry, 1953). When presented with concurrent sound stimuli, humans possess the capacity to separate sound sources into discrete units, each of which can be processed independently (Ding and Simon, 2012). Practically, this means that humans can focus entirely on a single sound object, while virtually attenuating other sources of sound. Thus, a healthy hearing individual can block out distractor speech, while maximizing intelligibility on a desired speech stream (Cherry, 1953)

The same cannot necessarily be said for hearing impaired individuals. Much of the information as to the causes and neural manifestations of hearing impairment have yet to be discovered due to the extreme difficulty in measuring low and middle level auditory structures. Nevertheless, behavioral evidence has shown a decrease in the capacity for hearing impaired individuals to parse the "cocktail party" problem, generally expressed alongside a global loss in speech and sound intelligibility (Shinn-Cunningham and Best, 2008a), which leads to an interesting question: Is the higher-level "cocktail party" deficit in hearing due to low level sound transduction? - i.e. you cannot accurately parse sound streams that you are unable to hear reliably to begin with – Or, is it due to higher level cognitive deficits that bar functionally solving the “cocktail party” problem in ways not explained through a basic loss in audibility?

Exploring how humans process speech, multiple studies have shown that there is cortical entrainment to the envelope of incoming speech. Further, they show that it is possible to use a reverse direction mapping approach (i.e., recreating the incoming stimulus based on recorded neural data) to recover a proportion of the speech envelope (Ding and Simon, 2012; Mesgarani and Chang, 2012; O’Sullivan et al., 2015a). Further, in multi-talker environments, MEG, EEG, and ECoG studies have explored how cortical entrainment of the speech envelope is affected when there are multiple competing speech streams. They have found that while you can find components of multiple streams in the neural data, the attended speech stream is preferentially encoded when compared to the unattended speech stream, measured based upon the degree to which one can recreate each of the original stimulus envelopes (Ding and Simon, 2012; Mesgarani and Chang, 2012; O’Sullivan et al., 2015a). Thus, the degree of stimulus envelope reconstruction is used as an indirect measure for the amount of neural processing dedicated to a specific stimulus.

In 2014, O’Sullivan and Lalor were able to show that with the envelope re-creation technique and with only a minute of EEG data it is possible to predict retroactively and reliably which of two competing talkers was the target of attention. With our study we explored how well we could recreate O’Sullivan’s work about healthy hearing individuals, both in prediction accuracy and the magnitude of correlation values (O'Sullivan et al., 2015), using a similar paradigm but with distinctly different stimuli. Additionally, the novel aspect of this project was to see how translatable this technique is with hearing impaired individuals, and, if successful, quantifying which ways the hearing-impaired sample differs from our healthy hearing group.

II. Methods

Six healthy hearing subjects (2 males; 4 females; mean age = 49.5) and 4 hearing impaired subjects (1 male; 3 females; mean age = 64.25) were each asked to perform a dichotic listening task in which they were told to attend either to a speaker in their left ear or to a speaker in their right ear. Hearing impaired subjects each were each characterized as having bilateral mild sloping to profound hearing loss. The subjects were told to attend either fully left for the duration of the study or fully right (in equal proportions in the healthy and hearing-impaired groups) and were not required to switch their attention at all. Each subject was presented with two male voices simultaneously (one per ear), each being a local radio personality reading a series of published short stories (3 stores in total). Further, each participant's trials only differed from the other's trials in the order the stories were presented (pseudo-randomly shuffled) and to which speaker (left or right) they were asked to attend. Each subject heard the exact same set of spoken stories, and the spoken stories perceived volumes were all normalized between ears such that each speech stream was equal in volume to the other. Likewise, the stories had gain evenly applied to set the audibility to a “loud but not uncomfortable” level, qualitatively reported by the subjects on a ten-point scale prior to the experiment.

The spoken recordings had their volumes balanced against one another previously through a series of qualitative assessments from many listeners, and any gap in sound longer than 0.5 seconds were reduced to 0.5 seconds to reduce any chance of the unattended stream capturing attention inadvertently. Each participant listened to a total of 25 minutes of story, which consisted of 3 separate stories, with each being broken up into ~1 min segments (25 total segments). Each segment was chosen in a logical stopping point in the context of the story, therefore no story segment ended in the middle of a word or phrase. Thus, each trial (n=25) had two stimuli speech tracks, one that will be referred to as the “attended stimuli”, and the other being the “unattended stimuli”. Likewise, after each ~1 min segment the subjects were asked to answer two semantic questions each being a two-choice question inquiring as to a fact in the story, i.e. “The woman’s dress was: A) red or B) purple?” Accuracy in answering the questions and the time it took to respond to the questions were both recorded and quantified.

The voices were played to our participants in a sound treated and semi-electrically shielded chamber through Etymotic ER-3a scientific grade earphones. This choice of sound delivery was desirable due to its ability to reject electrical noise that might corrupt the EEG recording. It accomplishes this due to keeping the electrical and magnetic components comparatively farther away from the head than conventional magnetic headphone or earbuds, as it uses pneumatic tubes to deliver the sound pressure changes rather than having the magnetic drivers immediately next to or in the ear canal. In this regard, we differed from the O'Sullivan study, as their data was collected using Sennheiser HD650 over-the-ear headphones.

EEG was recorded from each subject across 64 channels at a sampling frequency of 10khz. CZ was used as the reference electrode and was then interpolated later during the analysis phase of the study. All recording was done on Brain Products hardware.

Data analysis (detailed below) was performed with an end goal of recreating an "estimated" speech envelope that is correlated to the original envelope to some degree (generally with Pearson correlation values of 0.03-0.09). Using linear regression (detailed below), one can create two models for each trial, an attended model and unattended model, each of which can attempt to predict the attended envelope and the unattended envelope. The efficacy at which the models can create their corresponding envelopes, as opposed to the opposite envelope, is what is used as the metric for prediction of where the subject's attention was for that trial (further detailed below).

III. Data Analysis

A. Preprocessing

The EEG data was band-pass filtered and down sampled to 0.5-40hz and 128hz respectively. The EEG data was analyzed using MNE, any qualitatively chosen "bad" channels (due to excessive noise or drift) were removed and interpolated; likewise, the CZ electrode needing interpolation, as it was used as the reference electrode for all other channels and thus does not have its own native channel in the data. Further, using independent component analysis (ICA) eye blink components were isolated and removed. Lastly, the EEG data was broken into ~1min segments corresponding precisely to their ~1min story segments, which was determined to <1ms worth of accuracy by inscribing the EEG data with start and stop triggers that were time locked to the stimulus.

For generating the models, a linear regression toolbox written by Lalor and colleagues "mTRF" version 1.3 was used, and was the same toolbox used in the O'Sullivan study (https://sourceforge.net/projects/aespa/). The speech stimuli were Hilbert transformed and low pass filtered below 15hz after applying an antialiasing filter to compute the amplitude envelope; likewise, the EEG was low pass filtered below 15hz. This is slightly different that O'Sullivan did — they band pass filtered his EEG from 2-8hz and low passed his envelope at 8hz, whereas we low passed both at 15hz. We did this because we found that our prediction accuracy was significantly lower when we did not include the 0.5-2hz range in our regression, and Di Liberto and Lalor found that there were significant contributions to the decoder accuracy for frequencies up to 15hz (Di Liberto et al., 2015).

B. Linear Regression

There are two types of models that can be trained in this type of paradigm, a "forward model", i.e., recreating an estimate of neural data from a given stimulus, or a "backward model", i.e., recreating an estimate stimulus from a given neural data. For this study, we used backward models to obtain the information necessary for our prediction.

We trained our system to recognize the relationship between neural data (recorded during dichotic listening) and the attended speech envelope component that evoked said response; we then repeated these steps for that same neural data, but this time understand the relationship with the unattended speech envelope component. We trained the linear regressor on both attended and unattended speech envelopes because, in theory, both representations ‘should’ be present in the neural data, though with differing efficacy. Once the system was trained, we then were able reconstruct a “best estimate” of each speech envelope (both the attended and unattended) with any new dichotic neural data presented (detailed extensively below). To assess the accuracy of these predictions, a Pearson correlation was used to compare the estimated envelop as calculated by our model to the real speech envelope (reported as a Pearson correlation (r) value). The higher the Pearson correlation (ranging between -1 (anti-correlated) to 1 (perfectly correlated)) the “better” our system is at reconstructing the original stimulus envelope.

To maximize the amount of data over which we could regress, we used a “leave-one-out” paradigm on our 25 data sets for specific ranges of “lags” (see Lag Analysis). Using this paradigm, given that we had recorded 25 trials for each of our subjects, one at a time we stepped through each of the 25 trials, choosing to leave that n^{th} trial as our “test” data set as opposed to using it to train our model. Further, with each step we used each of the other 24 data sets to train two decoders, an “attended decoder” and an “unattended decoder”, the 24 attended decoders and 24 unattended decoders were then averaged to give two robust models that could well predict the left out “test” data set

(O'Sullivan et al., 2015). Thus, after 25 "leave-one-out" steps we had created 2x25 models, each attended and unattended pair differing only regarding which trial was excluded. This allowed us to always test our models on neural data that was not used in the regression and gave us models that were trained on both the attended speech and the unattended speech.

The mTRF toolbox was used to perform our linear regressions, over a specified window of time, to generate two backward models for each of our trials (an attended model and an unattended model), i.e., each subject has 2x25 models. For our backward models, each model consisted of a two-dimensional matrix, with rows corresponding to individual EEG channels, and columns corresponding to the number of "lags" (see Lag Analysis) used in the regression. Generally, the size of these matrices was 64x33 (for a 0-250ms regression window) or 64x77 (for a -100-500ms regression window), aside from the single lag analysis that was performed where the dimensions were 64x1 (explained later).

C. Prediction

In general, the concept behind this prediction technique is predicated on the hypothesis that there should be more neural activity devoted to processing the attended speech stream than there is in processing the unattended speech stream. And by extension, as it has been shown in previous literature that neural activity measured through EEG has entrainment to the envelope of speech stimulus (Aiken and Picton, 2008; Di Liberto et al., 2015, 2015), if this logic is valid, there should be a more significant EEG component to the envelope of our attended speech than to that of the unattended speech, though components of both should be present in some quantity.

The stimulus reconstruction technique is an indirect measure in how to determine the amount of stimulus-related neural activity. Further, determining the accuracy of stimulus reconstructions with Pearson (r) correlation values gives a convenient summary of the amount of stimulus-related activity in a complex, multi-dimensional neural data set. The mTRF toolbox was used again to apply the previously trained models to predict an estimate of both the attended speech envelope and the unattended speech envelope, keeping in mind to never use a model to predict the same data it was trained on. Each predicted envelope was then Pearson correlated with the 'real' counterpart envelope (i.e., estimated attended envelope to the attended envelope, and the estimated unattended envelope to the unattended envelope). Further, each estimated envelope was Pearson correlated against its mismatch (i.e., estimated attended envelope to the unattended envelope, and the estimated unattended envelope to the attended envelope) to see how much correlation could be attributed to the congruent model, an analysis that was necessary for our prediction.

If the above reasoning that there is proportionally greater neural activity (and therefore speech envelope entrainment) in processing an attended speech stream is true, it leads to the simple logic in which the attended model should be able to predict a greater proportion of the attended envelope over the unattended envelope; a successful prediction being if $r_{attended} > r_{unattended}$ - while the unattended model should be able to predict a greater proportion of the unattended envelope over the attended envelope; a successful prediction being if $r_{unattended} > r_{attended}$. Comparing the Pearson correlations for the two stimuli gives a convenient way to compare the amount of stimulus-evoked neural activity due to each stimulus, and subsequently the focus of attention. Further, by doing this analysis twice with both attended models and unattended models we are given two measures towards the focus of attention **-** the attended model predicts where the subject **was** attending, and the unattended model predicts where the subject **was not** attending.

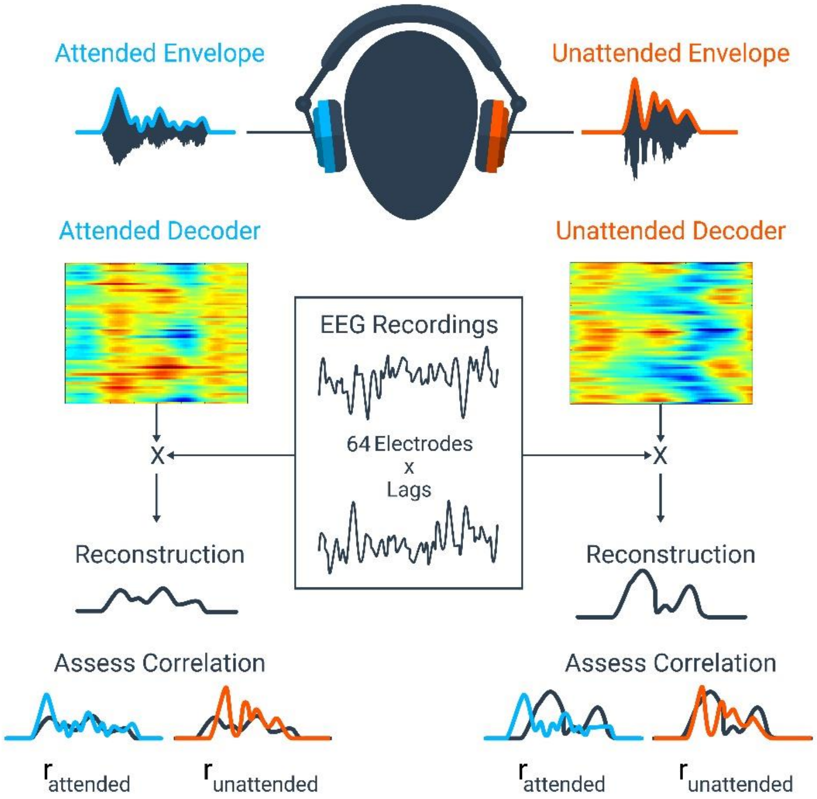


Figure 3.1. Credits to J. O’Sullivan et al., 2014 for the methodology used in this study.

D. Lag Analysis

Each regression was performed over a series of “lags”, a lag being a discrete unit of time from stimulus onset by which the EEG data is shifted to account for the transduction time of the signal through hierarchical processing centers in the brain. It takes time for cortex to receive and process input from the ear, thus at time T=0 it is causally improbable that any cortex activity was due in part to the stimulus. Now, if we were recording electrophysiology from a single level of cortex, you might find the correlation power as a function of lag distributed along a Gaussian curve, i.e., at some level of cortex there is an optimal “lag” that corresponds to the exact travel time of information from the ear to that level of cortex. However, as EEG records across many levels of cortex simultaneously, there will be a range of lag values that would need to feed into the regressor for maximal stimulus reconstruction power, as each level of cortex will have its own optimal lag. Thus, O’Sullivan determined his lag window to be from 0ms-250ms, which we initially used as well, but later found cause to extend the window out further in the case of hearing-impaired individuals.

Further, as a follow-up, we did a “single lag” analysis that regressed over a single “lag” worth of data (at 128hz sample rate, 1 lag = 7.8125ms) to obtain a time course measurement of which lag values contributed the most towards the magnitude of our Pearson (r) values. For this we generated 77 models for each trial of data from -100ms-500ms (600ms/7.8125ms=77 individual lags). As a note, we extended our lag window further back than t=0ms (to -100ms) to garner an assessment of how much correlation is due in part to the directly causal stimulus, and how much is due to generalized traits in the stimuli that could not be directly causally related. Thus, our total number of models generated from this analysis was 77 lag models * 25 trials * 6 subjects(healthy) or 4 subjects (hearing impaired) * 2 (both attended and unattended models) = 23100 individual models(healthy) and 15400 models (hearing impaired). Individual time lag correlations were averaged across trials and subjects.

