ABSTRACT

Title of Dissertation: DECODING THE BRAIN IN COMPLEX AUDITORY ENVIRONMENTS

Mohsen Rezaeizadeh
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Shihab Shamma
Department of Electrical and Computer Engineering

Humans have an exceptional ability to engage with sequences of sounds and extract meaningful information from them. We can appreciate music or absorb speech during a conversation, not like anything else on the planet. It is unclear exactly how the brain effortlessly processes these rapidly changing complex soundscapes. This dissertation explored the neural mechanisms underlying these remarkable traits in an effort to expand our knowledge of human cognition with numerous clinical and engineering applications.

Brain-imaging techniques have provided a powerful tool to access mental representations' content and dynamics. Non-invasive imaging such as Electroencephalography (EEG) and Magnetoencephalography (MEG) provides a fine-grained dissection of the sequence of brain activities. The analysis of these time-resolved signals can be enhanced with temporal decoding methods that offer vast and untapped potential for determining how mental representations unfold over time. In the present thesis, we use these decoding techniques, along with a series of novel

experimental paradigms, on EEG and MEG signals to investigate the neural mechanisms of auditory processing in the human brain, ranging from neural representation of acoustic features to the higher level of cognition, such as music perception and speech imagery.

First, we reported our findings regarding the role of *temporal coherence* in auditory source segregation. We showed that the perception of a target sound source can only be segregated from a complex acoustic background if the acoustic features (e.g., pitch, location, and timbre) induce temporally modulated neural responses that are mutually correlated. We used EEG signals to measure the neural responses to the individual acoustic feature in complex sound mixtures. We decoded the effect of attention on these responses. We showed that attention and the coherent temporal modulation of the acoustic features of the target sound are the key factors that induce the binding of the target features and its emergence as the foreground sound source.

Next, we explored how the brain learns the statistical structures of sound sequences in different musical contexts. The ability to detect probabilistic patterns is central to many aspects of human cognition, ranging from auditory perception to the enjoyment of music. We used artificially generated melodies derived from uniform or non-uniform musical scales. We collected EEG signals and decoded the neural responses to the tones in a melody with different transition probabilities. We observed that the listener's brain only learned the melodies' statistical structures when derived from non-uniform scales.

Finally, we investigated brain processing during speech and music imagery with Brain-Computer Interface applications. We developed an encoder-decoder neural network architecture to find a transformation between neural responses to the listened and imagined sounds. Using this map, we could reconstruct the imagery signals reliably, which could be used as a template to decode the actual imagery neural signals. This was possible even when we generalized the

model to unseen data of an unseen subject. We decoded these predicted signals and identified the imagined segment with remarkable accuracy.

DECODING THE BRAIN IN COMPLEX AUDITORY ENVIRONMENTS

by

Mohsen Rezaeizadeh

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Shihab Shamma, Chair
Professor Jonathan Z. Simon,
Associate Professor Behtash Babadi
Associate Professor Samira Anderson
Associate Professor Robert Slevc

Acknowledgments

It is impossible to thank every individual that made this dissertation possible. I thank you for everything you gave me across these years of grad school that made my life richer.

I cannot sum up or quantify what I have gained from working with Shihab over the years in my Ph.D. life. I do not think anybody else could push me this much to explore several scientific interests. Shihab was always a source of inspiration, and he made me fall in love with the brain. His authenticity and his energy level were contagious, and our meetings and discussions always put a smile on my face for the day. Shihab was always available, met whenever I asked, gave me space whenever I needed it, and steadily supported me, especially when I hit the tough spots. Even when I was doing something he did not think was worthwhile – of course, I was wrong most of the time. I believe there are not many advisors that just let you go on and make your own mistakes that way and at the same time get the best out of you. I feel lucky I was his student, and I do not know exactly why he kept believing in me, but I will never forget it, and I am forever grateful to him for that.

I learned a ton about everything from Jonathan. His attention to detail made many things easier for me. Jonathan offered support whenever I asked – as he did for anyone who asked for it. He always read everything I sent carefully; his patience, comments, and advice were priceless. For me, the more I know him, the fewer practical mistakes I end up making. I am grateful for the time and resources he put into helping me.

Thank you to Behtash for his friendly support, brilliant mathematical insights, and great ideas. I appreciate his mathematical approach to neuroscience, and I learned a great deal about applying them to neural data. On top of that, he was the best foosball team-mate in the world. Thank you for adding a lot of fun to my years of Ph.D.

I feel privileged because of working in a collaborative and high-performing culture these three wonderful guys made.

Thank you to Samira and Bob, who have kindly agreed to be part of my research committee. Thank you for your time and invaluable advice. Thank you to Andrew Oxenham for hosting me in his lab at the University of Minnesota and providing me with all the necessary resources. I am also grateful to Alain de Cheveigné, Mounya Elhilali, Malcolm Slaney, and Nima Mesgarani for all the fruitful conversations I had with them at workshops and conferences.

I am glad that I started my research siding with Claire Pelofi. She supported me, trusted me, and gave me the confidence to explore what I was interested in. She helped me a lot to have a good start in graduate school. She was also a great friend, who I enjoyed the company and a source of support when I was pretty unsure of myself.

Thank you to all the former and current folks at NSL and CSSL, to Jonathan for helping me set up the EEG lab, Neha and Kelsey for making the challenges much more manageable and for all the fun talks, Joshua for all the nerdy fantasy discussions, Shoutik for all nerdy scientific discussions, Ali, Dani, Diego, Kai, Daniel, Jaya, Pngbo, Christian, Rupesh, Lakshmi, and Sahar for all the time we spent during the coffee breaks and beyond. I thank all of you for helping me in so many ways and for being a friends.

Thank you to Ciaran, Ed, and Richard for helping me set up MEG experiments. To Ed for all his effort in making my experiment possible. To Ciaran for being a mesmerizing voice actor

in one of my stimuli. To Richard because of all his assistance during my last few projects.

I am lucky to have a partner who incredibly supported my career. Sonia always encouraged me to move forward, especially when I doubted what I was doing. She was always a source of motivation, made it possible for me to keep doing this, never give in, and when I was out of breath, she helped me to see the light at the end of the tunnel. I do not know if I would have made it through the Ph.D. if I had not had Sonia's constant support.

Thank you to all my wonderful friends for their support. To Fariba, Mahdi, Shahriar, Omid, Simon, Ehsan, Shima, Guilhem, Giovanni, Abhinav, Saber, Paniz, Ali ($\times 4$), Alireza, Maya, Sina, Matin, Sahar, Behrad, Pardis, Anousheh, Nariman, and those inadvertently missed. Thank you for your friendships and for making my life more colorful.

Lastly, I am just so lucky to have the family I have, thank you for believing in every dream I have had, and giving me the most significant reason to succeed.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A1	Primary Auditory Cortex
AAD	Auditory Attention Decoding
ASA	Auditory Scene Analysis
AUC	Area Under the [ROC] Curve
ANOVA	Analysis of Variance
BCI	Brain Computer Interface
BPM	Beats per Minutes
CHT	Concurrent Hierarchical Tracking
CNN	Convolutional Neural Network
CNS	Central Nervous System
DFT	Discrete Fourier Transform
DSS	Denoised Source Separation
ECoG	Electrocorticography
EEG	Electroencephalography
EOG	Electrooculography
ERP	Event-Related Potential
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
fMRI	Functional Magnetic Resonance Imaging
HG	Heschl's Gyrus
HEOG	Horizontal Electrooculogram
IC	Inferior Colliculus
ICA	Independent Component Analysis
IPC	Intraparietal Sulcus
IHC	Inner Hair Cells
JD	Joint Decorrelation

MEG	Magnetoencephalography
MGN	Medial Geniculate Nuclues
NN	Neural Network
OCT	Octave
PCA	Principal Component Analysis
PFC	Prefrontal Cortex
pSTG	Posterior Superior Temporal Gyrus
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SEM	Standard Error of the Mean
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
SQUID	Superconducting Quantum Interference Device
STG	Superior Temporal Gyrus
TC	Temporal Coherence
TET	Tone Equal Temperament
TFCE	Threshold Free Cluster Enhancement
TRF	Temporal Response Function
TSPCA	Time-Shift Principal Component Analysis
VEOG	Vertical Electrooculogram

Chapter 1: **Introduction**

1.1 A Remarkable Computation Tool

To navigate the world and perform our daily tasks, we depend on sensory inputs like vision, audition, somatic sensations, olfaction, and taste. Each sensory modality in humans and other animals has evolved to provide information derived from a specific form of energy. For example, the auditory system processes activity generated by local pressure fluctuations in the auditory environment. Our brain – an exceptional computation processor – transforms a complex mixture of acoustic input, and magically we can perceive segregated sound streams, localize different sound sources, attend selectively to a voice, extract semantic information from speech, and predict what can come next in a stream of sounds like speech or music. Although navigating the auditory world seems to be an effortless and undemanding task, all these steps are challenging problems that are disentangled and resolved through the auditory pathway from the ear to regions for higher levels of cognition in the brain.

In a cluttered environment – also known as the real world – the soundscape consists of several auditory objects, each emanates distinct or overlapping acoustical features, and all these features mix and compile together as a one-dimensional waveform by the time it reaches the outer ear. This complex procedure invites a lot of ambiguous situations where the information from the environment might be limited and insufficient. To overcome this and continuously

generate perceptions, the brain requires top-down modulations as much as it depends on the sensory input the peripheral nervous system provides. Therefore, the brain does not merely depend on continuous input from the external world to generate perceptions but only to modulate them contextually [1]. In other words, what we perceive is very much affected by attention, memory, emotional states, the inputs from other sensory modalities, and many other factors. Therefore to investigate how does brain help us to appreciate sitting in a concert hall and enjoying a piano concerto, we should study top-down and bottom-up information processing of the nervous system and how the brain puts all this information together.

1.2 Thesis outline

The advent of brain-imaging techniques has provided a powerful tool to access mental representations' content and dynamics. Non-invasive imaging such as Electroencephalography (EEG) and Magnetoencephalography (MEG) provides a fine-grained dissection of the sequence of brain activities. The analysis of these time-resolved signals can be enhanced with temporal decoding methods that offer vast and untapped potential for determining how mental representations unfold over time. In the present thesis, we use these decoding techniques, along with a series of novel experimental designs and paradigms, on EEG and MEG signals to investigate the neural mechanisms of auditory processing in the human brain, ranging from neural representation of acoustic features to the higher level of cognition, such as music perception and speech imagery.

This thesis is organized into six chapters. In the following chapter, we summarized the current evidence and the background regarding the brain's auditory processing. To have a funda-

mental view of different processing steps taken by the brain, we briefly touched on the auditory pathway from the peripheral to the central nervous system (CNS). We then introduced EEG and MEG techniques, how they work, and their differences. We discuss how they help address questions regarding cognitive neuroscience, particularly related to the auditory system.

In chapter 3, we discuss our findings regarding the role of *temporal coherence* in auditory source segregation. Numerous studies have suggested that the perception of a target sound source can only be segregated from a complex acoustic background if the acoustic features underlying its perceptual attributes (e.g., pitch, location, and timbre) induce temporally modulated responses that are mutually correlated, and that are uncorrelated from those of other sources in the mixture. This "temporal coherence" hypothesis asserts that listening attentively to one or a subset of attributes of a target source enhances their neural responses and concomitantly enhances all other coherent responses, thus binding them together while simultaneously suppressing the incoherent responses to the background features. Chapter 3 reports on EEG measurements in human subjects engaged in various sound segregation tasks that demonstrate binding among the temporally coherent features of the attended source regardless of their identity, harmonic relationship, or frequency separation.

Chapter 4 explored how the brain learns the statistical structures of sound sequences, such as music, in different contexts. The ability to detect patterns (or learn statistical structure) in the environment is central to many aspects of human cognition, ranging from perception to the enjoyment of music. Music offers an excellent opportunity to investigate statistical learning in an ecologically-valid setting, as music is an essential part of every culture, and evidence suggests that implicit learning underlies music acquisition. We used artificially generated melodies de-

rived from uniform or non-uniform musical scales[1] and collected EEG signals from participants exposed to those melodies. By decoding the neural responses to the tones in a melody with different transition probabilities, we observed that the listener's brain only learned the melodies' statistical structures when derived from asymmetric scales. In addition, this result suggests cognitive benefits associated with asymmetry in scales, which is a recurrent feature across cultures despite their rich diversity.

In chapter 5, we investigated the brain processing during speech and music imagery and attempted to *read the mind* during imagery tasks. Auditory imagery is voluntarily hearing sounds in our mind without external stimulation. Because of this lack of sensory input, EEG and MEG neural signals are much weaker during imagination. We used EEG data collected from professional musicians. We developed an encoder-decoder neural network architecture to find a transformation between EEG responses to the listened and imagined music. Using this map, we could reconstruct the imagery signals reliably, which could be used as a template to decode the actual weaker imagery neural signals. We observed that this is possible even when we are generalizing the model to unseen data of an unseen subject. We decoded these predicted signals and identified the imagined musical piece with remarkable accuracy. Furthermore, in a MEG experiment, we compared the speech imagery with music and showed that we could transfer the mapping train on music (or speech) stimuli to the speech (or music) stimuli.

Finally, In the last chapter, we provide an overview of the present thesis and a summary of its main findings. We then shall discuss future efforts to extend our understanding of the auditory brain further.