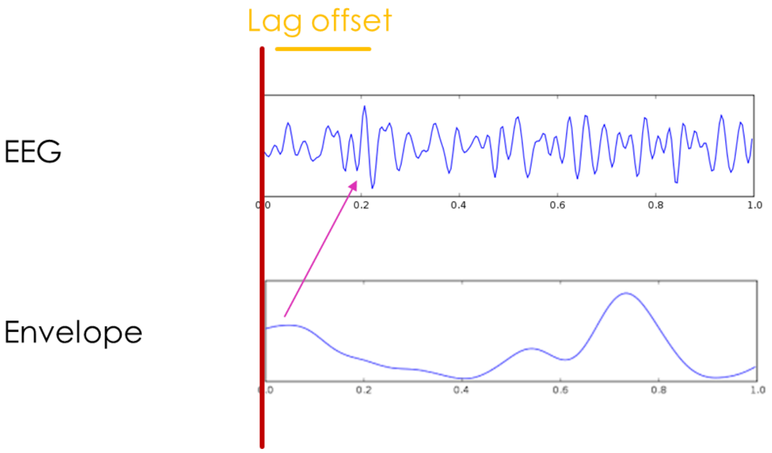


Figure 3.2. Example illustrating the lag offset between the auditory envelope and the resulting EEG signal.

E. Lag Window

Regressions were processed twice, once with an overall lag window from 0-250ms, which is what is commonly used in the literature, and then later with a lag window from -100-500ms. The reason for this larger lag window (which exponentially increases the processing time necessary to calculate backward models) was based in the results from the above “single lag” analysis.

IV. Results

A. Behavior

Both healthy and hearing-impaired subjects did equally well in answering the questions from the task and had very similar times it took to answer the questions. The healthy hearing subjects answered on average 86% percent of the questions correctly, with hearing impaired subjects answering 83% correctly. Likewise, the healthy hearing subjects on average took 5.5s to answer a question, whereas hearing impaired took 6.5s (Figure 3.3). In general, the hearing impaired did slightly worse in performance, but it is within the margin of error, thus we are treating both groups as performing equally. Further, we discarded any questions in which the subject took longer or shorter than two standard deviations from their mean time to answer a question, as they either never chose and the questions timed out, or they accidentally pressed multiple key presses simultaneously and skipped over a question without first reading it.

B. Prediction Accuracy

Prediction accuracies were looked at both 0-250ms lag windows and -100-500ms lag windows, the first range being what is commonly found in literature and the latter being a range we felt it necessary to explore based on the results in the single lag analysis.

When regressing from 0-250ms, for the healthy subjects, the attended decoder had an average prediction accuracy of 77%, while the unattended decoder had an average prediction accuracy of 60% (Figure 3.4). For the hearing-impaired subjects, the attended decoder had a mean accuracy of 65% and the unattended decoder had a mean accuracy of 64% (Figure 3.5).

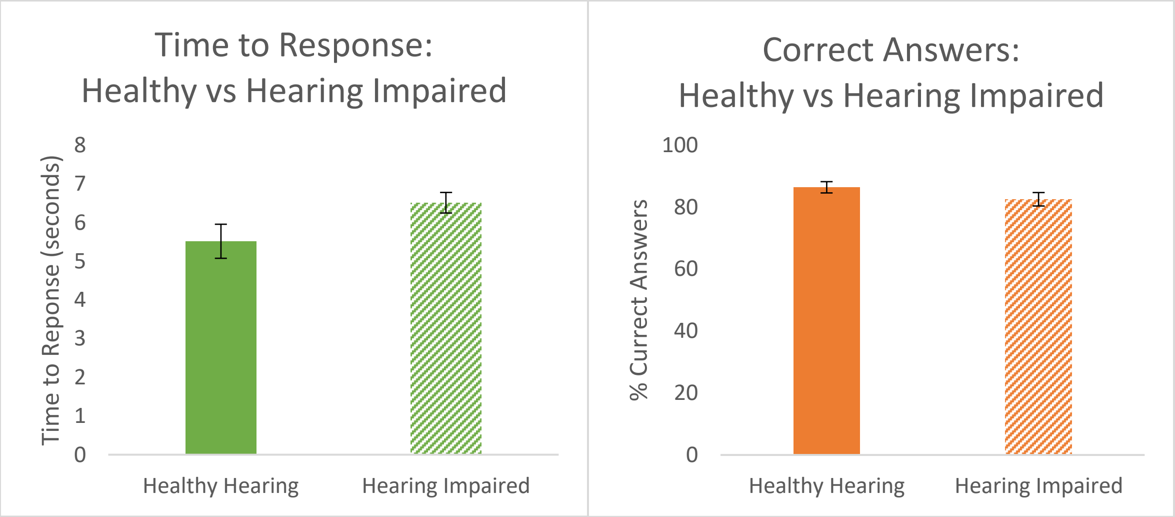
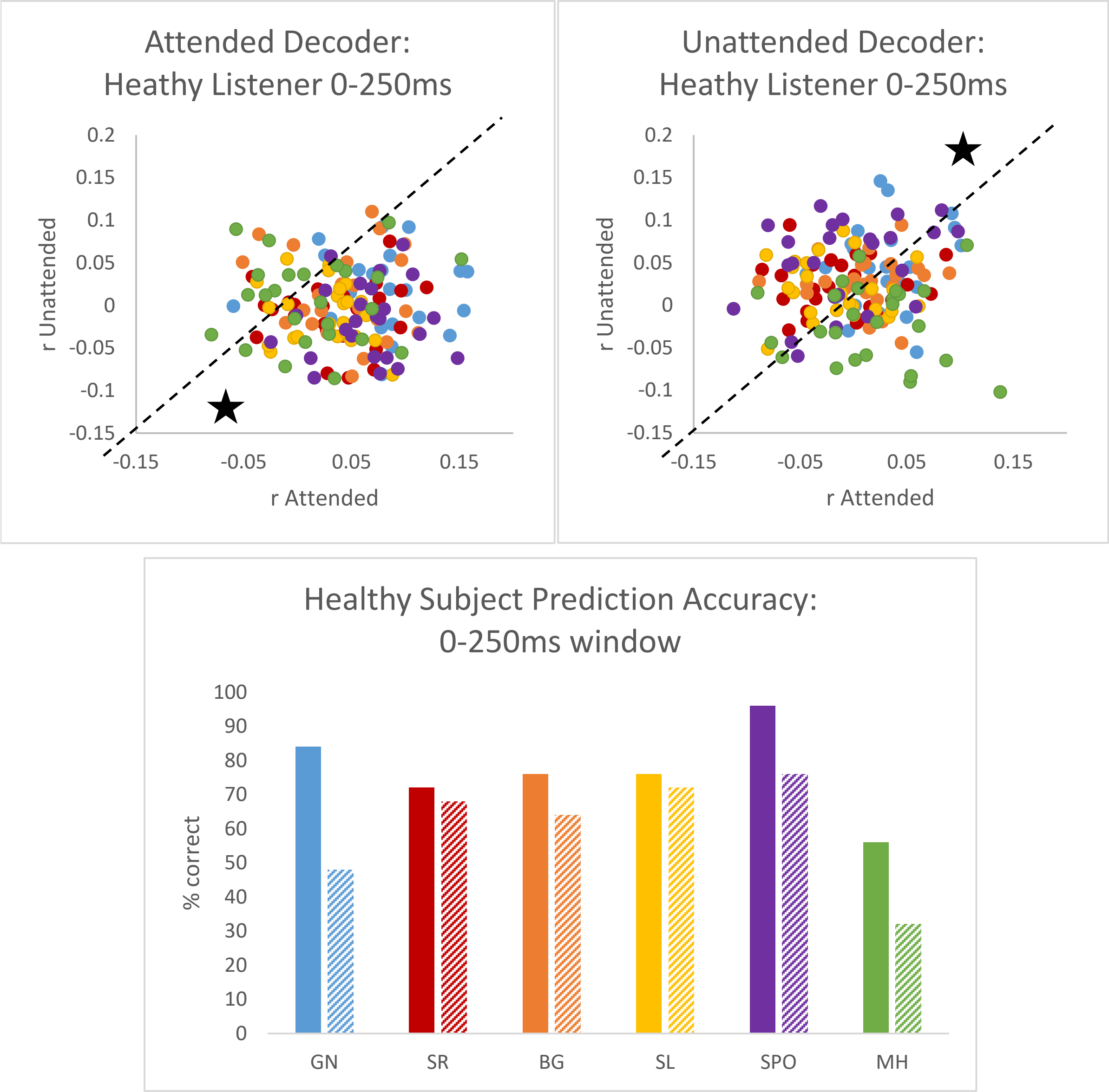


Figure 3.3. Illustration of the behavioral performance of both groups of subjects in the dichotic dual-speaker attention task. Healthy hearing (n=6) and hearing impaired (n=4) group performances in question answering correctness is illustrated in the top figure, and the time to keying in an answer to the question is illustrated in the bottom. Questions were asked after ~1min of dichotic listening to a single, pre-determined speaker, and they consisted of two, two-choice semantic questions.


Attended Decoder: Heathy Listener 0-250ms

Unattended Decoder: Heathy Listener 0-250ms

Healthy Subject Prediction Accuracy: 0-250ms window

Figure 3.4. Data from the healthy hearing subject group with a regression over a 0-250ms lag window. Top Left: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the attended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). Top Right: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the unattended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). All trials plotted across subjects (25x6). On both top graphs, the lines represent the decision boundary between a correct prediction and an incorrect prediction, and the star indicates the side of a successful prediction. Bottom: The attended (solid) and unattended (dashed) prediction accuracies across subjects(n=6).

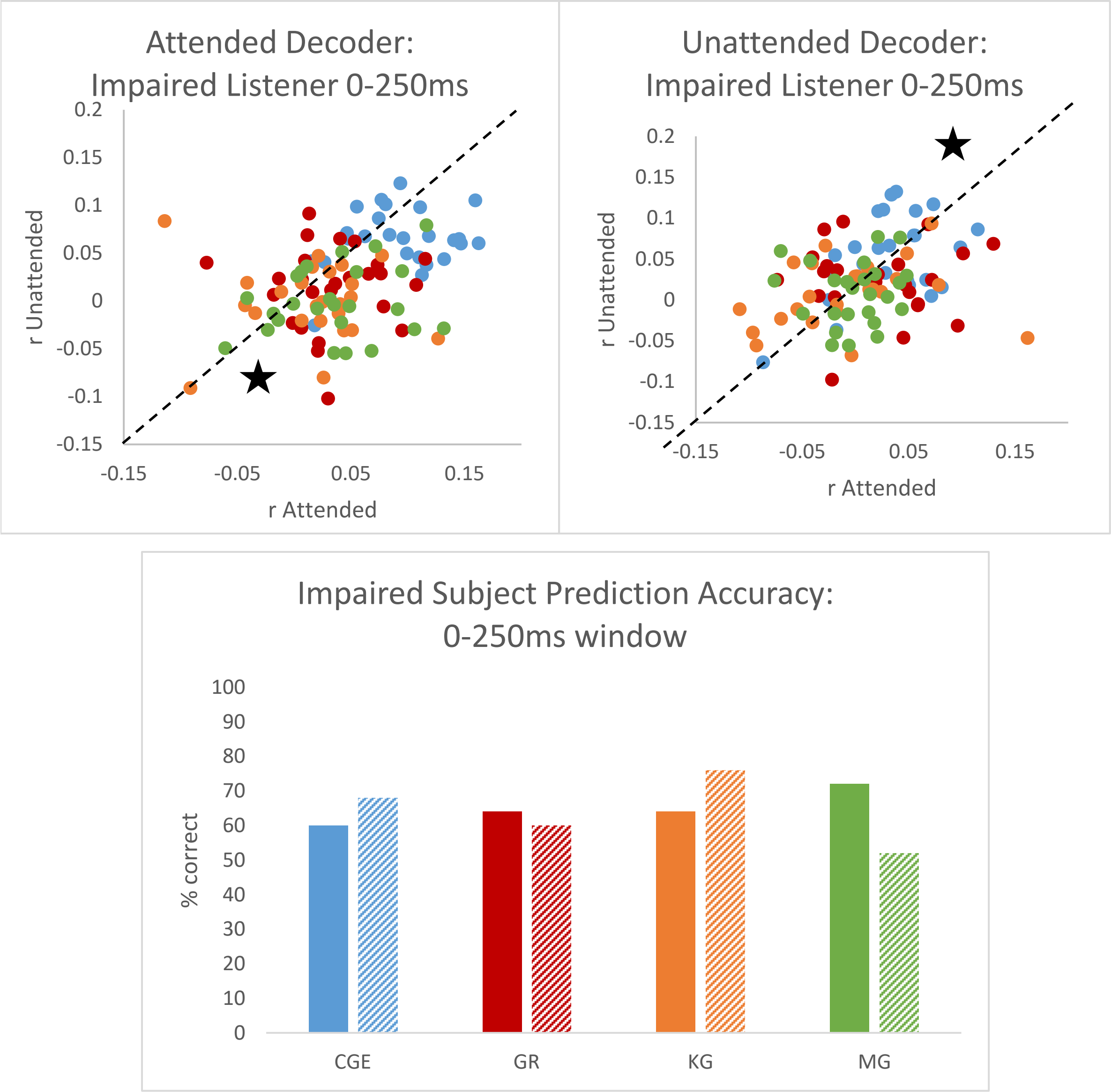


Figure 3.5. Data from the hearing-impaired subject group with a regression over a 0-250ms lag window. Top Left: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the attended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). Top Right: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the unattended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). All trials plotted across subjects (25x4). On both top graphs, the lines represent the decision boundary between a correct prediction and an incorrect prediction, and the star indicates the side of a successful prediction. Bottom: The attended (solid) and unattended (dashed) prediction accuracies across subjects (n=4).

However, when regressing from -100-500ms, the healthy hearing attended decoder accuracy increased to 80% (+3%) and the unattended decoder drops to 57% (-3%) accuracy (Figures 3.6 & 3.9). Likewise, the hearing-impaired subjects attended decoder accuracy increased to 74% (+9%), while the unattended decoder accuracies dropped to 48% (-16%) (Figures 3.7 & 3.9).

Interestingly, the absolute magnitude of the Pearson correlations seemed to change similarly between both groups, with a gain of ~0.013 for the attended decoders for both groups, and with essentially no change in Pearson correlation for the unattended decoders for both groups. This contrasts with the huge change in prediction accuracy in the hearing-impaired group (Figure 3.9).

C. Lag Analysis

Looking at figure 3.8, testing individual lag shifts (~7.8125ms) from -100-500ms, what we found is a somewhat steady time course of correlation across the range of tested lags in our healthy and unhealthy groups, even at lags preceding causality (i.e., -100-0ms). This is most likely since even at lag windows that could not have been correlated to our incoming stimuli there was still male speaker stimuli present (just not specific to the speech/sentence that was driving the cortical activity), and the $r > 0$ Pearson correlation is indicative of the generalizability of the model.

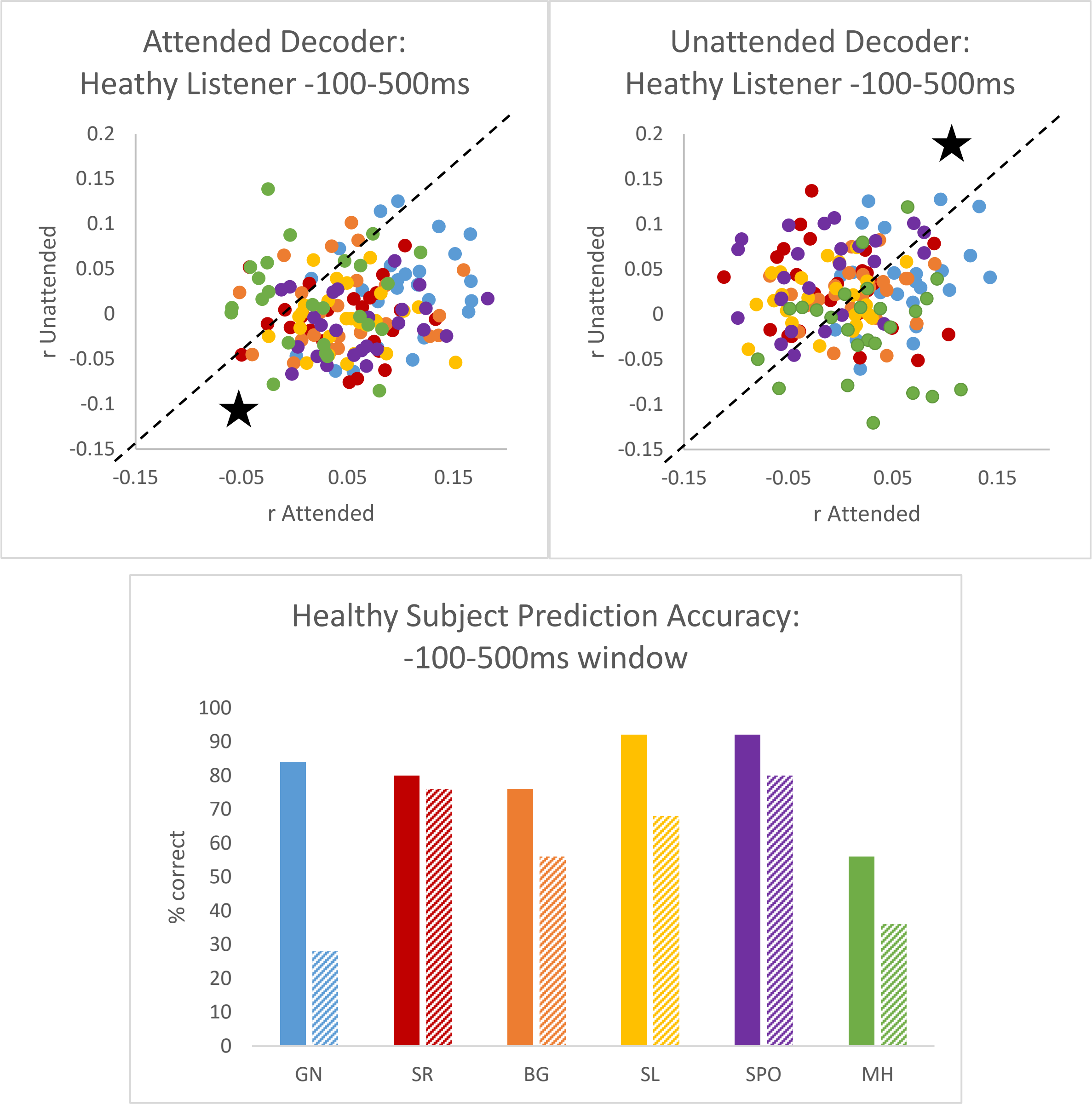
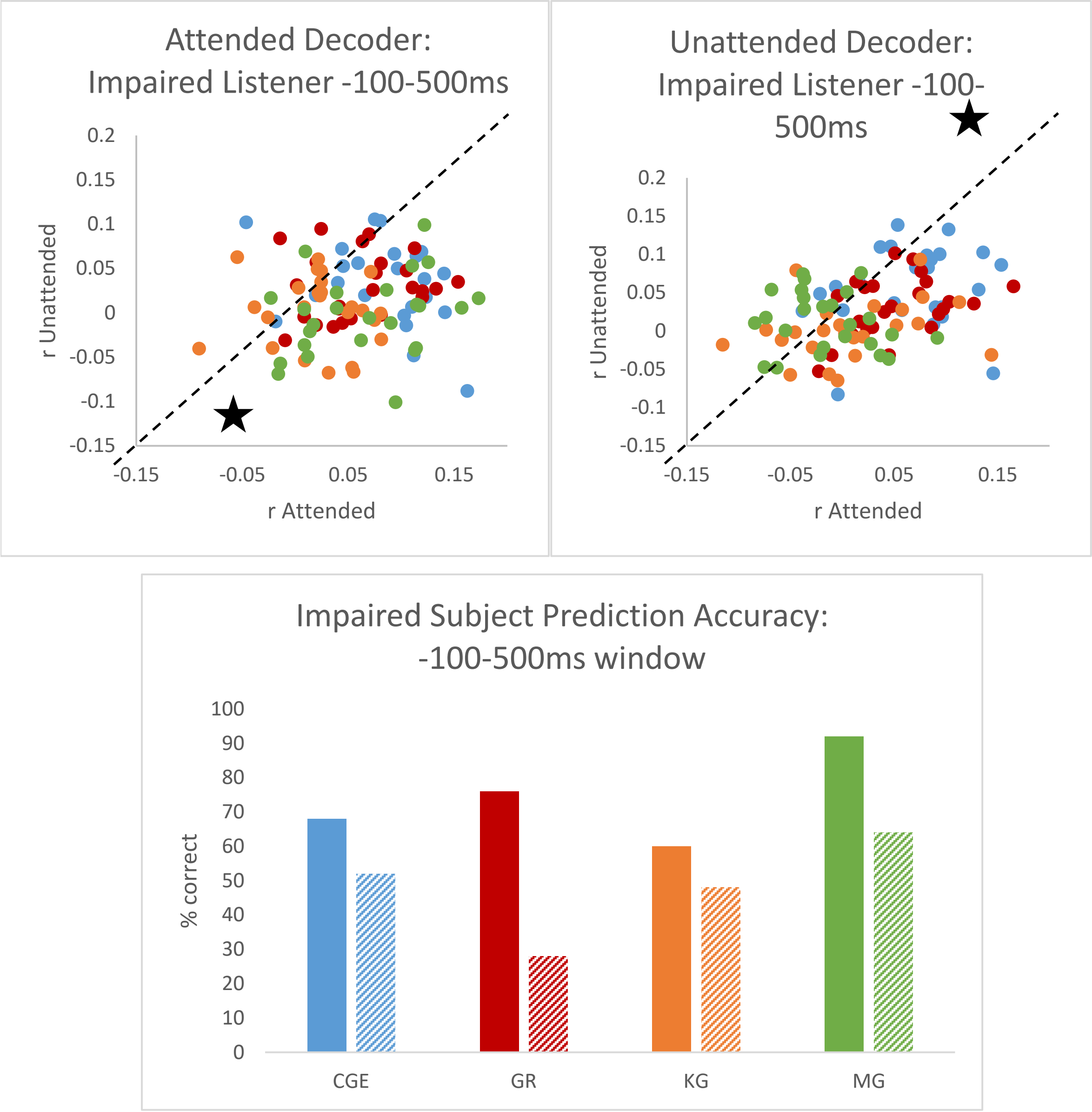


Figure 3.6. Data from the healthy hearing subject group with a regression over a -100-500ms lag window. Top Left: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the attended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). Top Right: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the unattended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). All trials plotted across subjects (25x6). On both top graphs, the lines represent the decision boundary between a correct prediction and an incorrect prediction, and the star indicates the side of a successful prediction. Bottom: The attended (solid) and unattended (dashed) prediction accuracies across subjects (n=6).


Attended Decoder: Impaired Listener -100-500ms

Unattended Decoder: Impaired Listener -100-500ms

Impaired Subject Prediction Accuracy: -100-500ms window

Figure 3.7. Data from the hearing-impaired subject group with a regression over a -100-500ms lag window. Top Left: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the attended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). Top Right: A scatter plot depicting the Pearson correlation coefficients for the envelope generated by the unattended decoder, and how that envelope correlates to the attended envelope (x axis) and the unattended envelope (y axis). All trials plotted across subjects (25x4). On both top graphs, the lines represent the decision boundary between a correct prediction and an incorrect prediction, and the star indicates the side of a successful prediction. Bottom: The attended (solid) and unattended (dashed) prediction accuracies across subjects (n=4).

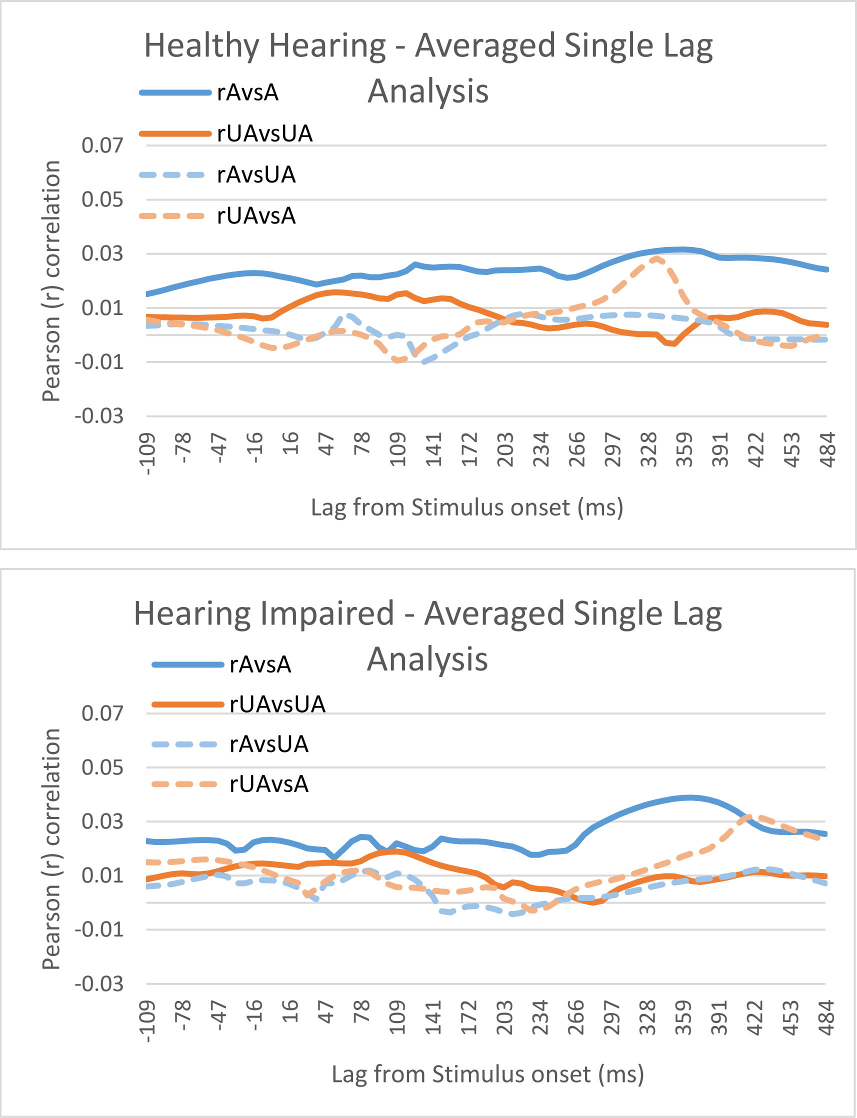


Figure 3.8. Data from a "Single-Lag" analysis taken over a range of time lags from -100-500ms for healthy hearing (top) and hearing impaired (bottom) groups. Plotted is the average across trials (n=25) and subjects (n=6; n=4) for 77 sets of models, each trained solely at a single lag point separated by 7.8125ms. Between the healthy hearing and hearing impaired, the rAvsA (solid blue) correlation curves between the two separate populations was highly correlated (r=0.78; p<.0001), as well as for the rUAvsUA (solid orange) correlation curves (r=0.74; p<.0001). The "mismatch" rAvsUA (dashed blue) and rUAvsA (dashed blue) curves were not correlated between groups.

rAvsA = Attended Decoder predicting the Attended Envelope

rUAvsUA = Unattended Decoder predicting the Unattended Envelope

rAvsUA = Attended Decoder predicting the Unattended Envelope

rUAvsA = Unattended Decoder predicting the Attended Envelope

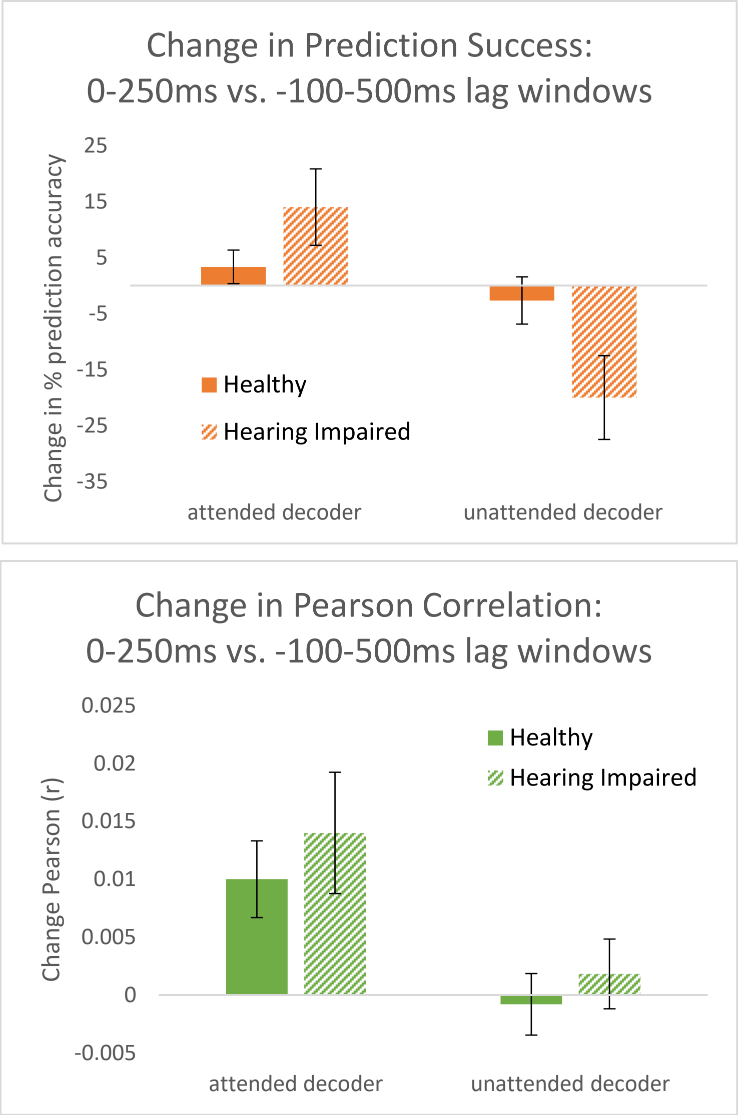


Figure 3.9. Illustrating the change in performance from regressions run from 0-250ms and -100-500ms. Top figure represents the group average change in prediction success between both healthy hearing and hearing impaired as well as between the attended decoder and unattended decoder. Bottom figure represents the group average change in Pearson correlation between both healthy hearing and hearing impaired as well as between the attended decoder and unattended decoder. Error bars represent the standard error of the mean (SEM).

What is more interesting though is that there seems to be a relative peak in correlation for the attended decoder (regarding the attended stimuli) at ~360ms, which is consistent between both groups. Likewise, the unattended decoder seems to have a peak correlation at 60-100ms (regarding the unattended stimuli), which is consistent in both groups. Further, the shapes of the curves for the attended decoder (regarding the attended stimuli) were 78% correlated between the hearing impaired and healthy hearing groups, and the unattended decoder (regarding the unattended stimuli) was 74% correlated (figure 3.8). This is true, even though there was large within group variance for the shapes of each of these individual lag curves. The group means between the groups seem to have settled on a high amount of correlation.

V. Discussion

The results from this study seemed to indicate that hearing impaired and healthy hearing individuals have speech processing occurring on differing time scales. As figure 3.9 illustrates, when expanding the lag window from 0-250ms out to 500ms, the healthy hearing individuals seemed not to have much of a change at all (only +-3%). On the other hand, with this expansion in the lag window, the hearing-impaired subject group had a tremendous change in performance, significantly increasing the performance in the attended decoder and decreasingly significantly the performance in the unattended decoder.

Figure 3.8 seems to allude to these results, as one can see in the hearing-impaired group the attended decoder curve reaches a maximum value out at the > 300ms time lag, whereas this is not seen to the same degree in the healthy hearing group. Thus, one would predict that keeping the later, and seemingly most significant, time lags in the analysis (instead of discarding them by only regressing from 0-250ms) would strongly boost the attended decoder's performance, which was the case.

Further, given that the purpose of this study was in large part to examine and replicate the findings by O'Sullivan and Lalor with regard to our experimental paradigm and machine learning toolkit, the results we found seemed to closely match their study (O'Sullivan et al., 2015a). Our study did differ in a few key components, however. First, O'Sullivan's study delivered stimuli using over-the-ear headphones, whereas to avoid speaker artifacts showing up in the EEG signal, we used Etymotic ER-3a scientific grade earphones. Second, O'Sullivan's study did not attempt to clean their EEG data of eye blink artifacts or bad channels from what we can tell, and from what Ed Lalor shared during a question-and-answer session, whereas we did remove eye blink artifacts and removed and interpolated bad channels. Further, they used 128 electrode channels, whereas we only used 64 electrode channels.