[1] Asymmetric scale is a robust universal feature that corresponds to the fact that the discrete pitches used in melodies are separated by intervals of different sizes. Conversely, a uniform scale consists of pitches with the same interval size. For more details, see chapter 4.

Chapter 2: **Background**

2.1 Auditory Pathway

Before studying higher-level processing in the auditory system, it is helpful to appreciate the early stage and periphery auditory system. The auditory periphery begins at the point where the acoustic wave meets the outer-most part of the ear and ends at the auditory nerves (Figure 2.1). The sound that enters the ears is transformed beyond recognition before it reaches the end of its journey.

The periphery consists of three areas: The outer, middle, and inner ears. The outer ear can be divided into the pinna (asymmetric weird shape external ear), the auditory canal (meatus), and the eardrum (tympanic membrane). The asymmetric shape of the pinna helps in spatial hearing (sound localization). Sound traveling through the auditory canal cause vibrations at the eardrum. These vibrations are transmitted and amplified to the cochlear fluid by tiny bones (ossicles) in the middle ear [2–4].

The cochlear in the inner ear consists of a fluid-filled tube divided along its length by *Reissner's membrane* and *basilar membrane*. The basilar membrane vibrates in response to the pressure changes caused by the *pushing* and *pulling* of the ossicles on the oval window. The physical structure of the cochlea is organized in terms of the frequency of the tone(s) being processed, and there is an ordered frequency axis along the length of the cochlea – in fact, this

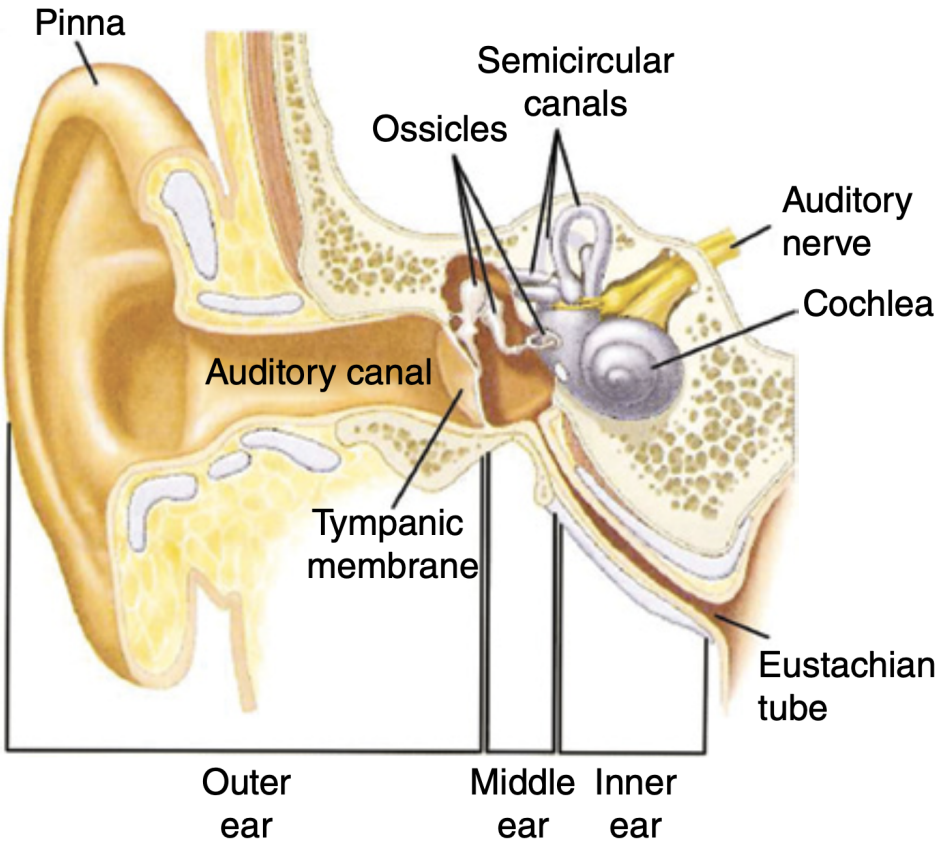

Figure 2.1: *Schematic of humans periphery auditory system.* The pressure wave enters the auditory canal, causes vibration at the eardrum and then passed on to the cochlea. Finally, cochlea transform the mechanical vibrations into action potentials that are generated in the tens of thousands of auditory nerve fibers that carry the auditory message to the brain stem. Adapted from [6].

relation between location and particular frequency is preserved all the way to the cortex, and it is called tonotopy – where it captures higher frequencies at the front and lower frequencies at the end of the membrane, and it decomposes the sound into frequency components like a Fourier transform; therefore, the basilar membrane is a mechanical analyzer of sound frequency. Translation of the mechanical energy of the basilar membrane into a neural signal occurs in inner hair cells (IHC) that are attached along the basilar membrane [5].

Auditory nerve activity at cochlear nuclei is provoked by the release of transmitter substance (glutamate) into the synaptic connections between the auditory nerve dendrites (ganglion

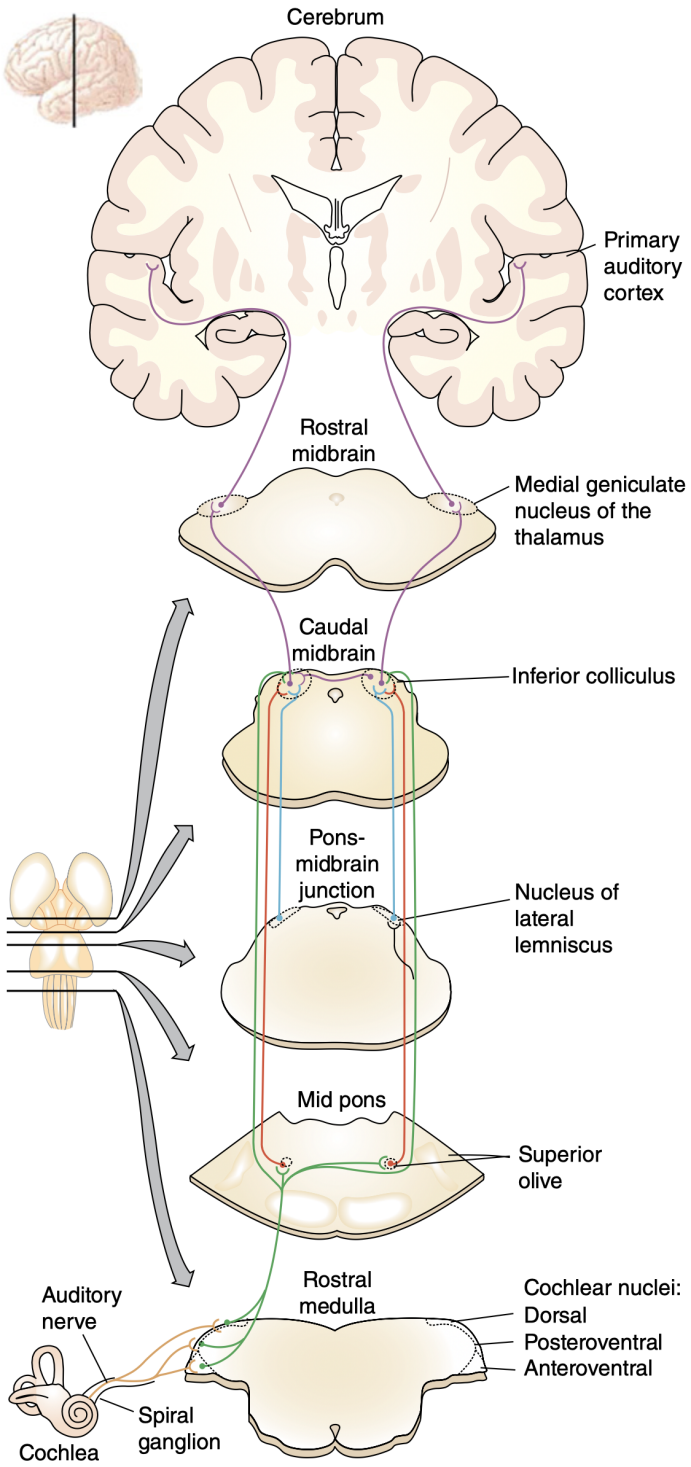

Figure 2.2: *Schematic of humans auditory system.* Auditory pathway from cochlea to primary auditory cortex (Heschl's gyrus). Adapted from [4]

cells) and the inner hair cells. Therefore two important properties of sound are already encoded in auditory nerves: frequency components of sounds and volume. The frequency of sound is encoded due to the tonotopical structure of the basal membrane, and consequently, the auditory nerve responds preferentially to certain frequencies. In addition, the volume of a sound is encoded by the rate at which neurons of the cochlear nucleus fire. i.e., Increasing the sounds' volume results in greater oscillations of the eardrum and, consequently, in greater deflection of the basilar membrane, larger shearing forces on the inner hair cells, and the release of more transmitters onto cochlear ganglion cells. [7]

Beyond the auditory periphery, the pathway leading to the primary auditory cortex (A1) passes through four neural structures: cochlear nuclei, superior olive, inferior colliculus (IC), and medial geniculate nucleus (MGN) (See Figure 2.2). Neurons throughout these nodes extract and encode several sound features, such as periodicity, frequency modulation, sound intensity, volume, interaural time, and intensity differences. Importantly, as we mentioned above, the tonotopic frequency established by the basilar membrane's properties is preserved at each of these nodes. Moreover, It is in these brainstem nuclei that many of the computations are performed that allow us, among other things, to localize where a sound in the environment is coming from.

The activity of the MGN projects onto A1, where A1 is located in the posterior superior temporal gyrus (pSTG), bilaterally, on the ventral wall of the Sylvian fissure. Along this projection, the tonotopy organization that originates in the cochlea is preserved. i.e., neurons are progressively tuned to higher (or lower) frequencies on the surface of the primary auditory cortex as we move along a particular direction. The areas of the auditory cortex in STG are interconnected, where neurons receive modest input from the MGN neurons and about 10-100 times more input from other brain regions [8]. This indicates the importance of neurofeedback and the

top-down modulations in sensory processing and perception [9, 10].

To summarize, the auditory pathway from the periphery to the cortex decomposes a sound wave into several acoustical features (pitch, frequency, location, volume, etc.) and encodes these attributes separately. The brain has to bind these features to form an auditory object and help us to perceive a single sound stream. We discuss these in more detail in chapter 3.

The following section (section 2.2) discusses the principles of Electroencephalography (EEG) and its use in studying the human auditory system.

2.2 Electroencephalography (EEG)

In 1929, Hans Berger reported a remarkable discovery in which he showed that the electrical activity of the human brain could be measured by placing an electrode on the scalp and plotting the changes in voltage over time [11]. This electrical activity is called the electroencephalogram, or EEG for short. Although these reported findings were controversial and neurophysiologists of the day were skeptical about the nature of the oscillations observed in EEG, over the years, EEG was proven as a distinguished way to study the human brain and its cognitive process non-invasively.

For state-of-the-art research, the EEG methods make use of several surface electrodes (up to 256). The electrodes placed on the scalp require good conductive contact with the skin to detect the variation in electrical fields caused by brain activities. The EEG signals derive from the aggregation of thousands or even millions of neurons (their dendritic field potentials) that vary together, as illustrated in Figure 2.3. The brain sustains ionic current flows, where neurons communicate with each other via action potentials (a current generated by a neuron). When there

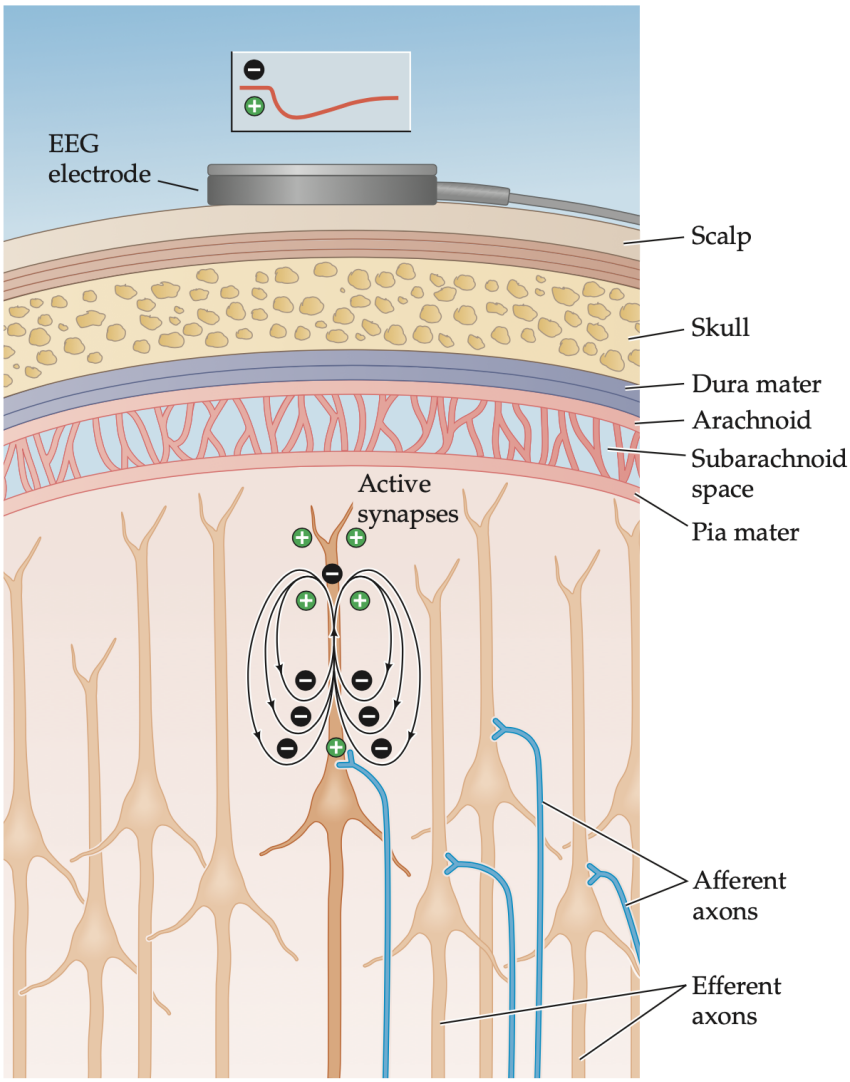

Figure 2.3: *Schematic of how EEG works.* Auditory pathway from cochlea to primary auditory cortex Adapted from [12]

is excitatory synaptic input to the dendrites near the cell body (see Figure 2.3), the voltage across the cell membrane is depolarized by the synaptic activity. The amount of current generated by a single cell is too small to be detected in a non-invasive measurement from outside of the head; however, if the currents are derived with a sufficient amount of synchrony, it gives rise to a sizeable net current that is detectable centimeters away, outside the head. The scalp electrodes pick up the fluctuating voltages associated with these dynamically changing return currents. This

interpretation of the basis of EEG is well supported by other data, such as direct intracranial recordings in experimental animals [13].