A third difference between O'Sullivan's study and our own was the way in which the data was preprocessed. O'Sullivan filtered the EEG data from 2-8Hz and found all his significance in that range; however, when we tried his analysis steps using that same filtering range our results were not significantly different from chance, i.e., guessing the attended talker. We found, however, if we broaden the filter to include very lower frequencies (0.5-8Hz, and later to 0.5-15Hz) (Di Liberto et al., 2015; O'Sullivan et al., 2015b), our prediction accuracy and the magnitude of our Pearson correlations increased tremendously, thus indicating the importance of such low frequencies in this analysis. Unfortunately, it leads to the question of how O'Sullivan attained the results he did without the use of these low frequencies.

And the last major difference between the O'Sullivan study and our own is illustrated in figure 3.8. In what was one of the last and final conclusions of O'Sullivan's paper, he made the claim that ~220ms was the most highly significant time lag regarding contribution to the overall correlation of the stimulus reconstruction. Our results in our healthy subjects seem to differ. While in his study he has a clear bell curve with strong peak right at ~220ms averaged across his subjects on his attended decoder, on our we see a flat curve with the only semblance of peak at ~360ms. Such data seems to loosely contradict with O'Sullivan's results ("loosely" as he never extended his analysis out to >250ms), but seems to match new literature in the field as to the long latency contributions to speech perception (Akram et al., 2014, 2016). Further, we found huge variability in the single lag analysis from one individual to another, and no single one of our subjects showed a strong peak at the same time points that O'Sullivan had shown (O'Sullivan et al., 2015). Though, with our small sample size it is hard to say how our results would change if we were to expand this out to the n=40 that was used in the prior study. However, while O'Sullivan did not report error bars as to his subject variance towards his ~220ms claim, he later responded (personal communication) that his study did in fact have large variability.

Nonetheless, the questions and observations raised by our study are highly interesting in that for the first time we can show that with an above change efficacy hearing impaired individuals can have their attention decoded akin to that of healthy hearing, albeit with a few deviations in the process therein. And what is even more interesting, is that our results seem to show a strong effect at >300ms latencies for cortical activity in response to speech in selectively our hearing-impaired sample.

Future directions could expand upon this work by increasing the hearing-impaired subject pool to see how these trends hold when taken to a statistically more powerful level. Likewise, while there has been work by Lalor et al. in combining the prediction information stored in both the attended and unattended decoders into a *combined* decoder, pursuing this line of work further might be beneficial for boosting the accuracy of our predictions even further. Lastly, using non-linear methods of reconstructing the speech stimuli could perhaps bring this performance up towards near 100% prediction accuracy across subjects. That, and optimizing the processing performance (perhaps by selectively choosing a small subset of crucial lag points) will be necessary to take this technology into the embedded, portable realm of application, where technology such as this could be a crucial element in the next generation of hearing prostheses.

Chapter 4

“Tracking fatigue dynamics during prolonged auditory attentional switching with EEG and pupillometry”

I. Background

Everyday acoustic scenes are complex and dynamic. When listening to others talk and when following a conversation in a realistic setting, we must exert sustained effort while rapidly switching our attention, which can lead to auditory attentional fatigue over prolonged periods. The neural mechanisms underlying auditory attentional switching are only beginning to be understood (Larson and Lee, 2013; Getzmann et al., 2015, 2016), particularly with regard to attentional fatigue, and the mechanisms have not been related systematically to the behavioral dynamics of realistic auditory scene. This knowledge gap raises a profound barrier to addressing real-world, daily communication challenges encountered by healthy listeners in noise, older adults, children with listening difficulties, and those with hearing loss. In this study we created a scenario to induce auditory attentional fatigue by forcing a prolonged, one hour session of near constant auditory attentional switching. Using this paradigm, we aim to look at the prolonged attentional switching and fatigue dynamics of older, healthy hearing adults (greater than age 50) compared to healthy hearing younger adults (less than age 30).

A. Perceptual Objects

When listening, our brain transforms complex time and frequency auditory information captured at the cochlea into meaningful perceptual objects (Griffiths and Warren, 2004) based on spatial and non-spatial cues, e.g. voice characteristics and patterns of stress and intonation, known as prosody. This perceptual object formation is crucial for auditory attention, and for auditory attentional switching. Perceptual “objects” are a concept that much of our understanding of cognitive attention builds from, and one of their better definitions comes from Barbara Shinn-Cunningham, who states that "[perceptual objects are a] *perceptual estimate of the sensory inputs that are coming from a distinct physical item in the external world*" (Shinn-Cunningham and Best, 2008b). The brain tends to group many single point sources of energy, be it light, touch, or, in our case, sound, into clusters which it predicts emanate from a singular, physical unit. This is a tremendously complex task, especially in the case for audition where a sound field almost assuredly contains echoes, which will each exhibit their own independent spatial information unique the physical layout of the environment, and this is not addressing the inherent ambiguity that exists in the case of auditory sound localization. Nevertheless, a healthy brain can continuously disambiguate the dynamically changing time, frequency, and locational information into singular groupings that it estimates are a part of one object, which could be a human voice, noise coming from a fan, a musical instrument, etc.

B. Selective Attention and Switching

Once auditory object formation has occurred, only then can the brain choose a given object as the focus of its attention. It does this by selectively suppressing the higher-level cognitive processing, in the case of conversation this would be language processing, on all objects and noise other than the lone desired target (Rudner et al., 2015). This process of sustained auditory selective attention and language processing in a noisy environment is referred to as the "cocktail party problem", and it is a demanding task for the brain to overcome. However, with the need for repeated attentional switches comes not only the burden of attention disengagement from the current auditory object and reengagement on the new object, but it also comes with the need for pre-attentional processing through which the brain must make on-line determinations about when to switch attention and to whom the target should go. Likewise, the performance of pre-attentive processing has been shown to be affected by neurological fatigue states (Yang et al., 2013).

Successful auditory attention and switching the target of auditory attention are crucial to understanding the semantic meaning in conversational environments, as “cocktail party” environments are rarely static and devoid background auditory distractions. Importantly, dynamic attentional switches significantly impair speech comprehension (Rudner et al., 2015), as much of the information during the period of time during and immediately following the attention switch in attention is lost, and it never reaches the higher level stages of speech processing. Thus, to successfully understand the content in an ongoing conversational speech stream, not only do the associated mental processes of engaging and disengaging attention from auditory objects need to continuously be taking place, but the brain must also be making repeat predictions about the information content being missed during the periods between switching. Unfortunately, the neural mechanisms of attentional switching to speech remain unclear, although several recent studies show that noninvasive neural measures such as electroencephalography (EEG) or magnetoencephalography (MEG)) can track sustained auditory attention and indicate to which talker a listener attends (Larson and Lee, 2013; O'Sullivan et al., 2015a). These signals therefore reflect either attentional control – the "top-down", volitional directing of attention, especially to the talker's location in space – or attentional modulation of the speech representation itself. Further, one well established measure of attentional control is EEG power in the alpha (~8-13Hz) band over occipito-parietal cortex. When a listener directs attention to a talker on the left, alpha power increases over the left scalp and decreases over the right (and vice versa) (Kerlin et al., 2010), similar to visuospatial attention (Worden et al., 2000).

C. Cognitive Fatigue

Over prolonged time, repeat attentional switches take a high demand on an individual's cognitive faculties, which is referred to as "mental fatigue" or "cognitive fatigue." Cognitive fatigue usually manifests in the form of deteriorated task performance, reduced motivation to continue on a task, and an increased difficulty in keeping attention focused with an increasing chance for distraction (Faber et al., 2012). Like auditory attention, the mechanisms underlying cognitive fatigue are yet to be fully understood. Many studies have tried to define cognitive fatigue in terms of progressive deterioration of cognitive resources, akin to muscular fatigue, while others have defined it in terms of terms of motivational state and effort/reward imbalance, i.e. cognitive fatigue is a consequence of the perceived effort in a task being proportionally larger than its associated benefit (Gergelyfi et al., 2015). However, neither of these paths to explanation have been fully explored or they do not match much of the evidence that exists. For example, in the case of cognitive fatigue being a product of motivational state, monetary incentives being introduced after fatigue had been induced failed to recover pre-fatigue task performance (Boksem et al., 2005; Gergelyfi et al., 2015). Nonetheless, despite not fully understanding ‘what’ drives cognitive fatigue, there are established and reliable measurements of cognitive fatigue that can be used to track fatigue onset over time. First, performance on mentally challenging tasks has been shown to decrease proportionally with the degree of fatigue (Gergelyfi et al., 2015). Secondly, cognitive fatigue has been shown to be tightly correlated with increased global frontal theta band (θ, 4-7Hz) rhythmic activity, as well as an increase in global parietal alpha (α, 8-12Hz) rhythms (Gergelyfi et al., 2015; Trejo et al., 2015).

Additionally, another key measure that has yet to be explored while studying cognitive fatigue is pupillometry, e.g., pupil size as a function of fatigue. Multiple studies have used changes in pupil size as a powerful measure for task difficulty (Piquado et al., 2010; Zekveld et al., 2014; Demberg and Sayeed, 2016); however, there have been little to no published studies that directly equate changes in pupil size to the level of cognitive fatigue. It has been found that pupil size increases as a function of task difficulty, to which people have associated the increase in pupil area as being proportional to cognitive load. Though, in the case of fatigue, is it not conceptually plausible that cognitive load could also increase in the case of an unchanging task difficulty but with the add-in of increasing cognitive fatigue? In which sense, pupil size would increase in-kind as a function of fatigue, just as it did with an increase in task difficulty. Disambiguating these two concepts, or whether a task done in a fatigued state has the same cognitive load as a harder task done in an unfatigued state, has yet to be shown in the literature.

D. Hypothesis

We hypothesize that over the course of the study, the onset of fatigue will be measured with a decrease in behavioral task performance, a decrease in pupil diameter as measured with pupillometry, and increases in parietal alpha (8-12 Hz), frontal alpha (8-12 Hz), and frontal theta (4-7 Hz) band activity as measured with EEG.

II. Methods

A. Stimulus Preparation and Delivery

The paradigm consists of two one-hour long audio/video recordings of a voice actor reading Jules Verne’s “Journey to the Center of the Earth” and “20,000 Leagues Under the Sea”, delivered concurrently side-by-side against one other. The voice on the left was pitch-shifted up six percent while the voice on the right was pitch shifted down three percent to aid in spectral separation of the two voices, since the same voice actor voiced both stories (Figure 4.1). To execute our attention switching dynamics, approximately every six seconds the story contents would switch from the left speaker to the right speaker, or vice versa, for which the participant was cued and would therefore switch their attention from one talker to the other. If the participant was properly executing the task they should experience one single fluid story, shared between the left and the right talker. If the participant did not execute the task and were to listen to solely a single talker, they would experience the voice switching between two separate stories continuously.

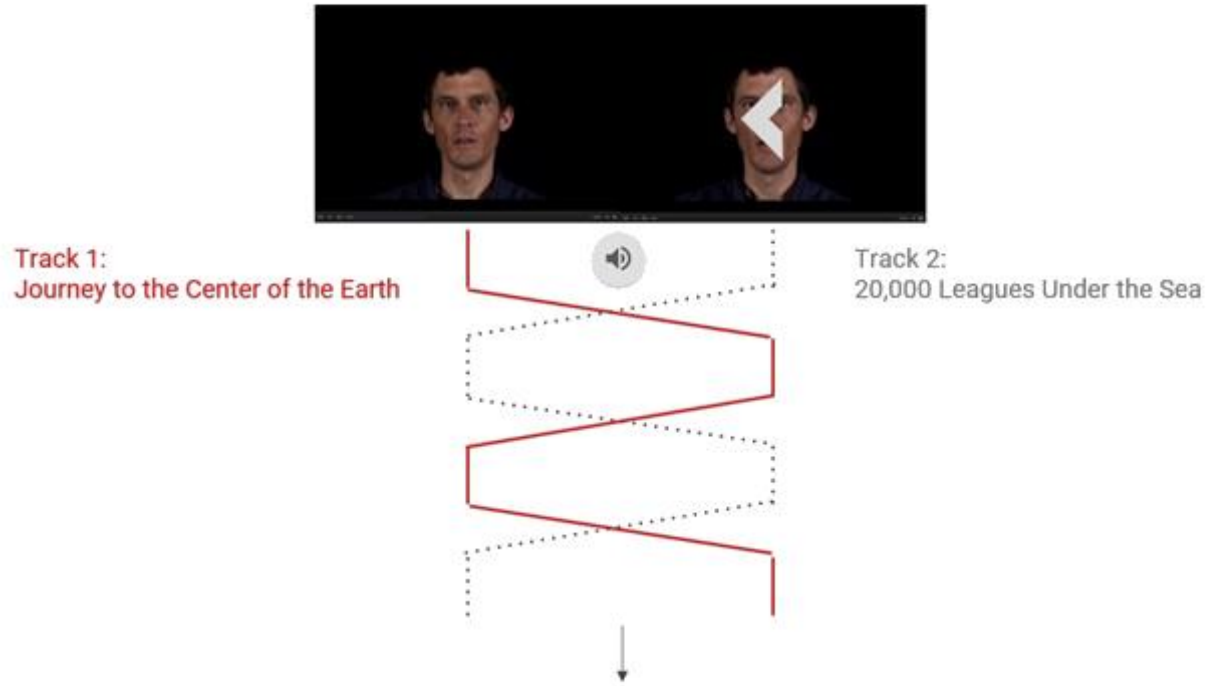

Track 1:
Journey to the Center of the Earth
Track 2:
20,000 Leagues Under the Sea

Figure 4.1. Example illustrating switching dynamics of the attentional paradigm. The red track is the story the participant is tasked to attend onto. Switches happen approximately every six seconds.

As a measure of attentional switching delay over the course of the study, we embedded 200 cue words into the “Journey to the Center of the Earth” story that the participants were instructed to listen to and to press the <space> bar on the keyboard when heard. These 200 cue words were simple, monosyllabic color words, i.e., red, blue, green, pink, white, etc. To accommodate these words into the semantic content of the story, the section of story was first scrubbed of any color references, and 200 new color words were inserted spaced approximately evenly, 100 in the first half of the story and 100 in the second half. Because the cue words were worked into the semantic flow of the story it was impossible to evenly distribute the cue words, but we felt this deficit was worth the benefits of having semantically appropriate sounding attentional cues worked into the story itself.

The stimuli were recorded in a sound-controlled booth using a Sony camcorder and RØDE microphone powered by a Fireface 800 sound interface at 30fps and 48kHz, respectively. The stimuli were edited, removing mistakes and extraneous silent gaps, in Adobe Premiere. Each video and corresponding audio track needed to be interlaced with the other at precise time points, one switch every ~6 seconds, and this was executed using custom scripts written in MATLAB. Likewise, each video switch was accompanied with five frames of blending, and each audio switch was accompanied with 0.166 seconds of audio crossover, the length of 5 video frames, to remove any hard transition effects and/or audio clipping. Lastly, the two ‘new’ tracks, i.e., the newly interlaced left and right voice/video tracks, were pitch shifted using Audacity.

Importantly, 200 of the 600 story switches corresponded to the 200 color words that were inserted into the story, and the switch times occur at precise intervals relative to the time of the color word. The switches happened at either 0.25s, 0.5s, 1.0s, 1.5s, or 2.0s before the cues, with these switch time intervals being pseudo-randomly distributed amongst the first half and the second half of the study in equal proportion.

The stimulus delivery paradigm was coded in MATLAB using Psychtoolbox for the audio and video delivery as well as for capturing keyboard inputs from the participants. The voice and video tracks were presented separately from one another, the audio making use of Psychportaudio for precision audio with <4ms of latency and the video making use of Screen (part of the Psychtoolbox package) for precise frame timings. The visual stimuli were presented on two Dell monitors, one face per monitor, centered. The audio stimuli were presented through two Tannoy Precision 6 studio monitors powered by a Stewart stereo power amplifier placed just to the outsides of the Dell monitors. The left voice was played through the left studio monitor and the right voice was played through the right studio monitor, giving the effect of the voice emanating from the face on the screen for each screen.

B. Experiment Design

The study consisted of six blocks, each approximately 10 minutes in length, and within each block there were 100 auditory/visual attentional switches. The participants were instructed to pay attention solely to the story “Journey to the Center of the Earth”, to which an arrow on the screen cued them as to which speaker they should be attending to. Upon each attentional switch, an arrow would appear on top of the newly distracting speakers face, signaling for the participant to switch their attention to the other talker (Figure 4.1).