Although the continuous EEG signal helps assess the overall state of the brain since it captures the signals from the entire brain, its utility in investigating specific cognitive functions is relatively limited. The reason is that the ongoing EEG record is not linked in time to any particular cognitive process or event. Therefore, a more effective way of relating scalp electrical activity to cognitive function is to implement a time-locked approach. The most common signals to extract from the ongoing EEG in this way are event-related potentials (ERPs). ERPs are small voltage fluctuations in a continuous EEG time-locked with a sensory response or cognitive events; they reflect the summed electrical activity of neuronal populations specifically responding to those events. Consequently, they can provide high temporal resolution (milliseconds; see Table 2.1) of the neural processing underlying various cognitive functions. However, because ERPs are generally smaller than the raw EEG signal, and since they are embedded in noise – the source of noise can be outside of the brain as well as unwanted neural activities from different brain regions –, it is necessary to average multiple trials, time-locked to repeated occurrences of a specific sensory, to extract ERP signals from the background noise. The ongoing EEG varies more or less randomly in amplitude relative to the timing of these events; these random fluctuations in the EEG cancel out in averaging, leaving only those voltage changes specifically associated with the processing of the event type of interest.

The average ERP waveform obtained in this way generally sustains a series of negative and positive peaks that are typically named according to their electrical polarity and latency. e.g., N100 in an auditory task is a negative peak 100 milliseconds after the tone onset (researchers sometimes show this pick as N1). These sensory evoked responses are effectively identical in all

Table 2.1: *Overview of methods in cognitive neuroscience, comparing their spatial and temporal resolutions.*

Method	Spatial Resolution (meter)	Temporal Resolution (second)
Single-unit Recording	10^{-5} (single neuron)	10^{-3}
EEG	10^{-1} (whole brain)	10^{-3}
MEG	10^{-2}	10^{-3}
fMRI	10^{-3}	10^{0}

neurologically healthy individuals. This makes them very effective tools for studying the neural bases of many different behavioral and physiological states. For instance, to study attention, how does the auditory evoked response differ in trials when one is attending to a concurrent stream of visual stimuli vs. trials when one is attending to the auditory stimuli?

We reviewed the auditory pathway in the previous section (Section 2.1). We considered the processes of the transduction of acoustic energy into a neural code and the stages of processing the auditory signal from the brainstem to MGN to the cortex, where different sound features are encoded. The synaptic activity associated with these many nodes along the pathway produces the many peaks in the early-latency components of the auditory ERPs that can be captured using EEG signals and ERPs. Figure 2.4 illustrates the components of the auditory ERP; each reflects a different processing stage, ranging from the early picks associated with the brain stem(as soon as one millisecond) to the latest peak at 300-400 milliseconds. The N1 and subsequent components are the first that correspond to cortical processing. Thus, despite all the circuitry of the auditory brainstem, auditory information gets from ear to cortex in less than 100 milliseconds! [14]

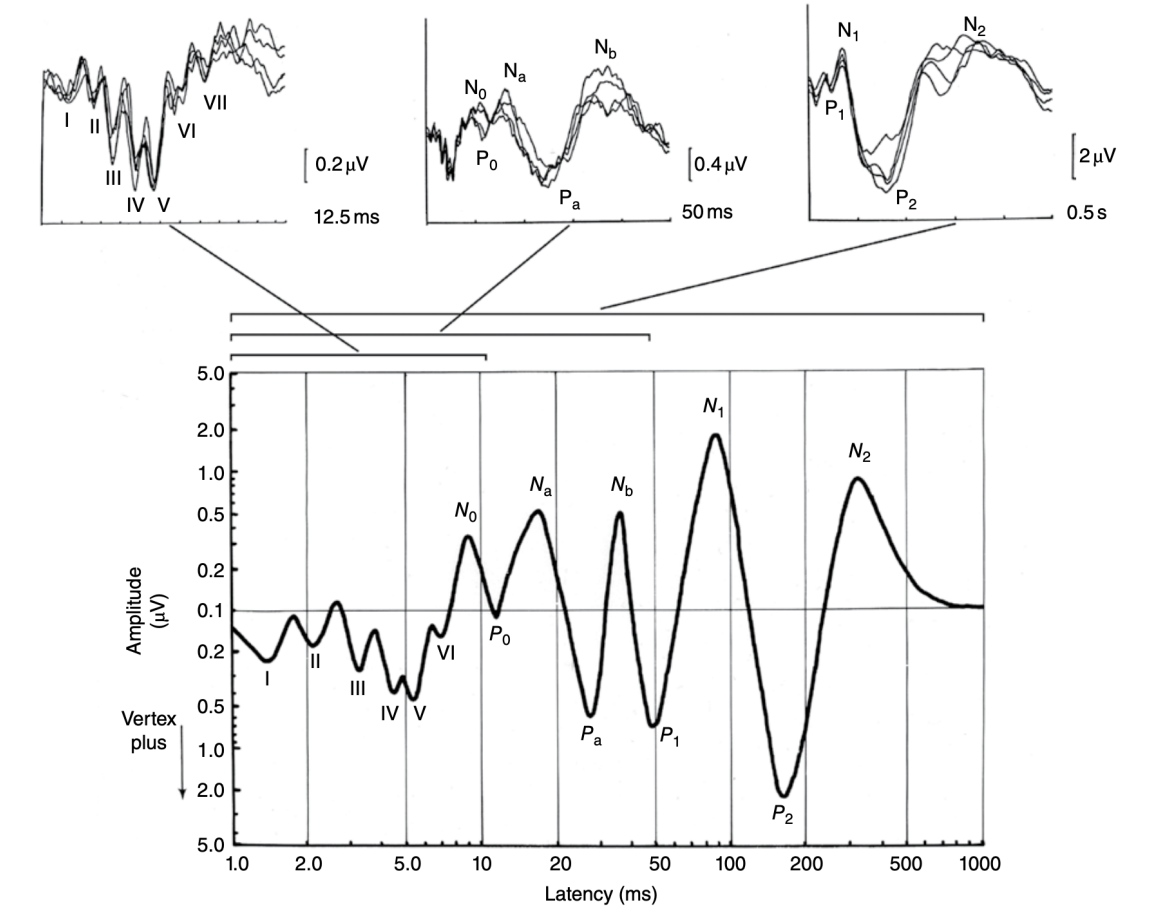

Figure 2.4: *Auditory evoked potential*. The auditory ERPs to a 60 dB_{SPL} click of 50 μs duration. Each of the panels in the top row shows four traces, each an ERP computed by averaging the EEG from 1024 trials, and each acquired from the same subject, but in four different recording sessions. Although each of these peaks is time-locked to the click, each differs with regard to the length of time following the click over which the EEG traces were averaged. Therefore, peaks are associated with neural activity ranging from brainstem responses (top-left) to higher level cortical processing (top-right). Note that for the bottom panel the scales for vertical and horizontal axis are logarithmic. Adapted from [14]

2.3 Magnetoencephalography (MEG)

A similar way to measure electrophysiological brain activity non-invasively is to record the magnetic counterpart of EEG, which is magnetoencephalography, or MEG. Like EEG, MEG is sensitive to the electrochemical current flows within and between brain cells. However, the

smearing of voltage caused by the high resistance of the skull can be largely circumvented by recording magnetic fields instead of electrical potentials. The transparency of the skull to the magnetic fields makes MEG a robust technique with a surprising spatial resolution as well as its temporal resolution; therefore, unlike the EEG method, MEG can provide a reliable source localized image of the brain (see Table 2.1).

As illustrated in Figure 2.5A, a varying electrical dipole is always surrounded by a magnetic field of proportionate strength. Therefore, when current flows in a conducting element, it produces a magnetic field in the volume surrounding the wire. A current tangential to the scalp's

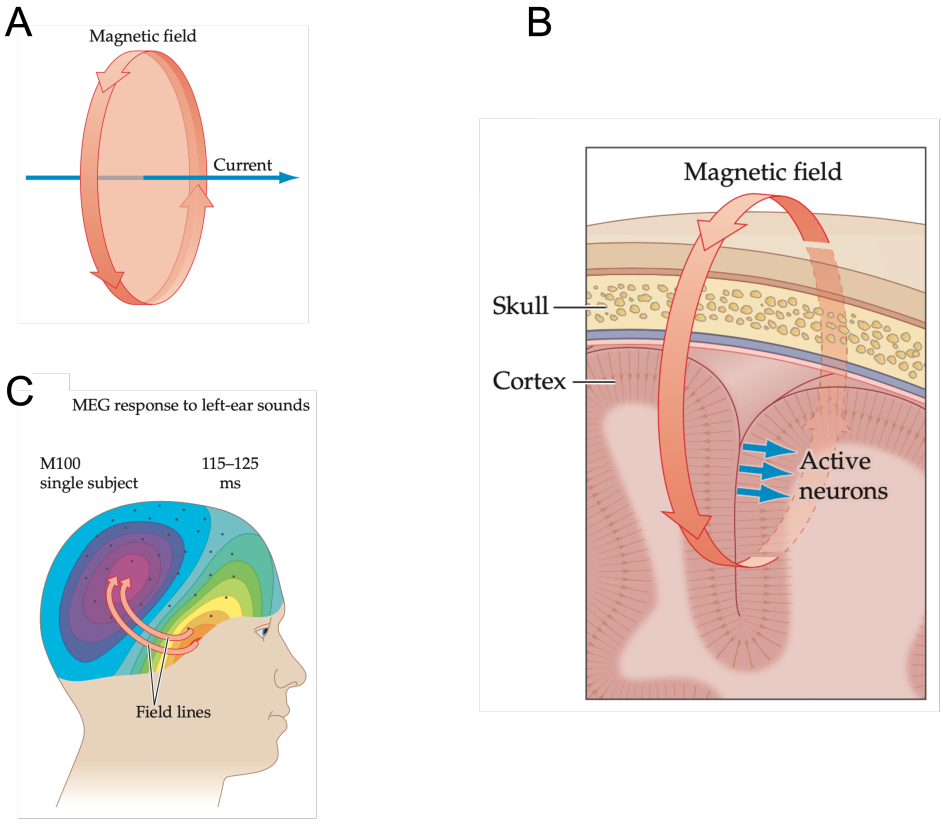

Figure 2.5: *Schematic of how MEG works.*(**A**) An electrical current in a line cause a circular magnetic field. **(B)** Analogous to part (A), the current in cortex produce circular magnetic field. The magnetic field can be captured using magnetometers with MEG. **(C)** Schematic topomap of the magnetic fields produced by brain responses 100 ms after a tone presented on left ear. Adapted from [12]

surface will be accompanied by a magnetic field that leaves the head on one side of the current and enters back again on the other side (Figure 2.5B). Suppose the field strengths are measured at different points on the surface of the head with a magnetometer. Consequently, the distribution of these values over space (aka topography) and time can be obtained, including whether they are positive (coming out of the head) or negative (going into the head). MEG measurements are conducted externally using a susceptible instrument called a superconducting quantum interference device (SQUID), which usually runs at a very cold temperature. The SQUID is a very low noise detector of magnetic fields that converts the magnetic flux threading a pickup coil into voltage, allowing detection of weak neuromagnetic signals. The magnetic signals emitted by the brain are 10^{10} order of magnitude smaller than the earth's magnetic field; thus, shielding from the external magnetic signals is necessary.

One drawback regarding MEG is its insensitivity to the current flow oriented perpendicular to the surface of the skull. Therefore, the neurons that MEG can capture tend to be located within the sulci, where the long axis of each apical dendrite tends to be oriented parallel to the skull surface. This sensitivity to sulcus activity is due to the orientation of the electrical currents that generate the magnetic fields. However, neuronal activity in a cortical gyrus produces currents that are perpendicular to the head's surface, inducing magnetic fields that are primarily parallel to the head surface and thus invisible to the nearest magnetometer. Conversely, the EEG electrodes detect voltage fluctuations produced by the volume currents, which are not affected by this limitation. Accordingly, EEG picks up voltage fluctuations from sources in both cortical gyri and sulci. Figure 2.6 shows a simulation of the EEG and MEG topomaps, which were responding to an identically simulated activity located in the frontal cortex on the right hemisphere. One can observe the orientations of Electric and Magnetic fields and their orthogonality.

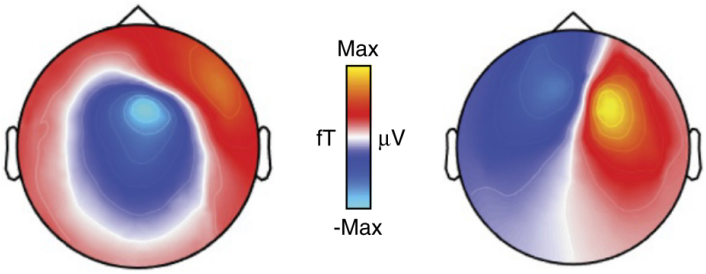

Figure 2.6: *EEG versus MEG.* An example comparing MEG and EEG. Synthetic data were generated by impressing a simulated. Adapted from [12]

Because the magnetic fields dissipate much faster than electrical fields, only EEG is sensitive to sub-cortical activities like auditory brainstem response. The earliest auditory-evoked responses picked up with MEG is in the auditory cortex peak near 50 ms after the stimulus presentation, followed by a deflection at about 100 ms (M100[1]) which has been the most investigated auditory response in MEG [15, 16]. Evidence suggests that the peak at M50 and M100 are associated with different functional systems, and also these peaks are localized to different brain regions. The M50 is mainly localized to the primary auditory cortex (A1) and activated with any sound, regardless of the task. Whereas M100 is associated with more task-dependent responses, such as detecting changes in the various attributes of the auditory inputs, or reflects attention modulated response [15–21].

[1]In MEG evoked responses, "M" is used to refer to the peaks of the event-related signal, whereas in the EEG, the letters "N" and "P" refer the polarity of the peak.

Chapter 3: All for One: Binding the Acoustic Features of an Auditory Source through Temporal Coherence

3.1 Introduction

Humans and other animals can segregate a target sound from background interference and noise with remarkable ease [22–24], despite the highly interleaved spectrotemporal acoustic components of the different sound sources (or streams) [25]. It is hypothesized that attention is important for this process to occur in a listener's brain, and that the consistent or coherent temporal co-modulation of the acoustic features of the target sound, and their incoherence from those of other sources, are the two key factors that induce the binding of the target features and its emergence as the foreground sound source [26–28]. Specifically, the temporal coherence principle implies that acoustic features underlying the perceptual attributes of a sound emanating from a single source (e.g., its pitch, timbre, location, loudness) evoke correlated neural responses, i.e., that fluctuate similarly in power over time, and that the attentive listener tracks and utilizes this neural coherence to extract and perceive the source. The definition of temporal coherence and other related terms is further elaborated upon in section 3.2.

Numerous studies have provided insights into the temporal coherence theory and tested its predictions. For example, psychoacoustic experiments have shown that perception of syn-

chronous tone sequences as belonging to a single stream is not appreciably affected by their frequency separation (from 3 semitones to over an octave) or small frequency fluctuations of the individual components, as long as the tones remain temporally coherent [29, 30]. Furthermore, it is far easier to detect the temporal onset misalignment between tones across two synchronized sequences, compared to between asynchronous (e.g., alternating) sequences [26], suggesting that temporally coherent tone sequences are perceived as a single stream [31–33]. Additional strong evidence for the temporal coherence principle was provided by a series of experiments utilizing the stochastic figure–ground stimulus, in which synchronous tones (referred to as the "figure") are found to pop out perceptually against a background of random desynchronized tones, with the perceptual saliency of the "figure" being proportional to the number of its coherent tones [34–36].

To account for the neural bases underlying the principle of temporal coherence, a recent electrocorticography (ECoG) study in human patients examined the progressive extraction of attended speech in a multi-talker scenario. It demonstrated that a linear mapping could transform the multi-talker responses in the human primary auditory cortex (Heschl's Gyrus, or HG, in humans) to those of the attended speaker in higher auditory areas. Furthermore, the mapping weights could be readily predicted by the mutual correlation, or temporal correlation between the responses in the HG sites [37]. This experimental finding is consistent with an earlier computational model for how temporal coherence could be successfully implemented by measuring the coincidence of acoustic feature responses to perform speech segregation [38]. It has also validated single-unit studies in ferret auditory cortex, which tested the importance of attention and temporal coherence in stream formation and selection, and further demonstrated that the responses and connectivity among responsive neurons were rapidly enhanced by synchronous stimuli and suppressed by asynchronous sounds, but only when the ferrets actively attended to

the stimuli [28]. Exactly the same idea has been shown to be relevant in the binding of multisensory auditory-visual streams both in cortical responses and in psychoacoustic tests [39–41], as well as to explain stream formation associated with comodulation masking release [42].

In this chapter, we sought to investigate the properties and dynamics of the temporal coherence principle using the more accessible EEG recordings in human subjects while performing psychoacoustic tasks with a wide variety of stimuli, including natural speech. The experiments tested several key predictions of the temporal coherence hypothesis (schematized in Figure 3.1) primarily the coincidence of the neural responses to any acoustic features is the fundamental and overriding determinant of the segregated perception of an auditory stream. Thus, it is not the specific nature of the features (e.g., being a single-tone, tone-complex, or a noise burst) or the harmonic relationship among the tones of the complex that determines their binding. Rather, it is the temporal coincidence among the components that matter. A second prediction of the hypothesis is that directing attention to a specific feature (e.g., a tone in a complex) not only enhances (or modulates) the response of the neurons tuned to it but would also bind it or similarly modulate the responses of all other neurons that are synchronized with it. Conversely, attending to a target sound is postulated to suppresses uncorrelated responses due to acoustic features that are modulated incoherently with the target. Another aspect of the temporal coherence hypothesis that has already been explored is the rapid dynamic nature of the binding among the components of a stream [28], which explains how listeners are able to switch attention and rapidly reorganize their auditory scene according to their desired focus. Nevertheless, the role of attention in stream formation can be somewhat ambiguous in that many studies have demonstrated streaming indicators even in the absence of selective attention [35,43]. However, even in these cases, deploying selective attention always enhances the responses, significantly confirming its important role in

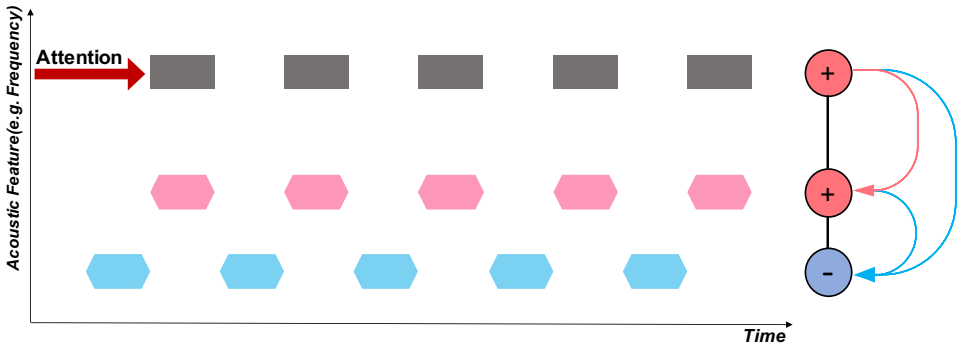

Figure 3.1: *Schematic of attention and the temporal coherence principle.* The horizontal axis is time, and the vertical axis depicts an arbitrary feature dimension of interest (e.g., spectral frequency, fundamental frequency (pitch), or location). Three separate sequences of sound tokens are depicted: two are synchronized (in black/pink), alternating (or de-synchronized) with another sequence (in blue). If attention is focused on the black sequence, its neural responses become enhanced. Because of its temporal coherence with the pink sequence, mutual excitation causes the two to bind, thus becoming enhanced together. By contrast, the blue stream is asynchronous (temporally incoherent), and hence it is suppressed by rapidly formed mutually inhibitory interactions (depicted in blue).

mediating the streaming percept.

It should be noted that conducting such experiments and analyses with EEG recordings is difficult because of the extensive spatial spread of the responses across the scalp. This introduces two types of challenges that must be overcome. First, it is hard to resolve and assess the binding of individual frequency components in a complex or in a speech mixture that contains many other nearby components. Second, because the responses from many neural sources interact and superimpose in EEG recordings, a response enhancement due, for example, to attention to a specific feature may be accompanied by suppression of responses from a competing feature. Hence, the total response becomes instead manifested as a complex, unintuitive modulation of the response patterns. Both challenges can be overcome by the techniques presented in this report. Specifically, we addressed the spectral resolution problem by presenting isolated tone probes soon after the end of the complex stimuli. By aligning a probe tone with various spectral components of

the preceding stimulus and measuring the persistent effects of attention on its responses just after the stimulus, we could detect the attentional effects on the responses to individual components within these complex stimuli. Furthermore, to decode and assess these changes directly from the complicated distributed EEG responses, we resorted to a pattern classification technique that quantified whether attention significantly altered (or modulated) the response patterns and for how long before and after.

3.2 Experimental Design and Procedures

3.2.1 Terminology

In the present chapter, several terms are used somewhat interchangeably to refer to the idea of temporal coherence, which more specifically states that neural responses that are temporally correlated over a short period of time on a pair of sensory pathways can evoke a unified percept or "become bound" into a single stream. Sometimes the reference is made instead to the stimuli that evoke these responses, and hence terms like synchronized tone sequences or coincident events occurring over a time period can mean the same thing as temporal coherence. In all these cases, the context will hopefully clarify the intent as it is by no means necessary that any of these stimuli can unambiguously evoke the necessary correlated activity. For instance, a single pair of synchronized events are irrelevant to stream formation since coincidence must occur multiple times over a short interval. Similarly, synchronized bursts of random tone complexes do not evoke coherently modulated activity on any pair of frequency channels and hence do not bind. For more examples of such conditions, please see [27].

3.2.2 Participants

76 young adults with normal hearing (ages between 19 and 31) participated in this study, consisting of four experiments. 18, 14, 21, and 23 subjects participated in experiments 1-4, respectively. Experiments 1 and 3 were conducted at the University of Minnesota and experiments 2 and 4 were conducted at the University of Maryland. All participants were given course credits or monetary compensation for their participation. The experimental procedures were approved by the University of Maryland and the University of Minnesota Institutional Review boards. Written, informed consent was obtained from each subject before the experiment.

3.2.3 Data Acquisition and Stimuli Presentation

Data were collected at two sites. At University of Maryland, Electroencephalogram (EEG) data were recorded using a 64-channel system (ActiCap, BrainProducts) at a sampling rate of 500 Hz with one ground electrode and referenced to the average. We used a default fabric head-cap that holds the electrodes (EasyCap, Equidistant layout). EEG data from University of Minnesota were recorded from 64 scalp electrodes in an elastic cap, using a BioSemi ActiveTwo (BioSemi Instrumentation). EEG signals were acquired at a sampling rate of 512 Hz and referenced to the average. We analyzed the EEG data offline. The stimuli were designed in MATLAB and presented to the participants with the Psyctoolbox [44–46]. The stimuli audio was delivered to the subjects via Etymotics Research ER-2 insert earphones at a comfortable loudness level (70dB).

3.2.4 Stimulus Design

We explored the consequences of temporal coherence on simple harmonic tone-complexes at a uniform rate to inharmonic complexes, irregular presentation rates, mixed tone, and noise sequences, and ending with speech mixtures. Details of the experimental design are described below. In addition, Figure 3.2 summarizes stimulus constructions for all experiments.