Alongside the attention switching, the participants were asked to listen for monosyllabic color words in the story, i.e., red, blue, green, pink, white, etc. When a color word was detected, the participants were told to press the space bar on the keyboard in front of them. The purpose of this was three-fold: 1) it allowed us to detect the time until attention-reengagement, with our hypothesis being that as fatigue sets in participants would be more likely to miss the cue words with shorter switch latencies, 2) it allowed us to detect their response time latency, and 3) it kept the participant engaged throughout the duration of the study.

Furthermore, between blocks participants were asked to perform a simple visual reflex task that consisted of five trials in quick succession where they would stare at a crosshair on the screen and push the space bar on the keyboard once a white circle flashed around the crosshair.

Lastly, participants were given instructions both verbally and visually on the screen and were provided a small one-minute practice round of attention switching before the blocks began to clarify their task. Between blocks, participants were provided a short rest period before continuing with the study. During this time the experimenter verbally spoke with the participants and received consent to continue.

C. Data Collection and Processing

Both multichannel EEG and pupillometry data was collected of the course of the study alongside the behavioral response data mentioned earlier.

EEG data was collected using a Biosemi Active 2 with the ActiView data acquisition software. Recording consisted of 32 channels at 2048Hz, re-referenced to the left and right average mastoid. The EEG data was imported into EEG Lab running on top of MATLAB 2020a. The data was first re-referenced to the mastoids from the original CMS/DRL as provided by Biosemi. Visibly bad channels were removed and interpolated, albeit in this limited series of data we had remarkably no bad channels to remove. ICA was performed to isolate and remove eye blink components from the data set. The data was band-pass filtered between 0.5 and 40Hz to remove high frequency noise and low frequency drift. And lastly, the channel location data was updated from the simple Biosemi naming convention A1->A32, to the international 10-20 system.

Likewise, for the purposes of this study “frontal” and “parietal” electrodes consist of left and right hemispheric clusters of three electrodes, whose activity was averaged for analysis purposes. The clusters were as follows:

Frontal Left: F7, F3, AF3

Frontal Right: F8, F4, AF4

Parietal Left: P7, P3, CP5

Parietal Right: P8, P4, CP6

Pupillometry was recorded using Pupil Labs’ Pupil Core headset and making use of the Pupil Capture software. We used a binocular setup with a front, world-view camera. Each camera recorded at a resolution of 1920x1080 at a sampling rate of 30Hz. Calibration was achieved using the “physical marker” calibration process (as opposed to the “screen marker” approach), and a target was placed in between the two monitors the participants were facing with fixation points appearing on each of the two screens. To sync the pupil data with the EEG data, participants were asked to blink 10 times at the beginning of the study, to which we could compare to the EOG channels in the EEG and then time-align both data sets. Blinks were detected using the Pupil Player software using the Blink Detection plugin with a confidence interval set to 0.8.

D. Participants

Two participants were used in this pilot study, henceforth known as P01 and P02. P01 was a 29-year-old female with healthy hearing and no vision correction. P02 was a 53-year-old male with healthy hearing and glasses, which were removed for the duration of the study. Both subjects reported moderate tiredness at the beginning of the study, but both completed the full study without issue.

III. Results

A. Behavioral Response

When comparing the percentage of cue words detected over time for subject P01 and P02, we find no noticeable trend either downwards or upwards for subject P01, but subject P02 trended consistently upwards (Figure 4.2). This seems to indicate that subject P02 did consistently better from block to block, which would seem to run counter to our hypothesis that fatigue leads to worse task performance.

When looking at the latency of response to the cue words, we see no noticeable trend in either subject, with them both hovering around 2 seconds across all blocks (Figure 4.3). The same is true when looking at the reflex response latency to a flash of light. No noticeable trend shows across blocks with both subjects responding between 0.25 and 0.3 seconds on average (Figure 4.4).

B. Pupillometry

Regarding the diameter of the pupil across blocks, we hypothesized that a subject's pupil would decrease in size as a function of fatigue. Unfortunately, there was an issue with the recording of subject P01's pupil data, so we do not have that dataset to compare against. However, the data collected from P02 does seem to show a downward trend in pupil size, plateauing starting at block 4 (Figure 4.5).

C. Parietal Alpha

When looking at the alpha (8-12 Hz) activity of the parietal cortex of our subjects, subject P01 showed no trend downwards or upwards across blocks (Figure 4.6). Contrary, subject P02 showed a very strong and consistent increase in the parietal alpha power, which is consistent with our hypothesis.

When looking at the average time course difference for the parietal alpha activity, time locked to the onset of the switch cues, we find that subject A01 had very strong downward inflections upon leftward switches and very strong upward inflections upon rightward switches (Figure 4.7). Subject P02 does not show the same consistency between blocks and no conclusion can be drawn from their data (Figure 4.8). This measure likely needs many more participants before trends begin to appear.

D. Frontal Alpha

When looking at the alpha (8-12 Hz) activity of the frontal cortex of our subjects, subject P01 showed no trend downwards or upwards across blocks (Figure 4.9).

However, subject P02 showed a very strong and consistent increase in the parietal alpha power, which, again, is consistent with our hypothesis.

When looking at the average time course difference for the frontal alpha activity, time locked to the onset of the stimulus, we find that both participants have positive upward inflections for both leftward switches and rightward switches (Figures 4.10 and 4.11). This is interesting in two ways: 1) this activity pattern is entirely inconsistent with our parietal alpha results, in that regardless of the direction of the attention switch the inflection direction is the same, and 2) this response is very strong and very consistent for both participants, whereas our parietal alpha results had no noticeable trend for subject P02.

E. Frontal Theta

Unlike the previous two measures, neither participant showed any noticeable trend when looking at the theta (4-7 Hz) activity of the frontal cortex (Figure 4.12).

When looking at the average time course difference for the frontal theta activity, time locked to the onset of the stimulus, we find that both participants have positive upward inflections for both leftward switches and rightward switches (Figures 4.13 and 4.14). This pattern looks very much like our frontal alpha results; however, the size of the peaks is twice more than twice that of our alpha results. This may be insignificant, but it is interesting, nonetheless.

IV. Discussion

With only two subjects we cannot draw any significant conclusions from this data. That said, the data and the results presented are encouraging in that for at least one of our subjects our results matched our hypotheses on multiple occasions. It is also interesting that our younger participant's data was more consistent to itself and less like our predicted hypothesis when compared to the older participant. An immediate explanation could be that perhaps the younger subject did not become fatigued during our paradigm to the same degree as the older participant. Given that our hypothesis is under the assumption that that the participant is in fact becoming fatigued, a subject resisting fatigue over the hour time course could explain the null results. Only with more participants will be able to determine if this is in fact the case, and whether the paradigm is fatiguing enough for younger participants.

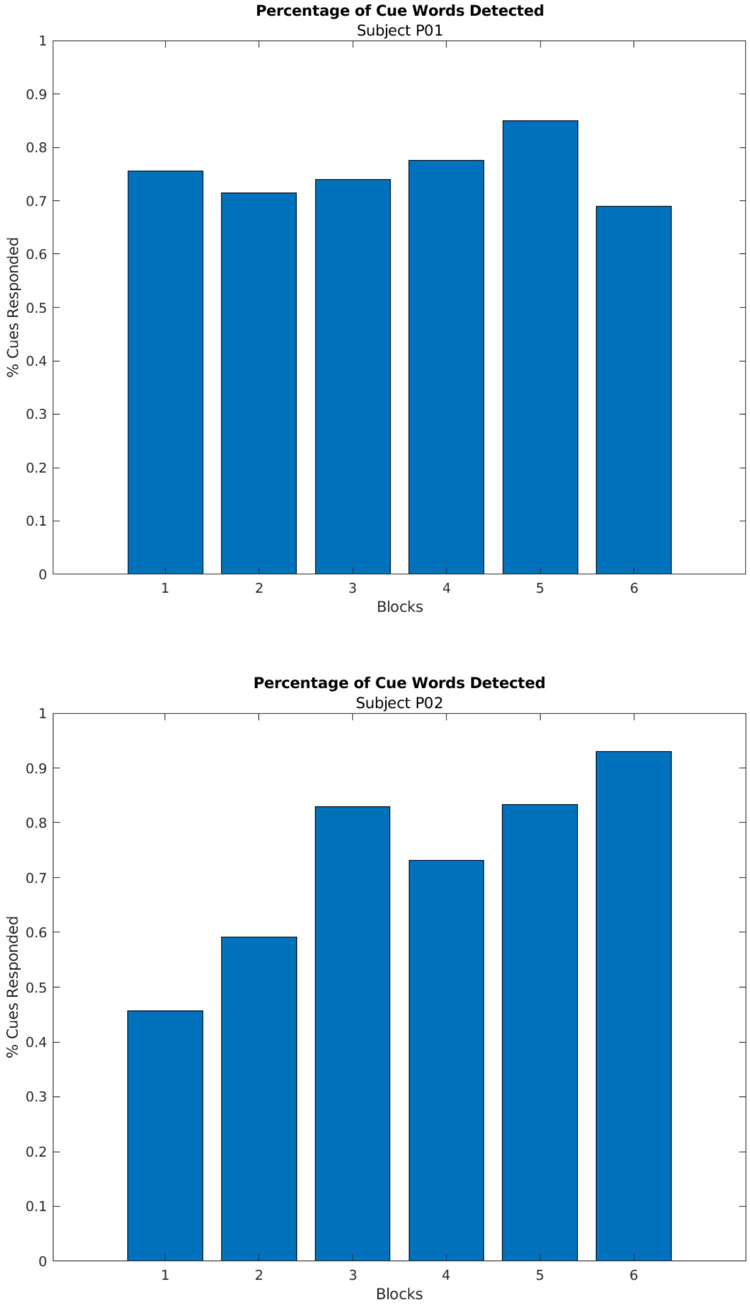


Figure 4.2. Comparison between subjects P01 (top) and P02 (bottom) as to their percentage of cue word detections over the course of six experimental blocks.

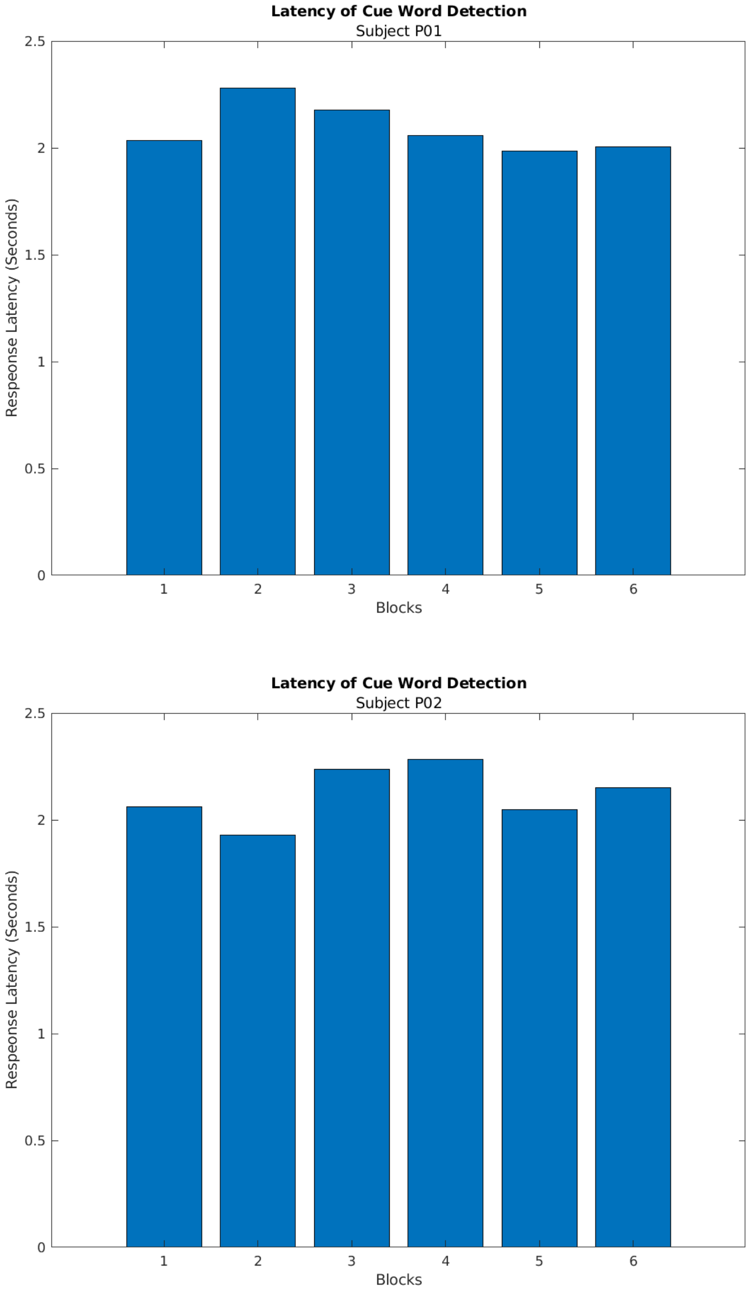


Figure 4.3. Comparison between subjects P01 (top) and P02 (bottom) as to their latency (seconds) of the cue word detection over the course of six experimental blocks.

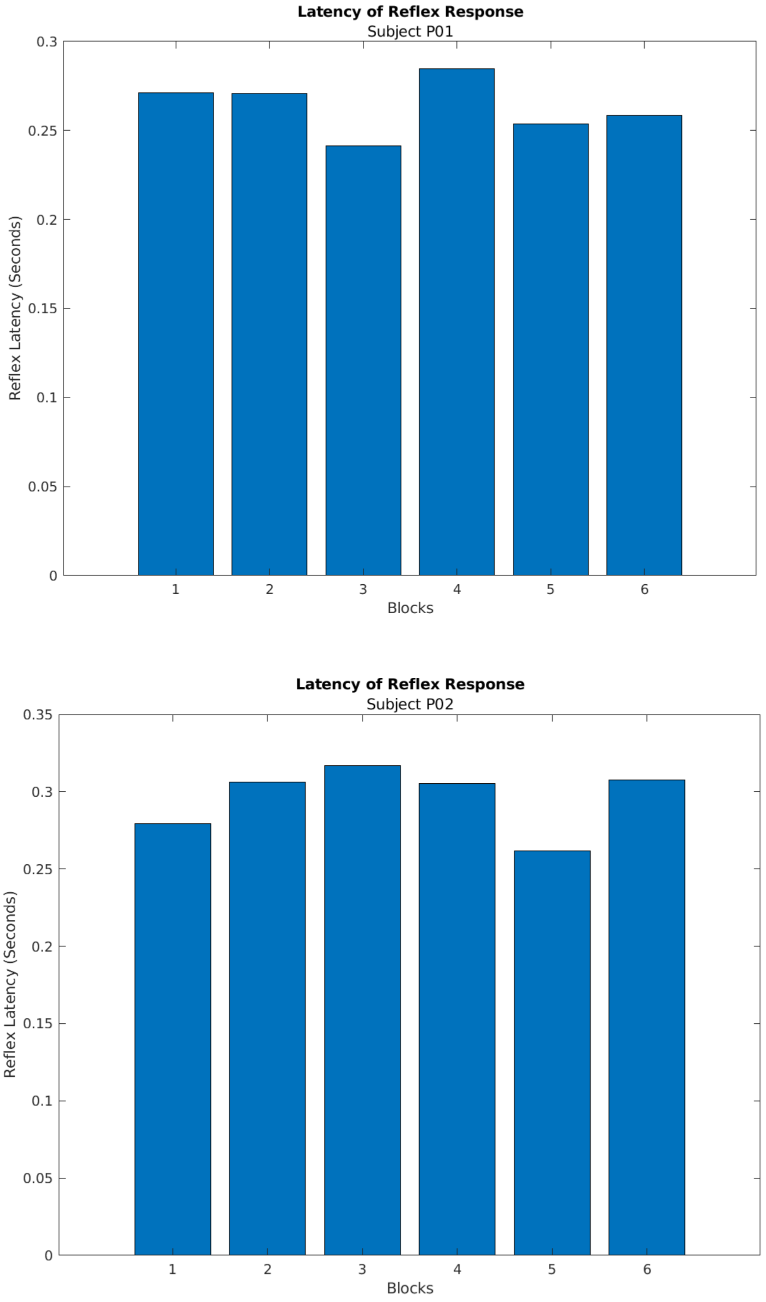


Figure 4.4. Comparison between subjects P01 (top) and P02 (bottom) as to the latency (seconds) of their reflex responses to a flash of light over the course of six experimental blocks.