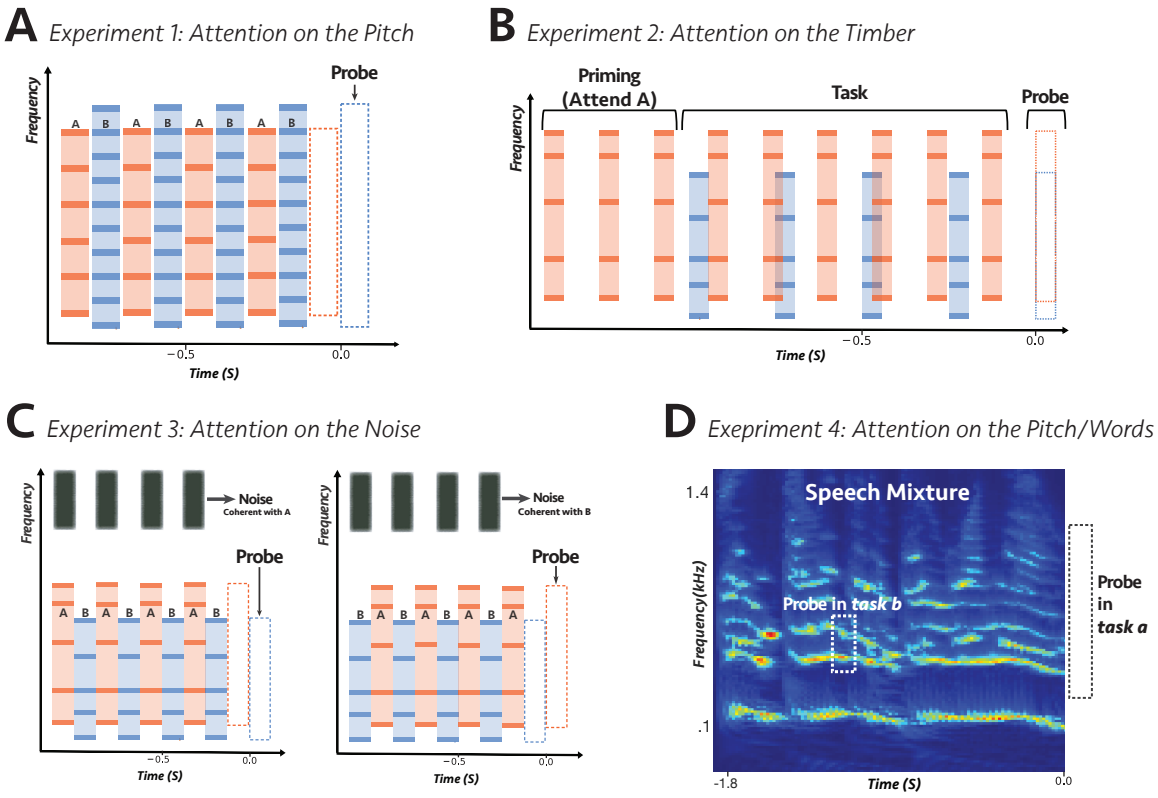

Figure 3.2: *A summary of stimulus construction in all experiments.* In all four experiments, the auditory scene consisted of two concurrent auditory sequences. Participants were asked to pay attention to one of the sequences and ignore the distractor. We used an intensity deviant detection task for the first three experiments (panels A, B, and C). In experiment 4, we asked participants to report a specific word in the target voice (panel D). To do the task, listeners had to focus their attention on the pitch of the harmonic complex sequence in experiment 1 (panel A), attend to the timber of the inharmonic complex sequence in experiment 2 (panel B), attend to the noise sequence in experiment 3(panel C), or pay attention to the target word and pitch of the target voice in experiment 4(panel D). We used a probe-tone paradigm to investigate the effect of attention on individual frequency components of auditory objects.

Experiment 1

The stimulus was presented as an alternating **ABAB** sequence with a sampling rate of 24414 Hz, followed by a short probe-tone. Segment **A** was a harmonic tone complex with a fundamental frequency (F0) of 400 Hz, while **B** was a harmonic tone complex with an F0 of 600 Hz. All **A** and **B** harmonic complexes were generated with random starting phases and were low-pass filtered to exclude frequency components higher than 4000 Hz with a 48 dB/oct filter slope. Each segment of **A** or **B** lasted 90 ms, including 10-ms raised-cosine onset and offset ramps. Segments were separated by 20-ms gaps (the repetition rate of both A and B tones alone was 4.55 Hz). The total number of **AB** harmonic tone pairs in one trial was randomly chosen from 27 to 33, so participants could not predict the sequence's total duration. The sequence was followed by a 100-ms silent gap and a 90-ms pure-tone probe. The level of probe-tones and each component of A and B harmonic complexes were at a rms level of 55 dB SPL. The ending tone complex in the sequence was balanced so that half the trials ended with tone complex A. The other half ended with tone complex B (see Figure 3.3A), thus based on these structures, we defined 2 different conditions for the relative position of the probe and the last complex tone in the sequence: (1) When the probe tone is a component of the last complex, and therefore it occurred at the supposedly expected spectral location (e.g., the sequence ended with tone complex A and we probed the frequency channel unique to complex A), and conversely (2) when the probe tone was a component of the penultimate complex tone and occurred at an unexpected spectral location (e.g., the sequence ended with tone complex B and we probed the frequency channel unique to complex A).

The experiment included six blocks of 100 trials each. Participants were instructed to selectively attend to the A tones (the higher pitch) in half the blocks and the B tones (the lower pitch) in the other half of the blocks by explicitly ask them to pay attention to the higher or lower

pitch sequence. The order of blocks was randomized. The 300 trials for each set of instructions were equally divided into two groups, depending on the probe-tone used. The two probe-tones were 2000 Hz (unique frequency component of A) and 3000 Hz (unique component of B). The order of trials with different probe-tones was randomized within the three blocks for each set of instructions.

We used intensity deviant in both A and B harmonic tones to monitor stream segregation and attention. There were 0 - 3 deviant tones in each sequence, and the number of deviants was uniformly distributed across trials. Out of 100 trials within a block, there were 25 trials with intensity deviant segments only in the A sequence, 25 trials with deviants only in the B sequence, 25 trials with deviants in both sequences, and 25 trials with no deviant. Deviant tones were presented at a level 6 dB higher than the regular tones. Participants were instructed to press a button (0, 1, 2, or 3) after hearing the probe-tone at the end of each sequence to answer how many intensity deviants they detected within the attended stream while ignoring deviants in the unattended stream. Deviants were prevented from occurring in the first five and last three tone pairs. To analyze the EEG data, we kept the trials in which participants reported correctly, e.g., the exact number of deviants in the target sequence.

Experiment 2

The auditory scene consisted of two tone-complexes denoted as **A** and **B** and presented at a sampling rate of 44100 Hz, 90 ms in duration, 10 ms cosine ramp, and 150 ms onset-to-onset interval for tone complex A (6.67 Hz repetition rate) and 250 ms interval for tone complex B (4 Hz repetition rate). Each complex tone consisted of 5 predefined inharmonic frequencies with one shared frequency component between the two complexes (F_A = [150, 345, 425, 660, 840] Hz and F_B = [250, 425, 775, 1025, 1175] Hz). The two sequences were presented at different

rates but converged on the last tone in the sequence. A single frequency probe-tone replaced the converged tones; thus, the probe-tone always occurred at the expected location (see descriptions for experiment 1). The probe-tone was centered at the frequency that was shared between the two complexes (425 Hz) or at a frequency unique to complex A (660 Hz) or complex B (775 Hz). The participants were instructed to pay attention to a complex sequence that was included in the priming epoch and report whether they heard an intensity deviant (5 dB increase in the loudness) only in the target sequence, with a 20% chance of having a deviant in complex A sequence and independently 20% chance of having a deviant in the sequence of complex B. Deviant was prevented from occurring in the first and last two tones; to analyze the EEG data, we only kept the trials in which participants reported correctly (i.e., all trials with misses and false alarms were discarded). Each trial started with a target stream alone (priming), and after three bursts, the distractor stream was added (Figure 3.4A). The priming phase was balanced for all trials, so half the trials started with tone complex A, and the other half started with tone complex B. Trials duration were uniformly distributed between 3.5-5 sec to avoid the formation of expectation of the ending.

The experiment was conducted in six blocks of 100 trials, and attention was fixed on one stream throughout an entire block. The probe-tone frequency was uniformly selected from [425 Hz, 660 Hz, 775 Hz] for all trials. Before neural data collection, a training module was provided. Subjects received feedback after each trial for both training and the test sets.

Experiment 3

This experiment consisted of a narrowband noise sequence with a passband of 2-3 kHz. The noise was accompanied by two inharmonic complex sequences, with one of them coherent with the noise sequence, and the other complex was alternating. We used the same frequency components

as experiment 2 for both inharmonic complexes (F_A = [150, 345, 425, 660, 840] Hz and F_B = [250, 425, 775, 1025, 1175] Hz presented at a sampling rate of 24414 Hz, and a level of 70 dB SPL). Each noise segment and tone complex were 125 ms in duration, including a 15 ms offset and onset cosine ramp with a 250 ms onset-to-onset time interval for all tones. We used 0 - 3 (with equal probability) intensity deviants, a 6 dB increase in amplitude, both in the target of attention (noise) and the distractor (alternating complex), there was no deviant in the complex tone coherent with the noise sequence. The subjects' task was always to pay attention to the noise sequence and count the number of intensity deviants only in the attention target (noise). To analyze the EEG data, we only kept the trials where subjects reported the exact number of deviants. The trials' duration was uniformly distributed between 3.5-5 secs, so participants could not form an expectation for the sequence's total duration. We inserted a single frequency probe-tone 125 ms after the last tone complex in the sequence. The probe-tone was at the frequency shared between the two complexes (425 Hz) or was unique to complex A (660 Hz) or complex B (775 Hz). To ensure that the EEG response to the last complex tone does not affect the probe-tone response under different attention conditions, trials ended with complex tone A when the probe-tone was unique to B and ended with complex tone B when the probe-tone was unique to A.

The experiment was conducted in six blocks of 100 trials. For 3 blocks, complex A and the rest of the blocks complex B were coherent with the noise sequence. The probe-tone frequency was uniformly selected from [425 Hz, 660 Hz, 775 Hz] for all trials within a block. Before neural data collection, a training module was provided. Subjects received feedback after each trial for both training and test sets.

Experiment 4

We used the CRM speech corpus (Bolia et al. 2000) for this experiment. In general, this speech database consists of 8 different speakers (4 female), and the format of each speech sentences is: "Ready [Callsign] go to [Color] [Number] now".

The callsign set is: {'Charlie', 'Ringo', 'Laker', 'Hopper', 'Arrow', 'Tiger', 'Eagle', 'Baron'}. The color set is: {'Blue', 'Red', 'White', 'Green'}, and the number set is: {1, 2, 3, 4, 5, 6, 7, 8}. Therefore, each speaker has 256 unique sentences. Two speakers (one female and one male) were chosen for the task. We manipulated each sentence's duration, so all the sentences had the same length (1.8 seconds at a sampling rate of 40000 Hz), and on average, the Callsign occurred 300 25 ms after the speech onset. The experiment consisted of two different tasks. For the EEG analysis of both tasks, we only included trials in which participants answered correctly, i.e., the listeners' reports matched with the color or number uttered by the target speaker (see below). The two tasks were:

Task a: During this task, the auditory scene consisted of two concurrent speech streams - we constrained the mixtures to have different 'callsigns', 'colors' or 'numbers' in male and female voices - that were followed by a probe-tone with complex harmonics. The probe tone's harmonic frequencies were aligned with the 4 loudest harmonic frequencies of either male or female voice at the end of the sentences; therefore, the probe-tone was unique to male or unique to female voices. The probe's duration was 90ms, with a 10 ms cosine ramp, and was played after a 10ms interval after the sentences (Figure 3.6A). The experiment was conducted in 4 blocks with 100 trials. Participants attended to either the male or female voice throughout the 100 trials of a given block, and after each trial reported the color or number of the attended speaker who was designated randomly at the end of each trial. The order of the blocks was shuffled across subjects.

Task b: In this second part of the experiment, we inserted a single frequency probe-tone in

the middle, following 600 ms of the speech onset and around 300 ms after the callsign onset. The probe-tone was 90 ms in duration, including a 10 ms cosine ramp, followed by a 10 ms gap of silence. The frequency of the probe-tone was aligned with the 2nd harmonic of the female voice (unique to female, average F_F = 391 Hz) or the 3rd harmonic of the male voice (unique to male, average F_M = 288 Hz). Although the probe tone was presented in the speech, the speech mixture was masked by complete silence for the probe-tone duration. The experiment was conducted in a block of 400 trials, and participants were instructed to pay attention to the speaker who uttered the target callsign ('Ringo') and report either the color or number (randomly selected for each trial) spoken by the target voice, at the end of each trial.

3.2.5 EEG Preprocessing

After loading, EEG data were mean-centered. The bad channels which exceeded a threshold criterion (standard deviation of channels amplitude) were detected and were interpolated based on the data from the neighbor channels. The slow varying trend in data was removed by robust fitting a polynomial on data and then subtracted from it [47]. For the DSS analysis (see below section "Denoising Source Separation (DSS)"), data were bandpass filtered between 1 Hz to 20 Hz with Butterworth window of order 4 using 'filtfilt' in MATLAB. Eyeblink components were isolated and projected out with the HEoG and VEoG channels using a time-shift PCA [48]. Data were referenced by subtracting the robust mean and epoched based on the triggers sent at the beginning of each trial. Finally, the outlier trials (bad epochs that exceeded the standard deviation of channels amplitude) were detected and discarded based on a threshold criterion.

3.2.6 Decoding

Decoding analysis was performed using sci-kitlearn [49] and MNE [50] libraries in python 3.6. We trained linear classifiers on EEG sensor space signals, band-passed 0.1-20 Hz at 250 Hz sampling frequency [51]. At each time point t, we trained a classifier using the matrix of observations $X_t \in R^{N\times 64}$, for 64 electrodes in N samples, to predict the vector of labels $y_t \in \{0,1\}^N$ at every time point t' in a trial. The labels correspond to the two attention conditions (attend A versus attend B in experiment 1, 2, and 3 or attend female versus attend male in experiment 4). For example, for each subject, we trained the decoders on EEG signals at time points encompassing the probe tone (-100 ms - 500 ms). Therefore, the decoder at each time point learns to predict the attended stream using the EEG sensor topography at the same time point. Then, we generalized the trained decoder by testing it on all other time points of the trial. Logistic regression classifiers were used, with 5-fold cross-validation, within-subject for all the trials. We used the area under the receiver operating characteristic curve (AUC) to quantify the classifiers' performance. In summary, within a subject, the classifiers' scores imply the robustness of the attentional effects on the probe-tone response topography. So, the significant time regions in all figures corresponding to decoder scores indicate the effect's consistency across all subjects.