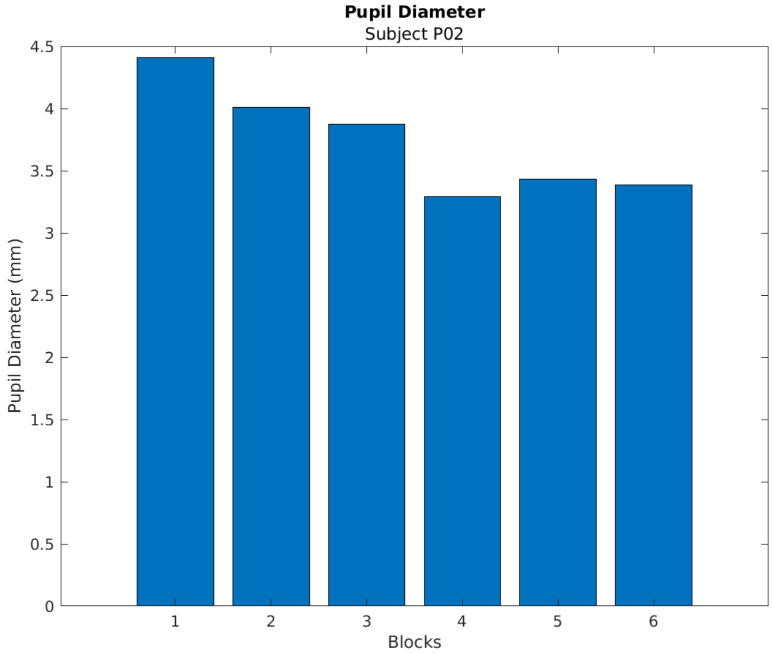


Figure 4.5. Subject P02’s pupil diameter (millimeters) over the course of six experimental blocks. *Note, there was an issue in the collection of subject P01. Their data was redacted accordingly.

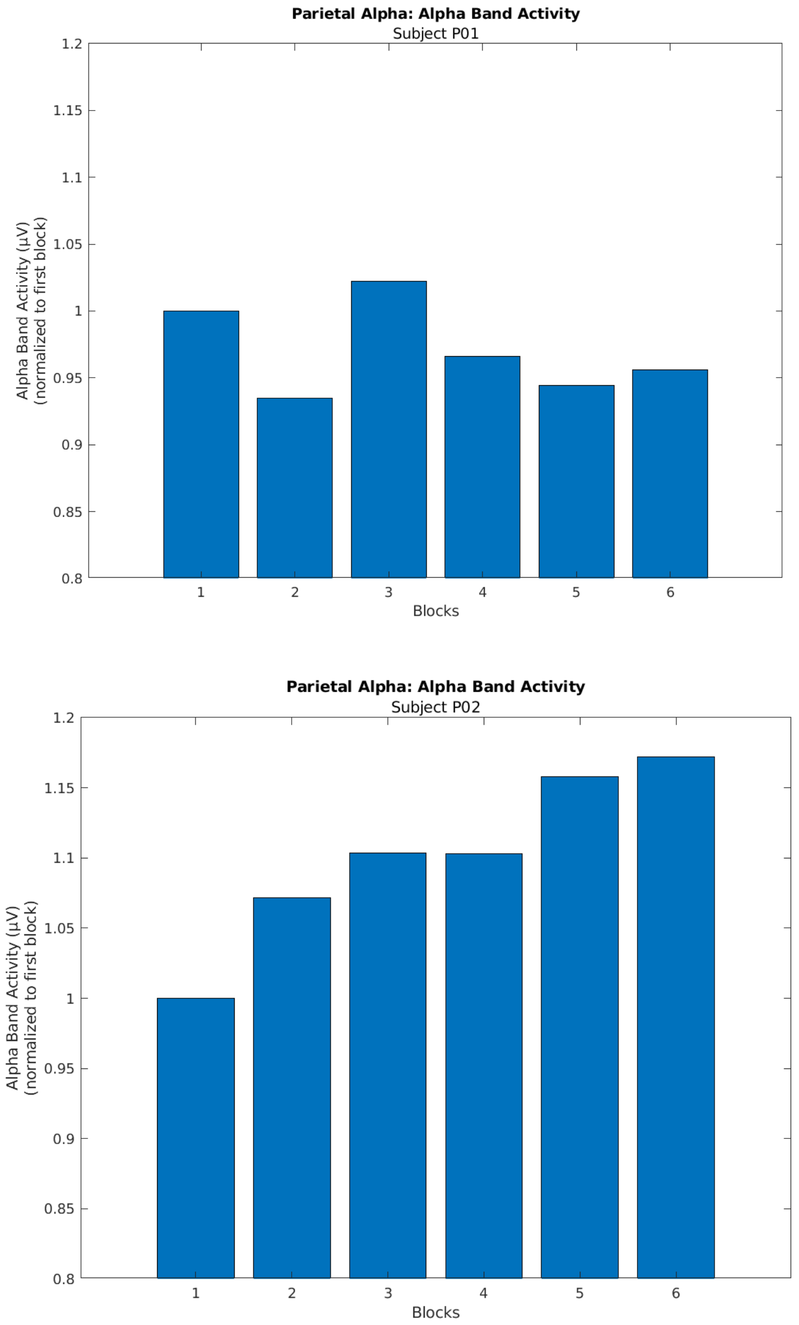


Figure 4.6. Comparison between subjects P01 (top) and P02 (bottom) as to their average parietal alpha band (8-12 Hz) activity (microvolts) over the course of six experimental blocks. Alpha band power was averaged across three left hemisphere and three right hemisphere parietal electrodes. Parietal Left: P7, P3, CP5; Parietal Right: P8, P4, CP6

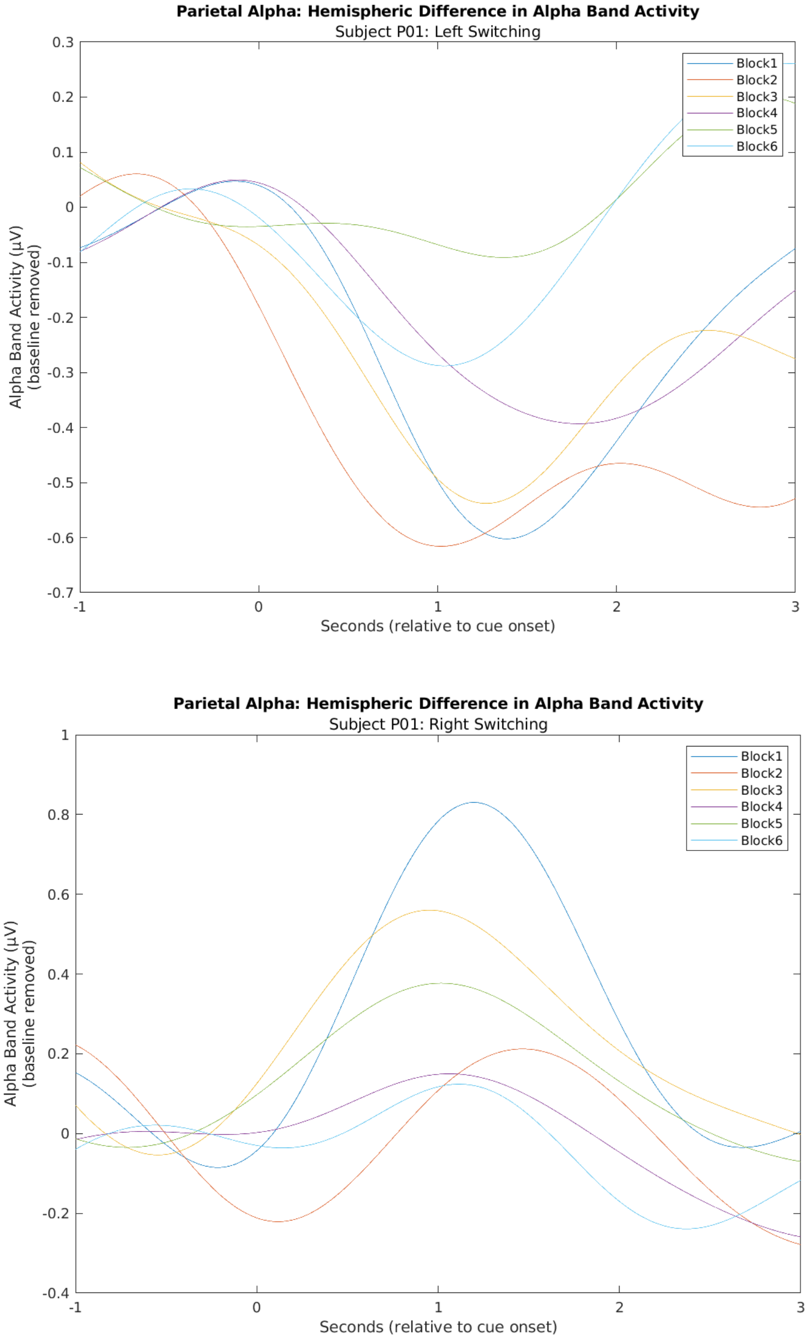


Figure 4.7. For subject P01, plotted is the average difference in parietal alpha band (8-12 Hz) activity (microvolts) between the left and right hemispheres over the time course of the trials. The upper graph represents the activity of the left hemisphere subtracted from the right hemisphere, while the bottom graph represents the activity of the right hemisphere subtracted from the left hemisphere. Time 0 (seconds) represents the onset of the switch cue. Blocks 1 through 6 are represented in different colors as referenced by the key. Alpha band power was averaged across three left hemisphere and three right hemisphere parietal electrodes. Parietal Left: P7, P3, CP5; Parietal Right: P8, P4, CP6

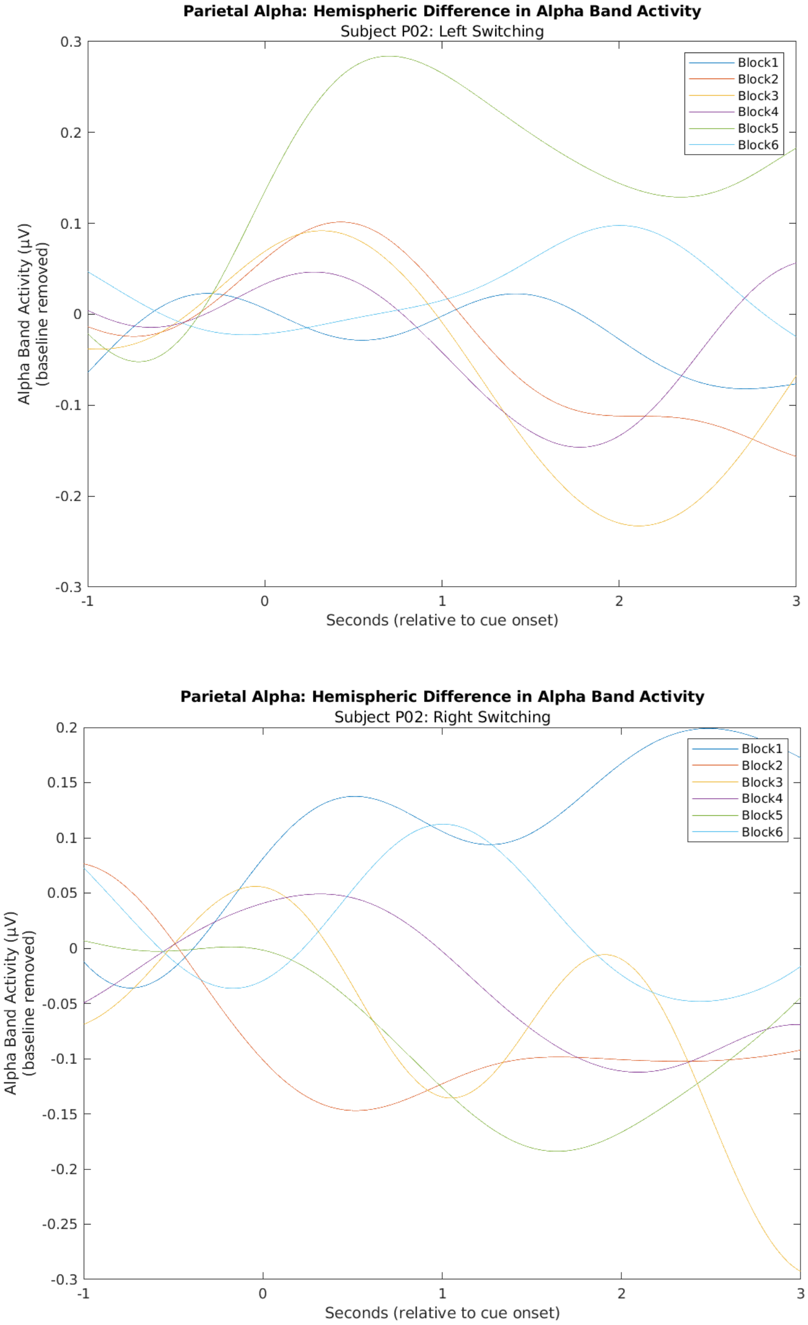


Figure 4.8. For subject P02, plotted is the average difference in parietal alpha band (8-12 Hz) activity (microvolts) between the left and right hemispheres over the time course of the trials. The upper graph represents the activity of the left hemisphere subtracted from the right hemisphere, while the bottom graph represents the activity of the right hemisphere subtracted from the left hemisphere. Time 0 (seconds) represents the onset of the switch cue. Blocks 1 through 6 are represented in different colors as referenced by the key. Alpha band power was averaged across three left hemisphere and three right hemisphere parietal electrodes. Parietal Left: P7, P3, CP5; Parietal Right: P8, P4, CP6