3.2.7 Denoising Source Separation (DSS)

A set of spatial filters are synthesized using a blind source separation method that allows the measurement of interest to guide the source separation. For detailed explanation see [52]. For our purpose, the Denoised Source Separation (DSS) filter's output is the weighted sum of the signals from the 64 EEG electrodes, in which the weights are optimized to extract the repeated neural

activity across trials. Therefore, for the experiment 1, 2, and 3, the first DSS component reflects a brain source of auditory processing, repeatedly evoked during the segregation task for the same set of sound frequencies. Our use of the DSS method required a large number of the same stimuli to extract the repeated activity. However, in our speech experiment (experiment 4a and 4b), since each trial consisted of various sentences with varying sound frequencies, different neural activities were driven by the stimulus in different trials; therefore, it is difficult to isolate the first DSS component as we did for tone experiments. Thereby, we only used the DSS method in order to denoise the data in experiment 4, in which we projected back the first 5 DSS components to the sensor space to form a clean and denoised dataset. Finally, we compared the evoked responses at the Cz channel (placed on the center of the mid-line sagittal plane) for experiment 4.

3.2.8 Statistical Analysis

Statistical analysis of the decoder results was performed with a one-sample t-test with random-effect Monte-Carlo cluster statistics for multiple comparison correction using the default parameters of the MNE SPATIO_TEMPORAL_CLUSTER_1SAMP_TEST function [53]. To compare the differences in evoked responses due to the probe-tones, we performed bootstrap resampling to estimate the standard deviation (SD) of the difference between the attention conditions. We checked whether at each time point the difference between attention conditions exceeded $2\times$ estimated SD (2σ). In supplement figures, for the first DSS component's strength comparison between two attention conditions, we used a one-tail non-parametric Wilcoxon signed-rank test [54]. Error bars in all figures are ±SEM (standard error of the mean).

3.3 Results

The results described in the present chapter are of EEG experiments conducted on normal-hearing subjects. Details of the experimental setup, subjects, and stimuli are provided in each subsection below, as well as in section 3.2. The experiments begin by exploring the consequences of temporal coherence on simple harmonic tone-complexes at a uniform rate and progress to inharmonic complexes, irregular presentation rates, mixed tone, and noise sequences, and ending with speech mixtures.

Note that all of the stimulus paradigms in this study were selected to closely resemble "classical" paradigms of streaming, e.g., alternating and synchronous tones and complexes. Thus, properties of the streaming precepts associated with these stimuli are already well-established and have been studied extensively, as we shall point out. Furthermore, the objective segregation measures we employ closely follow widely-used "deviant-detection" paradigms [23, 26, 55, 56].

3.3.1 Experiment 1: Binding the Harmonic Components of Complex Streams

In this experiment, we manipulated the streaming percepts evoked by alternating harmonic complexes of different pitches [22, 57, 58]. We specifically investigated how attention to one of the two streams modulates the neural response to the complexes' individual constituent tones. For example, consider the two alternating sequences of harmonic complex tones in Figure 3.3A. The complexes in the two streams had fundamental frequencies of F_A = 400 Hz and F_B = 600 Hz, 90 ms in duration, and were separated by 20 ms gaps, with 10 ms raised-cosine onset and offset ramp. When attending to one stream, it is known that the EEG responses of the attended complexes become relatively enhanced [59–61]. However, it is unclear whether this enhancement is due to

enhanced responses to the individual tones within the attended complex or just an enhancement of the channels selectively responding to the complexes' pitch. Conceptually, we shall hypothesize that the attentional focus on one stream effectively confers a steady (persistent) enhancement of the responses in the frequency channels of the constituent tones, specifically those tones that are unique to the attended stream. In the next experiment, we explore the fate of tones that are shared between the two streams.

Because of the poor spatial resolution of the EEG recordings, it was difficult to investigate the attentional effects on individual frequency components during the simultaneous streams. Instead, we probed the persistent modulatory effects of attention on individual frequency channels immediately following the end of the streams. This was done by presenting a 90 ms pure probe-tone after a 100 ms silent gap. The probe tone was aligned with the frequency of a harmonic that is either unique to complex A (3000 Hz) or tone complex B (2000 Hz). There were two conditions for the timing of the probe-tone (Figure 3.3A): "expected", in which the probe was a component of the last complex tone in the sequence (note that there is a gap between the last complex in the sequence and the probe-tone) or "unexpected" where it was a component of the penultimate complex tone in the sequence. The reason for these two conditions was to ascertain that the modulation of the probe-tone responses was not related to its violation of expectations (akin to the effects of "mismatched negativity") but rather to the persistent effects of attending to one stream versus another. 18 normal-hearing participants were instructed on each trial to selectively attend to tone complex A or B and report the number of intensity deviants in the attended stream - with the deviant tone -complex is 6 dB louder than other tones in the sequence. Subjects reported hearing the two streams and being able to attend reliably to one as the behavioral results indicate, with all subjects reported the correct number of deviants above the chance level (Figure

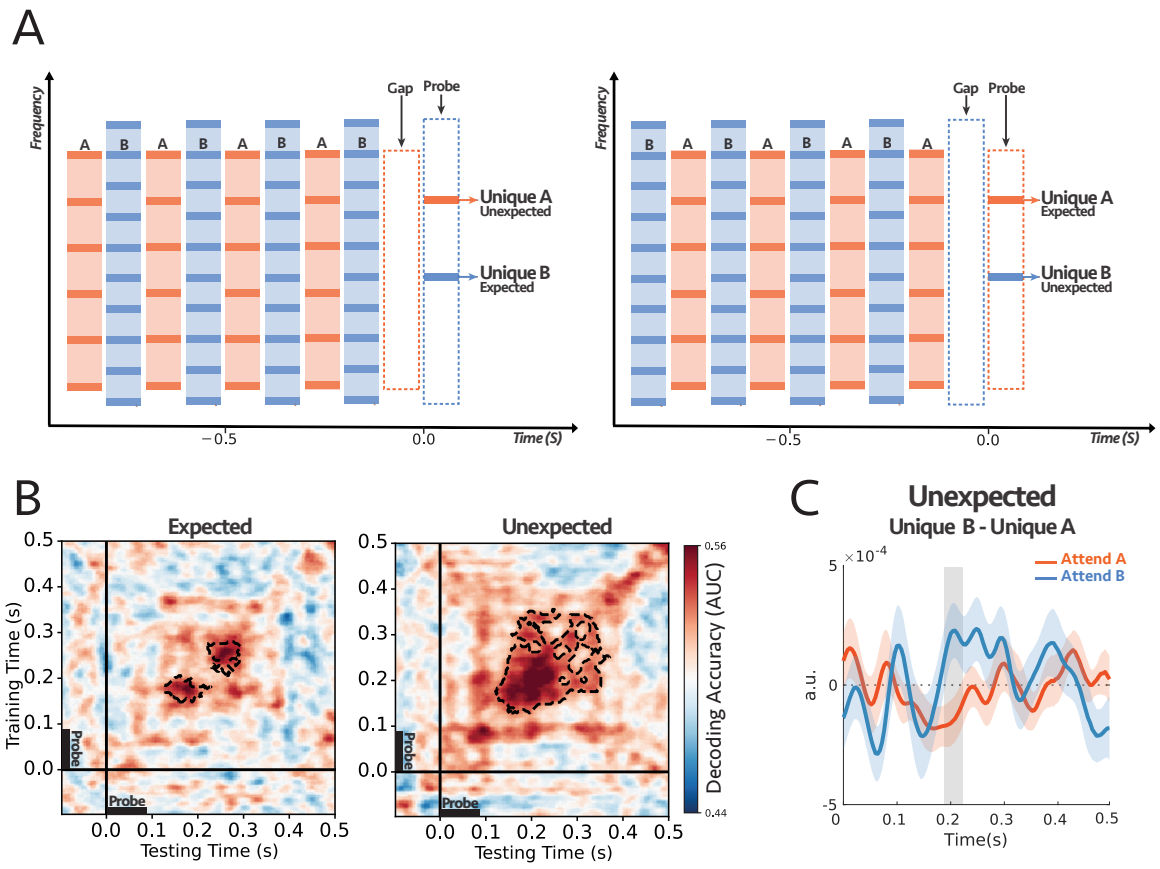

Figure 3.3: **(A)** The stimulus consisted of two complex sequences with harmonically related tones. In each trial, subjects were instructed to selectively attend to high or low pitch sequence and counted intensity deviants in the target stream. One of the frequency channels unique to tone complex A (unique A) or complex B (unique B) was probed at the end of each trial. **(B)** Decoding performance for Expected (left) and Unexpected (right) probe-tones (see the text for further explanation on the difference between the two paradigms). For each subject, classifiers were trained on the signals from all 64 EEG sensors and tested separately at each time in a 600 ms time window encompassing the probe tone (-100 ms to 500 ms). The trained classifiers tried to decode the target of attention within the mentioned time window. The cluster-corrected significance was contoured with a dashed line, p = 0.038 for expected and p = 0.004 for unexpected. Thus, both probes demonstrated significant attentional modulations of individual harmonic components, although the classification patterns were different between the two conditions. **(C)** Comparing the difference in responses to unique A and unique B probe-tones for two attentional conditions for the first DSS component (see text and section 3.2.7 for details). The comparison was significant for unexpected probes at around 200 ms after the probe onset (shaded area larger than 2σ), with reverse polarity for attend A and attend B suggesting the opposite effect of attention on coherent and incoherent tones (i.e., enhancement versus suppression). **(n=18)**.

3.8). To dissect the attentional effects on the responses to the two streams, we trained a set of independent logistic regression classifiers using data from all EEG sensors as explained in detail in section 3.2.6 [51, 62, 63]. These classifiers trained on the EEG responses to the probe-tones at each time point t and tested at time t' - where t and t' were within the probe-tone time window (-100 ms to 500 ms) - in order to predict the target of attention. In other words, the trained classifiers tried to linearly separate the attentional conditions based on the differences within the probe topomaps. At the subject level, the classifier scores reflected the robustness of the effect across the trials (see the example in Figure 3.9), and at the second level, we checked the consistency of the effect size across all subjects (Figure 3.3). It should be noted again here that, because of the complex spatial spread of the EEG, the decoders can detect if response patterns across the scalp are modulated by the attentional focus on the unique components, but they cannot readily indicate whether the effects are simple response enhancements. For this kind of additional information, we resorted to a Denoising Source Separation (DSS) procedure to extract and examine the principle response component as detailed later below.

The performance of the decoders is depicted in Figure 3.3B. The scores of the classifiers are significantly above the chance level for both Expected (p=0.038) and Unexpected (p=0.004), starting from about 150 ms to 350 ms after the probe-tone onset. This performance level (up to 0.60) is commensurate with that reported in previous studies [64, 65]. Thus, regardless of probe-tone timing and its different response dynamics due to its (expected or unexpected) context, the results demonstrate the persistent, significant differential modulatory effects of attention on the unique individual harmonic components of the attended and unattended sequences. We should note that the decoder significant regions differ between the two conditions of ”expected” and ”unexpected” probes, likely because of the differences in the detailed temporal response patterns

of the probes, as well as the effects of EEG noise which may render insignificant the response modulations at different epochs following the onset. Nevertheless, in both cases of the expected and unexpected probes, there were significant attentional modulations of the responses.

Finally, we attempted to extract an additional comparison among the probe responses under the different attentional conditions using a denoising procedure on the EEG recordings. Specifically, we isolated the most repeatable auditory component from the EEG responses to the probe-tones across trials using the DSS spatial filter (see section 3.2.7; [66]) and compared the average waveform of the first DSS component and its amplitude over subjects under different conditions. Importantly, we compared the difference in responses to the probes: $UniqueB - UniqueA$, under the attend A and attend B conditions. We hypothesized that, although the DSS waveform is a complex mixture of the EEG responses on all electrodes, if attention enhances coherent and suppresses incoherent responses, then the difference between the probes' DSS responses would be modulated by attention in opposite directions, i.e., the difference: $UniqueB - UniqueA$ would have the opposite signs for attend A and attend B, reflecting the enhancement and suppression due to attention. This was indeed the case as seen in Figure 3.3C for the unexpected case, where the difference in probes' responses was significantly modulated with a reversed polarity, at around 200 ms following the probe's onset (shaded interval). It should be noted that the extracted responses used for the DSS differences (Figure 3.3C) and the classifiers (Figure 3.3B) are of a very different nature, and hence it is unsurprising that the detailed timing of the significance epochs following the probe-tone onsets would differ.

In the next experiment, we repeat the measurement of the probe-tone responses but now using A and B complexes with different temporal structure. We also explore probe-tone response modulations when it is aligned with components shared by both A and B complexes and deter-

mine whether harmonicity is necessary altogether to induce these differential attentional effects and hence play a role in segregation.

3.3.2 Experiment 2: Binding of Inharmonic Components in Complex Streams

We extended here the results of the previous experiment in several directions. First, we examined whether the modulatory effects of attention on the individual components of a tone complex depended on the harmonicity of the complex. Second, we monitored whether components shared between the two complex sequences experience any differential modulation. This is an important question because we had hypothesized that attention is a slow or steady-state enhancement of the components of one sequence. A shared frequency channel (by definition) belongs to both the attended and unattended streams, and hence if it is subjected to attentional effects, it must experience rapidly alternating enhancement and suppression, which would violate our hypothesis. Instead, our hypothesis predicts that shared components would not be differentially affected by selective attention to either stream. Third, temporal coherence is independent of the exact temporal rates or regularity of the sequences (as long as they are roughly between 2-20 Hz [67]. Consequently, we expected temporal coherence to be equally effective for streams of different rates (tone complexes that are temporally incoherent with each other), regardless of whether the tone complex is harmonic or inharmonic. It is, of course, expected that the streaming percept is modulated by all these parameters, i.e., whether the components of a sound are intrinsically more glued together by harmonicity regardless of streaming. By using unequal sequence rates, it was possible (as we shall elaborate) to eliminate any difference in timing expectations between the two attentional conditions and hence confirm the validity of the earlier results con-

cerning the attentional modulation of the probe tone responses regardless of whether they were due to expected or unexpected contexts.

Normal-hearing adults (21) selectively attended to one of two streams - each 90 ms in duration, with 150 ms and 250 ms inter-stimulus interval for complex A and B - based on the priming phase at the beginning of each trial and reported whether they detected a deviant in the attended stream, ignoring deviants in the unattended stream. The two streams clearly differed by their timbre, and it was relatively easy for the listeners to track one or the other stream (see behavioral results Figure 3.11). The streams ran at different rates but converged to be synchronous on the last tone in the sequence. We replaced the last tone with a single frequency probe-tone to always occur at the expected time, regardless of which stream was the target of attention (Figure 3.4A). We measured the neural responses to the probe as a function of whether its frequency belonged to one (unique) or both complexes (shared) (see Figure 3.4A). We should note that this paradigm was effectively used previously to explore the effects of streaming on detecting timing misalignments between streams [26].

Similar to the previous experiment's analysis, a set of linear estimators were trained to determine the effect of attention encompassing the period of probe-tone (-100 ms to 500 ms; Figure 3.4B). The classifiers were trained on the EEG signals to predict the attentional conditions for probe-tones, therefore for each subject, the scores summarized the effect of attention within the probe-tones (see Figure 3.10). For the unique probe-tones, the performance of the classifiers in decoding attention was reliably above the chance level at 50 ms and lasted until about 400 ms after the onset of the probe, with a peak around 120 ms ($p=0.0003$), indicating the persistent effects of attention on the unique components. However, the classifiers could not distinguish between the attention conditions when the probe-tone was shared between the two tone-complexes, sug-

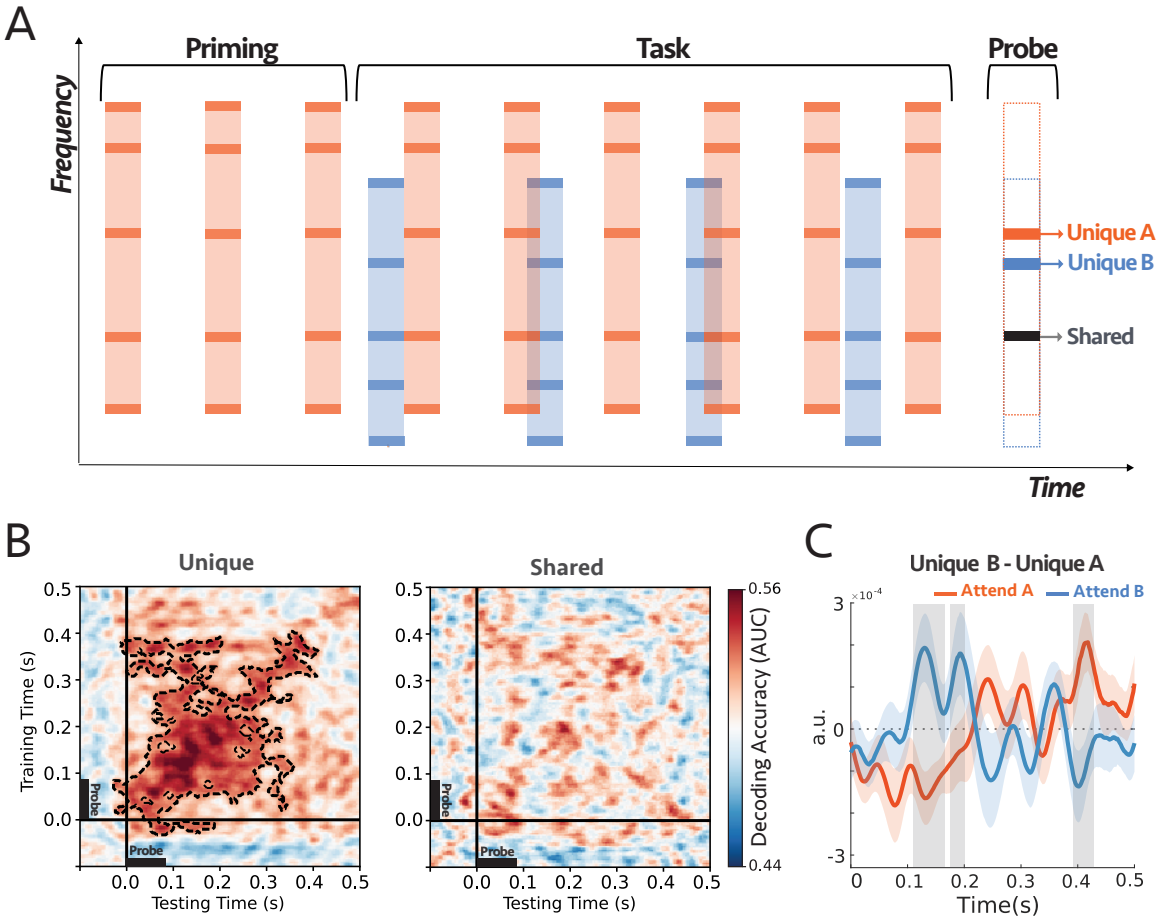

Figure 3.4: **(A)** The stimuli started with a target sequence alone (Priming epoch), and after three bursts, the distractor stream was added. Both tone complexes consisted of inharmonic frequencies (see 3.2). The two sequences were presented at different rates but converging on the last tone in the sequence, which was replaced by a single frequency probe-tone centered at the frequency that was either shared between the two complexes, unique to complex A (Orange), or unique to complex B (Blue). In each trial, subjects were instructed to selectively attend and detect an intensity deviant in the target sequence. **(B)** Decoding performance for unique (left) and shared (right) probe-tones. Classifiers trained and tested separately at each time in a 600 ms time window of the probe-tone (-100ms to 500ms). Cluster-corrected significance is contoured with a dashed line. The classifiers could decode attention only when the probes were unique components (p=0.0003). The two sets of scores were statistically different as depicted in Figure 3.12. **(C)** Comparing the difference in the first DSS component of the responses to unique A and unique B probe-tones in the two attentional conditions (Figure 3.13; see section 3.2.7). The comparison was highly significant for shaded areas (larger than 2σ), with reverse polarity for A and B, suggesting the opposite effects of attention on coherent and incoherent tones (i.e., enhancement versus suppression). **(n = 21)**.

gesting that the shared frequency channels remained on average undifferentiated by the selective attention.

3.3.3 Experiment 3: Binding Noise Sequences with Coherent Tones Sequences

l Experiments 1 and 2 confirmed that the binding of the components within a stream relies primarily on their temporal coherence and not on any harmonic relationship among them and that different sequences segregate well when running at different rates. Here, we investigate binding one step further to demonstrate that temporally coherent sound elements bind perceptually to form a stream even when they are of a different nature, e.g., tones and noise-bursts, and even when placed far-apart along the tonotopic axis. Specifically, Experiment 3 tested the hypothesis that attending to a distinct stream of noise bursts not only will modulate its neural responses it also affects all others that are temporally coherent with it, effectively binding the percept of the noise with the coherent tones to form a single unified stream. Moreover, we tested again if shared tones remain uncommitted as we found earlier, hence contributing equally to both streams.

Figure 3.5A illustrates the stimuli and procedures used with 14 normal-hearing subjects who were instructed to focus attention on a sequence of narrow-band noise bursts - between 2-3 kHz, 125 ms tone duration, and 250 ms inter-stimulus interval - and report the number of deviants that occurred in the noise stream. Two sequences of inharmonic tone complexes (A and B) accompanied the noise sequence, one coherent with the noise (the attended stream) and one alternating (incoherent) with it. At the end of each trial, a single frequency probe-tone was presented at frequencies aligned with either a unique component of complexes A or B or a component shared by both complexes. All subjects perceived the streaming of the alternating stimuli and

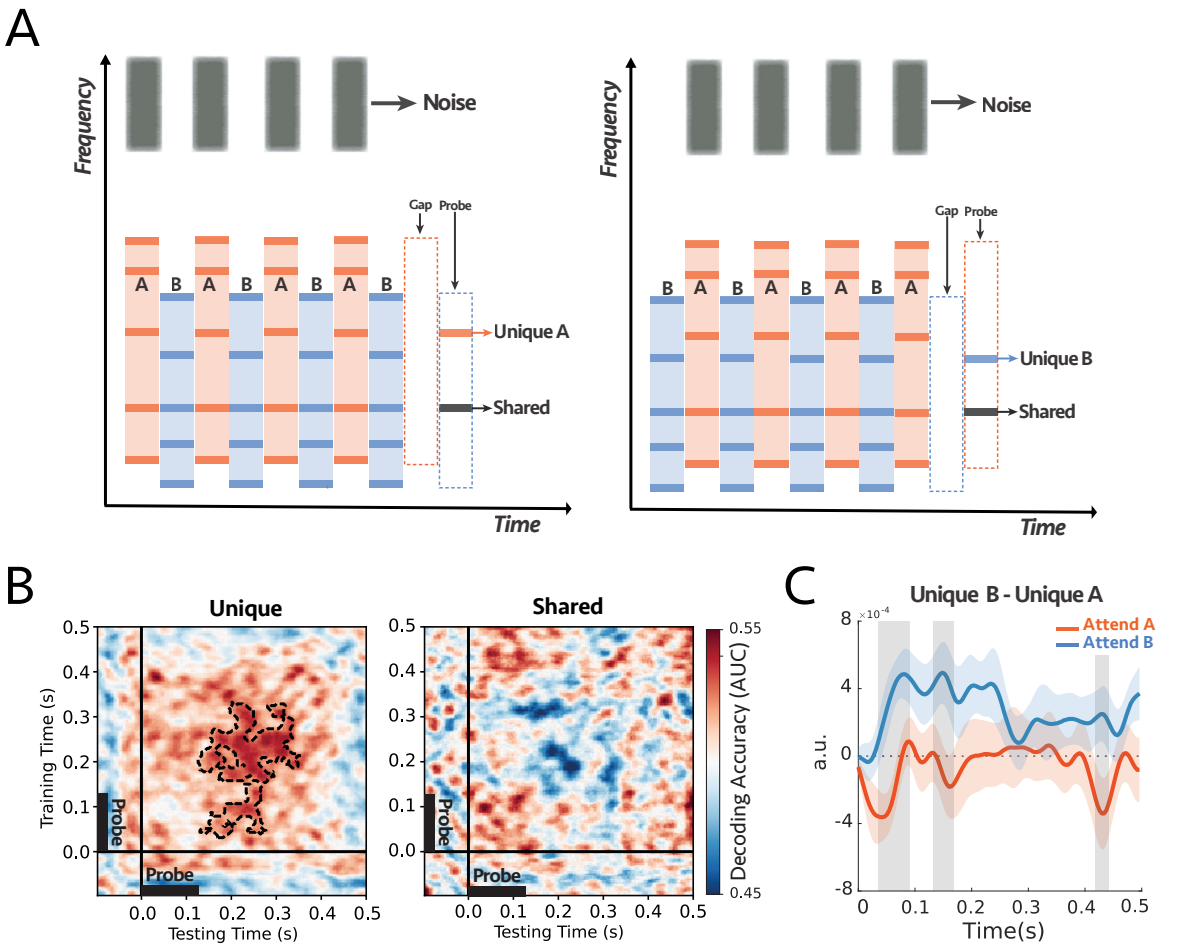

Figure 3.5: **(A)** The stimulus consisted of two complex sequences with inharmonic frequencies - the frequency components of the complex tones were the same as experiment 2 - and a sequence of noise as a target stream. In each trial, subjects were instructed to always attend and count intensity deviants in the target noise sequence. At the end of each trial, we probed a unique frequency channel to tone complex A (e.g., left panel) or complex B (e.g., right panel) or shared between the two complex tones (e.g., both panels). **(B)** Decoding performance for unique (left) and shared (right) probe-tones. Classifiers trained and tested separately at each time in a 600 ms time window of the probe-tone (-100ms to 500ms). The significant cluster is contoured with a dashed line. The classifiers could decode attention only when the probes are unique components ($p = 0.004$). There is a statistical difference between the unique and shared scores as depicted in Figure 3.15. **(C)** Comparison between the difference in responses to unique A and unique B probe-tones for two attentional conditions for the first DSS component (Figure 3.16; see section 3.2.7). There were significant differences for shaded areas (larger than 2σ), with reverse polarity for attend A and attend B, suggesting the opposite effect of attention on coherent and incoherent tones (i.e., enhancement and suppression). **(n=14)**.

performed the task above the chance level as demonstrated by the behavioral results in Figure 3.14.