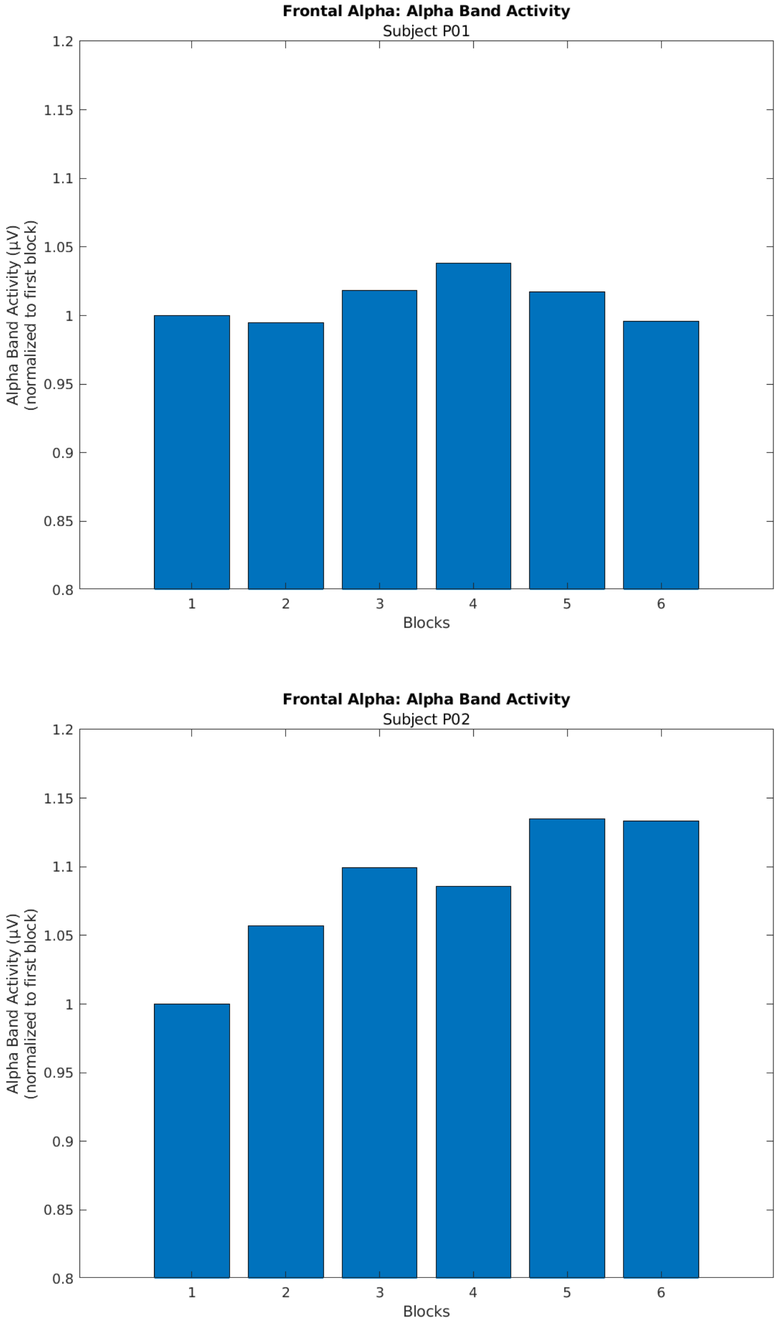
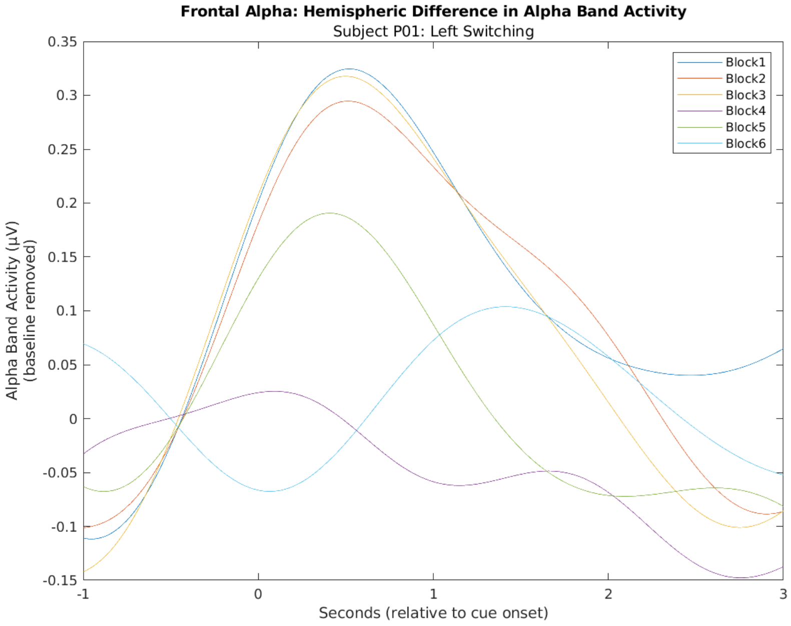


Figure 4.9. Comparison between subjects P01 (top) and P02 (bottom) as to their average frontal alpha band (8-12 Hz) activity (microvolts) over the course of six experimental blocks. Alpha band power was averaged across three left hemisphere and three right hemisphere frontal electrodes. Frontal Left: F7, F3, AF3; Frontal Right: F8, F4, AF4



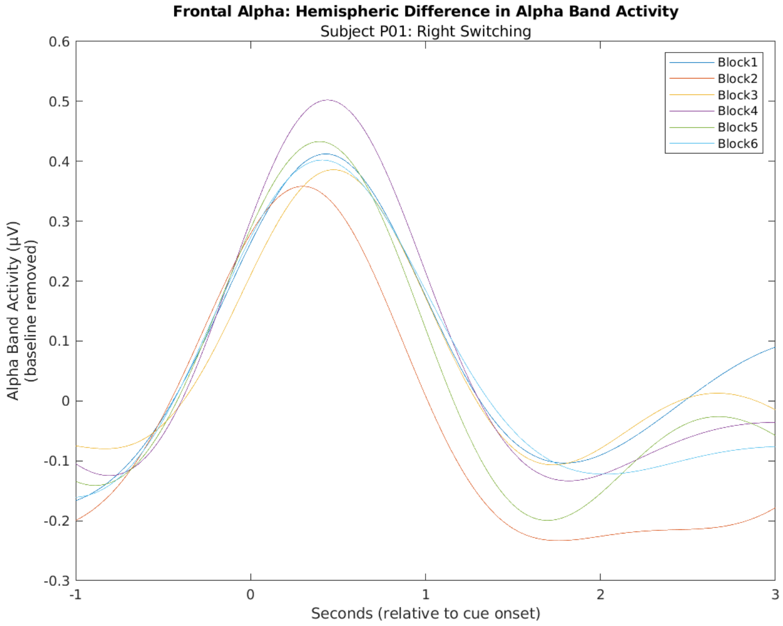
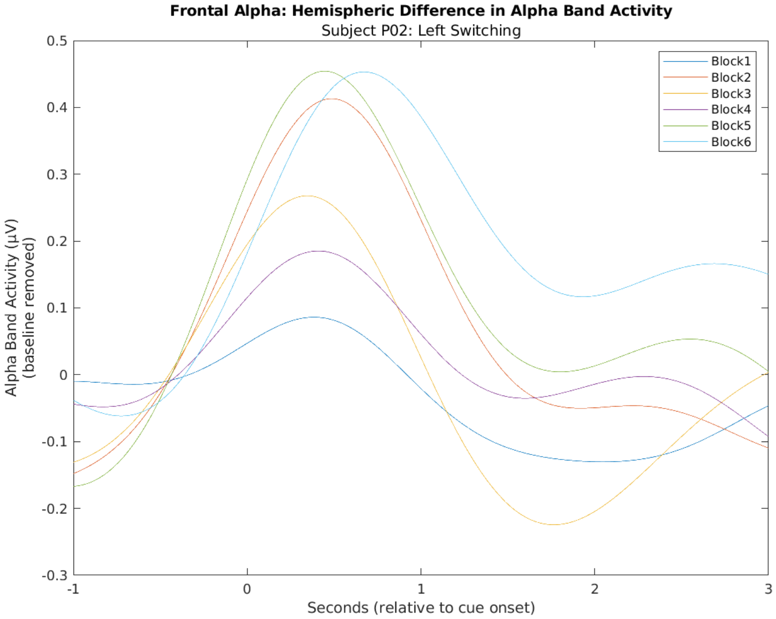


Figure 4.10. For subject P01, plotted is the average difference in frontal alpha band (8-12 Hz) activity (microvolts) between the left and right hemispheres over the time course of the trials. The upper graph represents the activity of the left hemisphere subtracted from the right hemisphere, while the bottom graph represents the activity of the right hemisphere subtracted from the left hemisphere. Time 0 (seconds) represents the onset of the switch cue. Blocks 1 through 6 are represented in different colors as referenced by the key. Alpha band power was averaged across three left hemisphere and three right hemisphere frontal electrodes. Frontal Left: F7, F3, AF3; Frontal Right: F8, F4, AF4



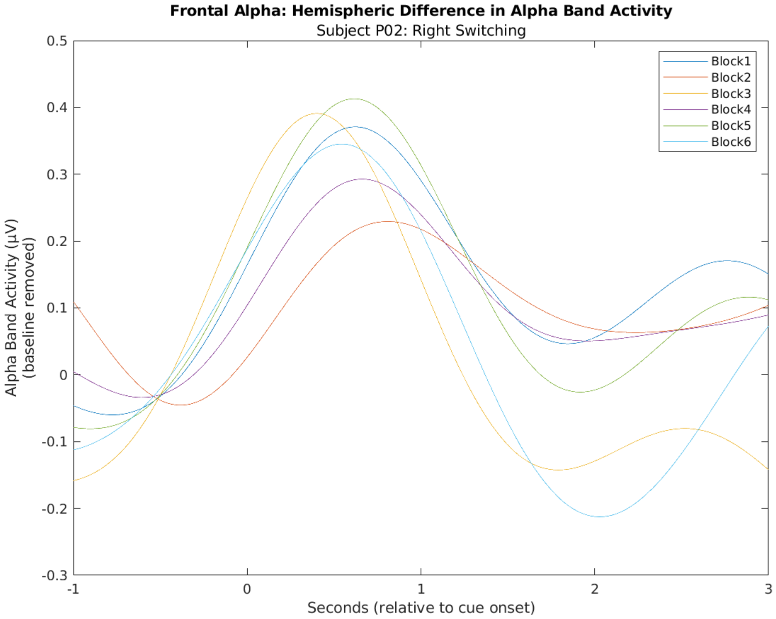


Figure 4.11. For subject P02, plotted is the average difference in frontal alpha band (8-12 Hz) activity (microvolts) between the left and right hemispheres over the time course of the trials. The upper graph represents the activity of the left hemisphere subtracted from the right hemisphere, while the bottom graph represents the activity of the right hemisphere subtracted from the left hemisphere. Time 0 (seconds) represents the onset of the switch cue. Blocks 1 through 6 are represented in different colors as referenced by the key. Alpha band power was averaged across three left hemisphere and three right hemisphere frontal electrodes. Frontal Left: F7, F3, AF3; Frontal Right: F8, F4, AF4

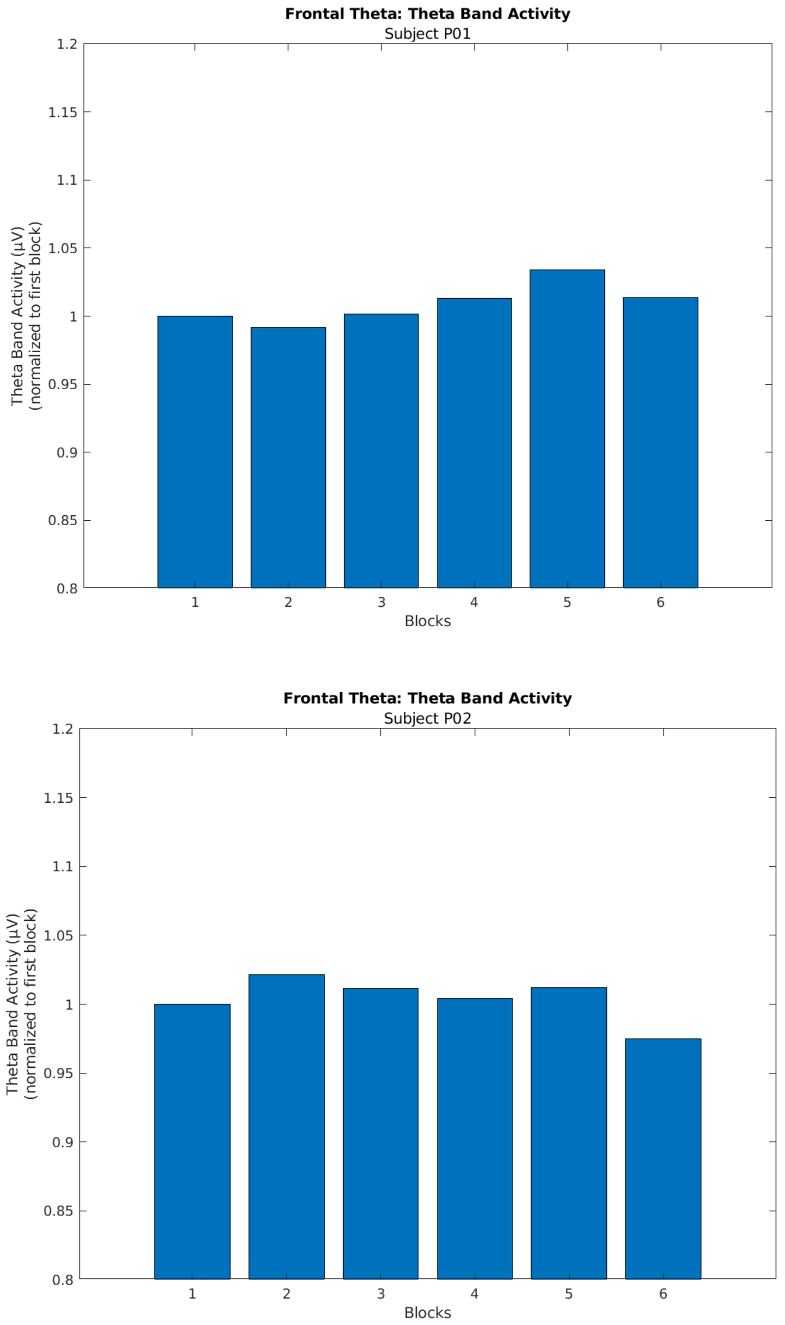


Figure 4.12. Comparison between subjects P01 (top) and P02 (bottom) as to their average frontal alpha theta (4-7 Hz) activity (microvolts) over the course of six experimental blocks. Theta band power was averaged across three left hemisphere and three right hemisphere frontal electrodes. Frontal Left: F7, F3, AF3; Frontal Right: F8, F4, AF4

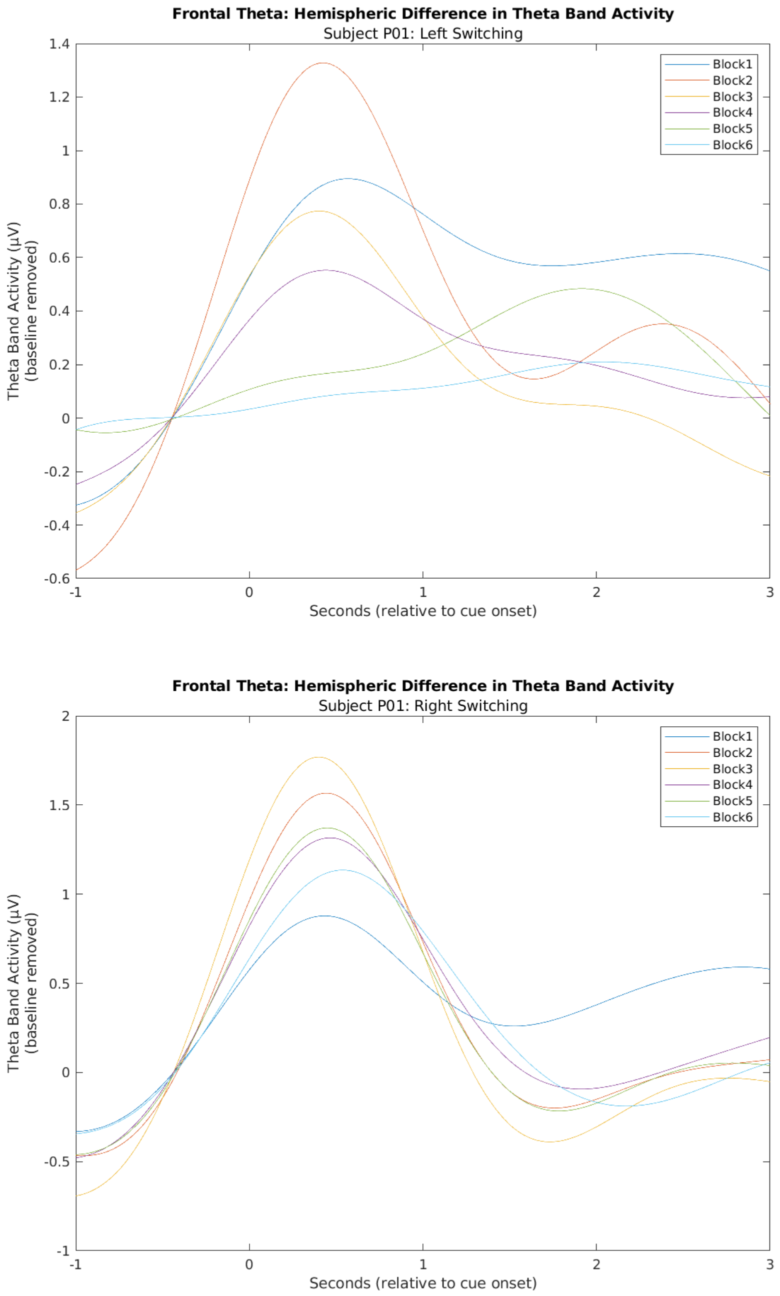
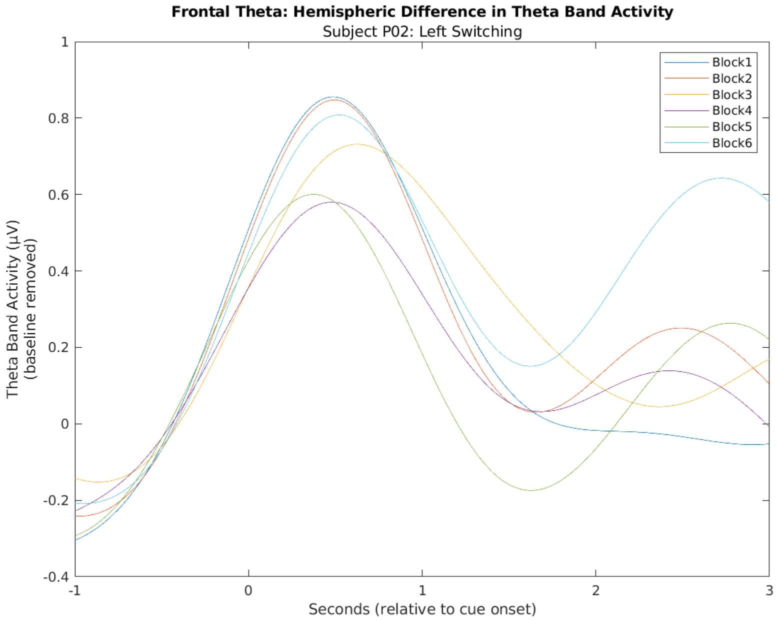


Figure 4.13. For subject P01, plotted is the average difference in frontal theta band (4-7 Hz) activity (microvolts) between the left and right hemispheres over the time course of the trials. The upper graph represents the activity of the left hemisphere subtracted from the right hemisphere, while the bottom graph represents the activity of the right hemisphere subtracted from the left hemisphere. Time 0 (seconds) represents the onset of the switch cue. Blocks 1 through 6 are represented in different colors as referenced by the key. Theta band power was averaged across three left hemisphere and three right hemisphere frontal electrodes. Frontal Left: F7, F3, AF3; Frontal Right: F8, F4, AF4



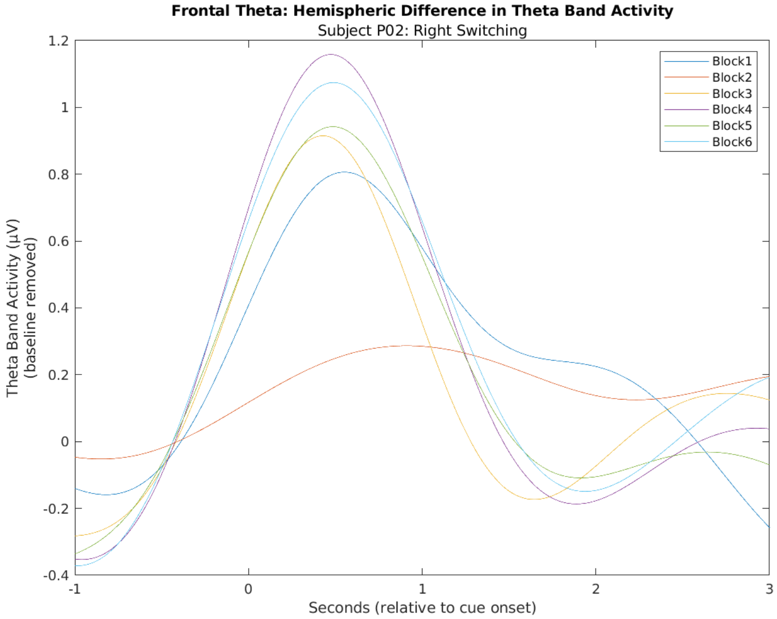


Figure 4.14. For subject P02, plotted is the average difference in frontal theta band (4-7 Hz) activity (microvolts) between the left and right hemispheres over the time course of the trials. The upper graph represents the activity of the left hemisphere subtracted from the right hemisphere, while the bottom graph represents the activity of the right hemisphere subtracted from the left hemisphere. Time 0 (seconds) represents the onset of the switch cue. Blocks 1 through 6 are represented in different colors as referenced by the key. Theta band power was averaged across three left hemisphere and three right hemisphere frontal electrodes. Frontal Left: F7, F3, AF3; Frontal Right: F8, F4, AF4

Chapter 5

“Conclusion”

Our work on Cochlearity was a success, and we showed that mobile platforms are capable of the computation necessary for auditory beamforming and powering attentional prosthetic devices. Both the Delay and Sum (DS) and the Minimum Variance Distortionless Response (MVDR) beamforming algorithms showed great strengths in spatially filtering our audio. While we made great strides over the development of Cochlearity, we were unable to finalize our “dual beamformer” paradigm, which is where we route the high frequency audio to the DS beamformer and the low frequency audio to the MVDR beamformer before combining them back together into a single output. We believe this approach could be particularly potent at noise cancellation, and we encourage further research into this method of combining beamformers together to maximize their effects. Nevertheless, our work successfully pushed the limits of using mobile technology for such scientific and therapeutic applications, and we expect to see more work in this class of research soon.

Alongside our mobile version of Cochlearity, we also made great strides on a second iteration of Cochlearity built using MATLAB on top of a standard x86 PC. This version of Cochlearity allowed us greater flexibility at implementing new algorithms and features and provided much more computational power at the expense of portability. However, we ran into issues due to the single threaded nature of MATLAB processes, and not being able to separate subprocesses into separate processing threads. Towards this end we decided to abandon this new version of Cochlearity, and instead a newer iteration is underway in development using the Python programming language which does not suffer the same limitations as MATLAB. We have great confidence that this iteration will be a success, and along with the developing codebase we are developing our own custom prototype microphone array boards using state-of-the-art MEMS (microelectromechanical systems) microphones.

Likewise, the work we performed on auditory attentional decoding at the Starkey Hearing Research Center has continued to reinforce the idea that the locus of auditory attention can be decoded using EEG using the envelope of the incoming speech. The work being done in the lab of Edmund Lalor, for which this work was based, had showed that tracking the locus of attention using the envelope of speech sound was possible in the healthy hearing population (Di Liberto et al., 2015; O'Sullivan et al., 2015b). Our work extended this knowledge, showing that this same technique could be applied to the hearing-impaired population with success, albeit slightly lower prediction scores.

An application for this research for which we are very interested is to use this attention decoder technique as a control scheme for steerable auditory prostheses. Up until now, the work being pioneered by Kidd et al. has made use of visually guided paradigms, using either eye tracking or electrooculography, which are powerful measures of attention, but are indirect measures of auditory attention that have issues with wearability, reliability, and fail with regards to covert listening. While EEG is in no way a perfect control system, particularly in real-life settings with noisy electromagnetic interference, it does provide us with a direct measure of neural activity, which has been shown to be correlated with the incoming auditory stimulus in such a way as to be predictable and useful (Di Liberto et al., 2015; O'Sullivan et al., 2015b).

Our attention switching paradigm sought to recreate a behaviorally relevant set of circumstances that could induce auditory attentional fatigue over the course of a one-hour study. This novel paradigm chose to use semantically fluid storytelling as well as cue words embedded seamlessly into the stimuli to match the dynamics of a conversational setting involving a story switching between talkers. Our goal was to create as realistic of an experience as possible to isolate the true metrics of auditory attentional fatigue. Towards that end, what we created was a marked success. The paradigm, while simple in its nature, is very fatiguing over the course of the study due to the near constant attentional switching and the need to suppress the distracting talker. That said, we were unable to conduct as robust of a study as we wanted due to events outside of our control, though we did successfully pilot our research paradigm with a single young healthy hearing participant and a single older healthy hearing participant. Or research would seem to indicate that this new paradigm is ready to be used in a myriad of different audio/visual attention experiments, and we hope the work done provides a strong backbone for future studies.

The technology behind Cochlearity, while novel at the time, has since grown more common in the market of auditory protheses. This is a tremendously positive move for consumers and patients with hearing loss. More companies are moving towards 'smart' hearing aids that make use of a combination of beamforming and clever signal processing in order achieve the goal of full suppression of unwanted sounds in otherwise crowded auditory scenes. This work has critical applications for those with hearing impairment, but also for healthy listeners. The last piece of this research that unfortunately was never realized would have been to model the effects of fatigue on both a healthy and a hearing-impaired sample of participants, and then to re-run the same paradigm using Cochlearity as an attentional prosthetic. If we were able to bring the rate of fatigue of our hearing impaired subject closer to their healthy hearing counterparts, that could indicate that attentional prostheses, such as what Cochlearity offers, could ease the cognitive burden of prolonged attentional scenes in those with hearing impairment. Likewise, if an improvement in rate of fatigue was shown for both groups, it would be evidence to support attentional prostheses for use outside of just hearing-impaired groups, but also in the general population.

Lastly, while not yet shown, there exists the possibility for auditory prostheses to solve the "cocktail party" problem in a way that outdoes normal human performance, which could provide wearers with super-human performance in tasks requiring listening in noisy scenes. If this were to be the case, it could have untold use cases in society, as well as police, military, and medical applications. However, this is particularly hopeful thinking on part of the author, and this reality would be many years off. That said, the future is bright for the field of auditory and attentional neuroprostheses, and more people than ever before can now afford access to this truly life changing technology.

References

Acar B, Yurekli MF, Babademez MA, Karabulut H, Karasen RM (2011) Effects of hearing aids on cognitive functions and depressive signs in elderly people. Archives of Gerontology and Geriatrics 52:250–252.

Aiken SJ, Picton TW (2008) Human cortical responses to the speech envelope. Ear and Hearing 29:139–157.

Akram S, Presacco A, Simon J, Shamma S, Babadi B (2016) Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. NeuroImage 124:906–917.

Akram S, Simon J, Shamma S, Babadi B (2014) A State-Space Model for Decoding Auditory Attentional Modulation from MEG in a Competing-Speaker Environment. Advances in Neural Information Processing Systems:460–468.

Anderson MH, Yazel BW, Stickle MPF, Espinosa iñguez FD, Gutierrez N-GS, Slaney M, Joshi SS, Miller LM (2018) Towards mobile gaze-directed beamforming: a novel neuro-technology for hearing loss. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 5806–5809.

Blazer DG, Domnitz S, Liverman CT (2016) Hearing Health Care for Adults. Available at: http://www.nap.edu/catalog/23446.

Boksem MAS, Meijman TF, Lorist MM (2005) Effects of mental fatigue on attention: An ERP study.

Capon J (1969) High-resolution frequency-wavenumber spectrum analysis. Proceedings of the IEEE 57:1408–1418.

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

Demberg V, Sayeed A (2016) The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. PLoS ONE 11:e0146194.

Denk F, Ewert SD, Kollmeier B (2018) Spectral directional cues captured by hearing device microphones in individual human ears. The Journal of the Acoustical Society of America 144:2072–2087.

Denk F, Ewert SD, Kollmeier B (2019) On the limitations of sound localization with hearing devices. The Journal of the Acoustical Society of America 146:1732–1744.

Di Liberto GM, O’Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology 25:2457–2465.

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences 109:11854–11859.

Faber LG, Maurits NM, Lorist MM (2012) Mental Fatigue Affects Visual Selective Attention. PLoS ONE 7:e48073.

Favre-Felix A, Graversen C, Dau T, Lunner T (2017) Real-time estimation of eye gaze by in-ear electrodes. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp 4086–4089. IEEE. Available at: http://ieeexplore.ieee.org/document/8037754/.

Gergelyfi M, Jacob B, Olivier E, Zénon A (2015) Dissociation between mental fatigue and motivational state during prolonged mental activity. Frontiers in Behavioral Neuroscience 9:176.

Getzmann S, Golob EJ, Wascher E (2016) Focused and divided attention in a simulated cocktail-party situation: ERP evidence from younger and older adults. Neurobiology of Aging 41:138–149.

Getzmann S, Hanenberg C, Lewald J, Falkenstein M, Wascher E (2015) Effects of age on electrophysiological correlates of speech processing in a dynamic &quot;cocktail-party&quot; situation. Frontiers in neuroscience 9:341.

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

Habets EAP, Benesty J, Cohen I, Gannot S, Dmochowski J (2010) New Insights Into the MVDR Beamformer in Room Acoustics. IEEE Transactions on Audio, Speech, and Language Processing 18:158–170.

Hart J, Onceanu D, Sohn C, Wightman D, Vertegaal R (2009) The attentive hearing aid: Eye selection of auditory sources for hearing impaired users. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 19–35. Springer, Berlin, Heidelberg. Available at: http://link.springer.com/10.1007/978-3-642-03655-2_4.

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional Gain Control of Ongoing Cortical Speech Representations in a “Cocktail Party.” Journal of Neuroscience 30:620–628.

Kidd G (2017) Enhancing Auditory Selective Attention Using a Visually Guided Hearing Aid. Journal of Speech Language and Hearing Research 60:3027.

Kidd G, Favrot S, Desloge JG, Streeter TM, Mason CR (2013) Design and preliminary testing of a visually guided hearing aid. The Journal of the Acoustical Society of America 133:EL202–EL207.

Kidd G, Mason CR, Best V, Swaminathan J (2015) Benefits of acoustic beamforming for solving the cocktail party problem. Trends in Hearing 19:233121651559338.

Larson E, Lee AKC (2013) The cortical dynamics underlying effective switching of auditory spatial attention. NeuroImage 64:365–370.

Levitt H (2018) Digital Hearing Aids: Wheelbarrows to Ear Inserts. The ASHA Leader Available at: https://leader.pubs.asha.org/doi/10.1044/leader.FTR4.12172007.28 [Accessed November 15, 2021].

Li CM, Zhao G, Hoffman HJ, Town M, Themann CL (2018) Hearing Disability Prevalence and Risk Factors in Two Recent National Surveys. American Journal of Preventive Medicine 55:326–335.

Marrone N, Mason CR, Kidd G (2008) Evaluating the Benefit of Hearing Aids in Solving the Cocktail Party Problem. Trends in Amplification 12:300–315.

Marzetta T (2008) Self-steering directional hearing aid and method of operation thereof. Available at: https://patents.google.com/patent/US20100074460A1/en.

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

Mills M (2011) Hearing Aids and the History of Electronics Miniaturization. IEEE Annals of the History of Computing 33:24–45.

O’Sullivan J, Power A, Mesgarani N, Rajaram S, Foxe J, Shinn-Cunningham B, Slaney M, Shamma S, Lalor E (2015a) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cerebral cortex (New York, NY : 1991) 25:1697–1706.

O’Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015b) Attentional Selection in a Cocktail

Party Environment Can Be Decoded from Single-Trial EEG. Cerebral Cortex 25:1697–1706.

Piquado T, Isaacowitz D, Wingfield A (2010) Pupillometry as a measure of cognitive effort in younger and older adults. Psychophysiology 47:560–569.

Rudner M, Ellis RJ, Koelewijn T, Lin G, Carlile S (2015) Costs of switching auditory spatial attention in following conversational turn-taking. Frontiers in Neuroscience | www.frontiersin.org 9:124.

Shinn-Cunningham BG, Best V (2008a) Selective Attention in Normal and Impaired Hearing. Trends in Amplification 12:283–299.

Shinn-Cunningham BG, Best V (2008b) Selective Attention in Normal and Impaired Hearing. Trends in Amplification 12:283–299.

Tollin DJ, Yin TCT (2009) Sound Localization: Neural Mechanisms. In: Encyclopedia of Neuroscience (Squire LR, ed), pp 137–144. Oxford: Academic Press. Available at: https://www.sciencedirect.com/science/article/pii/B9780080450469002679 [Accessed November 16, 2021].

Trejo LJ, Kubitz K, Rosipal R, Kochavi RL, Montgomery LD (2015) EEG-Based Estimation and Classification of Mental Fatigue. Psychology 06:572–589.

van Wassenhove V, Grant KW, Poeppel D (2007) Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45:598–607.

Vu NV, Ye H, Whittington J, Devlin J, Mason M (2010) Small footprint implementation of dual-microphone delay-and-sum beamforming for in-car speech enhancement. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp 1482–1485. IEEE. Available at: http://ieeexplore.ieee.org/document/5495493/.

Worden MS, Foxe JJ, Wang N, Simpson G V (2000) Anticipatory Biasing of Visuospatial Attention Indexed by Retinotopically Specific-Band Electroencephalography Increases over Occipital Cortex.

Yang B, Xiao W, Liu X, Wu S, Miao D (2013) Mental fatigue impairs pre-attentive processing: A MMN study. Neuroscience Letters 532:12–16.

Zekveld AA, Heslenfeld DJ, Johnsrude IS, Versfeld NJ, Kramer SE (2014) The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. NeuroImage 101:76–86.

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