As before, we trained classifiers to detect modulations on the probe tone responses during, pre, and post the probe tone onsets (at 0 ms). Results displayed in (Figure 3.5B) demonstrate that attention significantly modulates the probe-tone responses starting 130 ms following onset but only when aligned with unique tones of the complex-tone sequences (p=0.004; Figure 3.5B, left panel). The classifier failed to decode any modulations due to attention when the probe-tone aligned with a shared component (Figure 3.5B, right panel). Moreover, there was a statistical difference between the decoder scores of Unique and Shared probe tones, as illustrated in Figure 3.15.

We then analyzed the EEG responses to extract the most repeatable auditory component across trials by looking at the first DSS component (see section 3.2), which was averaged across all subjects (Figure 3.16A). Figure 3.5C illustrates the difference between the unique probes "$UniqueB - UniqueA$" under attend A (noise coherent with A; orange) and attend B (noise coherent with B; blue) conditions. For most of the time period encompassing the probe tone, the difference exhibited the opposite polarity in the two attention conditions, with significance near 70 ms, 160 ms, and 420 ms following the probe's onset. Furthermore, we looked at the average of the first DSS components over the time window of 60 ms to 200 ms. The average power is significantly larger when the probe-tone is a unique component of the attended stream ($p = 0.04$ for unique A and p=0.01 for unique B). However, there is no significant difference in response to the shared frequency channel ($p = 0.24$; Figure 3.16B).

To summarize, the key finding of this experiment is that attending to the noise-bursts, which are perceptually different from the tones and spectrally located at least 2.5 octaves apart, never-

theless caused the coherent complex-tone sequences to become modulated as if they became bound to the noise-bursts and included in the focus of attention. This is consistent with the earlier experiments' findings in the present study that coherent tones are modulated when subjects attended directly to the complexes. This we take as evidence of the perceptual binding of all coherent acoustic components to form a unified attended stream.

3.3.4 Experiment 4: Segregating Speech Mixtures

Real-world auditory scenes often consist of sound streams of unequal rates, many shared spectral components, and gradually changing parameters (pitch, location, or timbre). In all previous experiments, we have demonstrated that temporal coherence plays a crucial role in stream formation. However, all stimuli used were well-controlled, relatively simple tone-complexes and noise bursts with stationary parameters. Here, we extend the temporal coherence principle tests to a more naturalistic context using speech mixtures. In a speech, the signal is modulated in power during the succession of syllables, just like the tone and noise sequences used in the previous experiments, i.e., one can abstractly view a speech signal as a sequence of bursts separated by gaps of various duration. Each burst encodes features of the speaker's voice, such as his/her pitch, location, and timbre, which temporally fluctuate in power coherently. In a mixture of two different speakers, female (F) and male (M), saying different words, the sequences of bursts begin to resemble the alternating A & B complexes of our simpler stimuli. Consequently, the power in each speaker's features would fluctuate coherently, but they are both different and out-of-sync with those of the other speaker. Furthermore, simultaneous speech segregation can potentially rely on the incoherence between the power modulations of the two speech streams since speakers

utter different words, and hence their modulations are often de-synchronized [38].

To confirm these assertions, we first tested that the same approaches using probe-tones and trained classifiers can be readily applied to decoding speech responses. The probe-tones were restricted to the end of the sentences and were always harmonic complexes, as detailed below. In the second part of experiment 4, we refined the probe-tones to investigate the attentional modulations on single frequency components, and more importantly, to insert the probe anywhere in the midst of the speech mixture.

a) Probe at the end:

Single-speaker sentences were selected from the CRM corpus [68] and then mixed to produce two-speaker mixtures, each containing a male and a female voice. All sentences in this corpus have the same format, including color and a number (see section 3.2). During the task, subjects were instructed to attend to a specific speaker on each trial and then report the color and number uttered by this target speaker. The mixture in each trial ended with a 90 ms harmonic-complex tone as a probe, consisting of the 4 lowest harmonics aligned with corresponding frequency components of either the attended or unattended voice. Therefore, the probe-tone was uniquely aligned with one speaker, as were the single-tone probes in the earlier experiments (Figure 3.6A).

16 Participants were asked to report the color and the number mentioned in the target sentence to make sure that they were attending correctly to the target speaker, all the subjects were able to do the task with ease (average accuracy = %93; Figure 3.17). Meanwhile, we measured the neural responses to the probe-tone with EEG. The responses were compared under different attention conditions using the same linear classifiers described earlier, with decoder scores significantly above the chance level (Figure 3.6B). Additionally, we generalized the modulatory pattern of attention by using classifiers trained and tested at various times relative to the probe

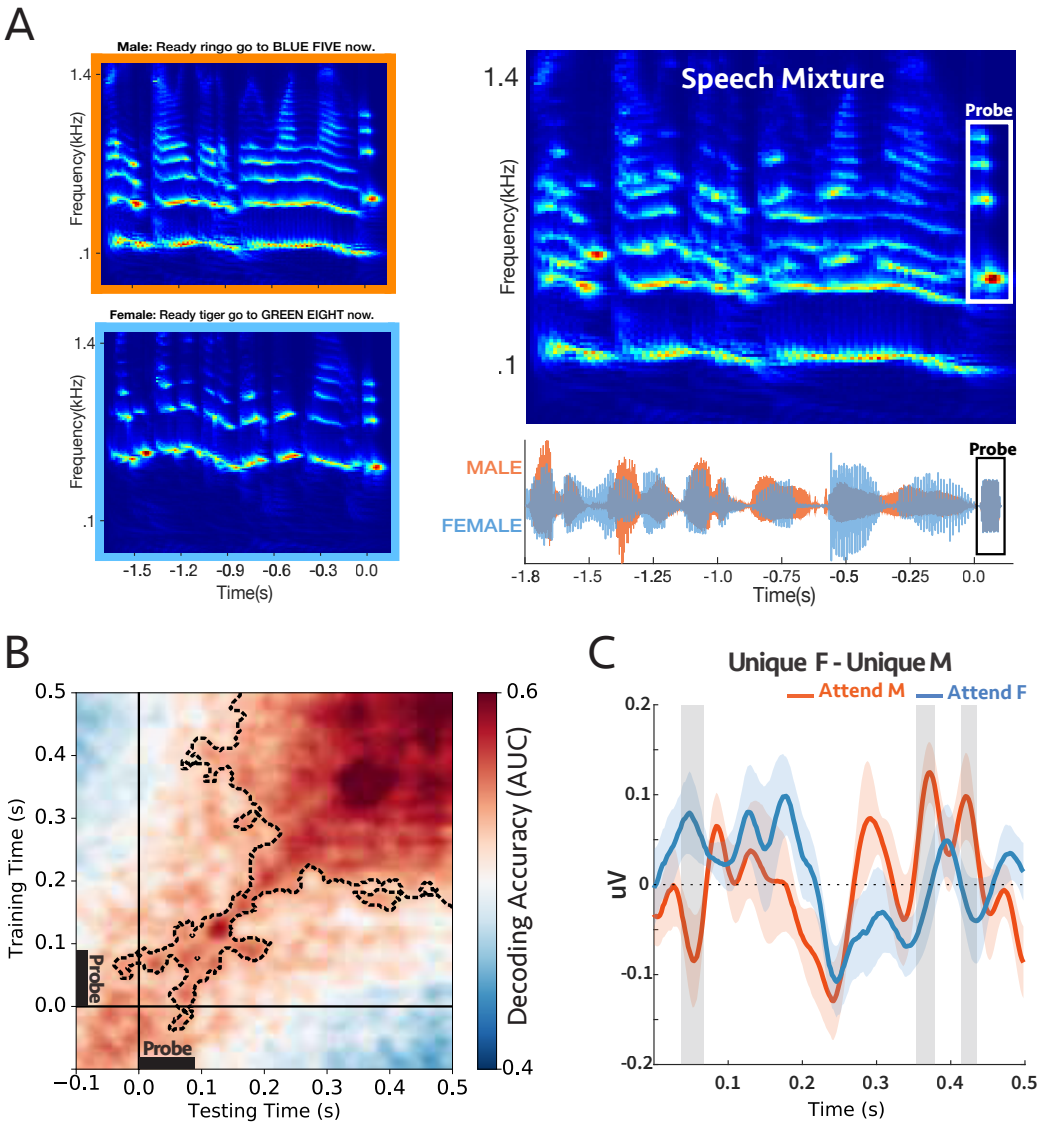

Figure 3.6: **(A)** A sample stimulus for experiment 4a. Top-left: cochlear spectrogram of the male voice. Bottom-left: cochlear spectrogram of the female voice. Note the probe-tone at the end. Right: cochlear spectrogram and acoustic waveform of the mixture. All the sentences had the same length (1.8 sec). The participants were instructed to report the color or the number by the target voice. During the task, the auditory scene consisted of two concurrent speech streams followed by a 90 ms probe-tone with complex harmonics. The probe-tones harmonic frequencies were aligned with the four loudest harmonic frequencies of either male or female voice at the end of the sentences; therefore, the probe-tone was either unique to male or unique to female. **(B)** Decoding performance for the probe-tones trained and tested during the probe time window (-100ms to 500ms). The significant clusters were contoured with a dashed line. The classifiers could decode attention significantly above chance ($p = 0.0002$) **(C)** Comparison between the difference in evoked responses to unique F and unique M probe-tones for two attentional conditions at "*Cz*" channel (placed on the center of the mid-line sagittal plane) (Figure 3.19; see 3.2.7). There were significant differences for shaded areas (larger than 2σ), suggesting the opposite effect of attention on coherent and incoherent tones.**(n=16)**.

onset, e.g., trained at the beginning of the speech mixture (-1.8 s to -1.2 s) and tested around the probes (-0.1 to 0.5s; Figure 3.18 left panel), trained near the probe (-0.1 s to 0.5 s) and tested at the onset of the speech mixture (-1.8 s to -1.2 s; Figure 3.18 right panel) In all cases above, the decoding scores were significantly above chance as is evident in the figures ($p = 0.0002$, $p = 0.009$, and $p = 0.009$ respectively). It is interesting to note that the decoder exhibited "ramping" dynamics [51] which may reflect the gradual buildup of the harmonic pattern responses to the probe complex which resembles the spectral characteristics of the preceding sequences. This ramping was not seen in the earlier experiments as the probe was only a single tone. Finally, we also contrasted the evoked responses to the probe-tones ($UniqueF - UniqueM$) at the Cz channel and observed significant differences between attending to male and female, with opposite polarity (Figure 3.6C). Therefore, the results thus far indicate that: 1) The neural response to the harmonic frequency components aligned to the pitch of the two speakers is reliably modulated by attention. 2) The pattern of brain activity at the onset of the attended/unattended speech sentence is similar to the activity during the probe tone at the end, and consequently, the trained decoders were generalizable for these two-time windows even when separated by a sizable interval. These results are consistent with the temporal coherence hypothesis because if attention to the pitch of one voice enhances the pitch signal, it will enhance its harmonics (all being coherent with it; [38]) and will relatively suppress the harmonics of the unattended speaker, which are incoherent with it.

b) Probe in the middle:

Using the same speech corpus as stimuli, this experiment probed the modulations of a single frequency channel potentially anywhere within the duration of the speech mixture. The probe frequencies in these experiments were chosen centered at the 2^{nd} harmonic of the female or at the

3rd harmonic of the male, unique components in the midst of the speech mixtures as illustrated in Figure 3.7A. Participants were instructed to report the color or number spoken by the talker who uttered the target call-sign ("Ringo"; see section 3.2), all participants did the task successfully (average accuracy = %79; Figure 3.20). On average, the onset of the call-sign occurred 300 ms (25 ms) following sentence onset, and the probe-tone was inserted 600 ms after speech onset.

We trained linear classifiers to ascertain the modulation that attention induced in the probe responses during the time window (-200 ms to +400 ms with probe onset defined as 0). It is evident in Figure 3.7B) that the decoding scores were significantly above the chance level even prior to the onset of the probe tone, reflecting the emergence (or streaming) of the target speaker spectral components. The significant decoding continued for up to 280 ms after the probe onset, with a peak at 150 ms. We also extracted the evoked EEG signal at channel Cz due to the probe-tones to determine the direction of the modulation induced by attention Figure 3.21. Figure 3.7C shows the difference in response to unique F and unique M was significantly and rapidly modulated by attention within about 250 ms, consistent with previous findings in ECoG recordings [69].

Therefore, in conclusion, we measured significant attentional modulations of the probe-tone responses that are frequency-specific (distinguishing between alignments with closely spaced male and female harmonics). These findings indicate that during speech segregation, the components of the attended speaker are differentiated from those of the unattended source quite rapidly, or specifically, as soon as 250 ms after the onset of the target callsign. This delay is commensurate with that observed in analogous ECoG experiments involving switching attention between 2 speakers.

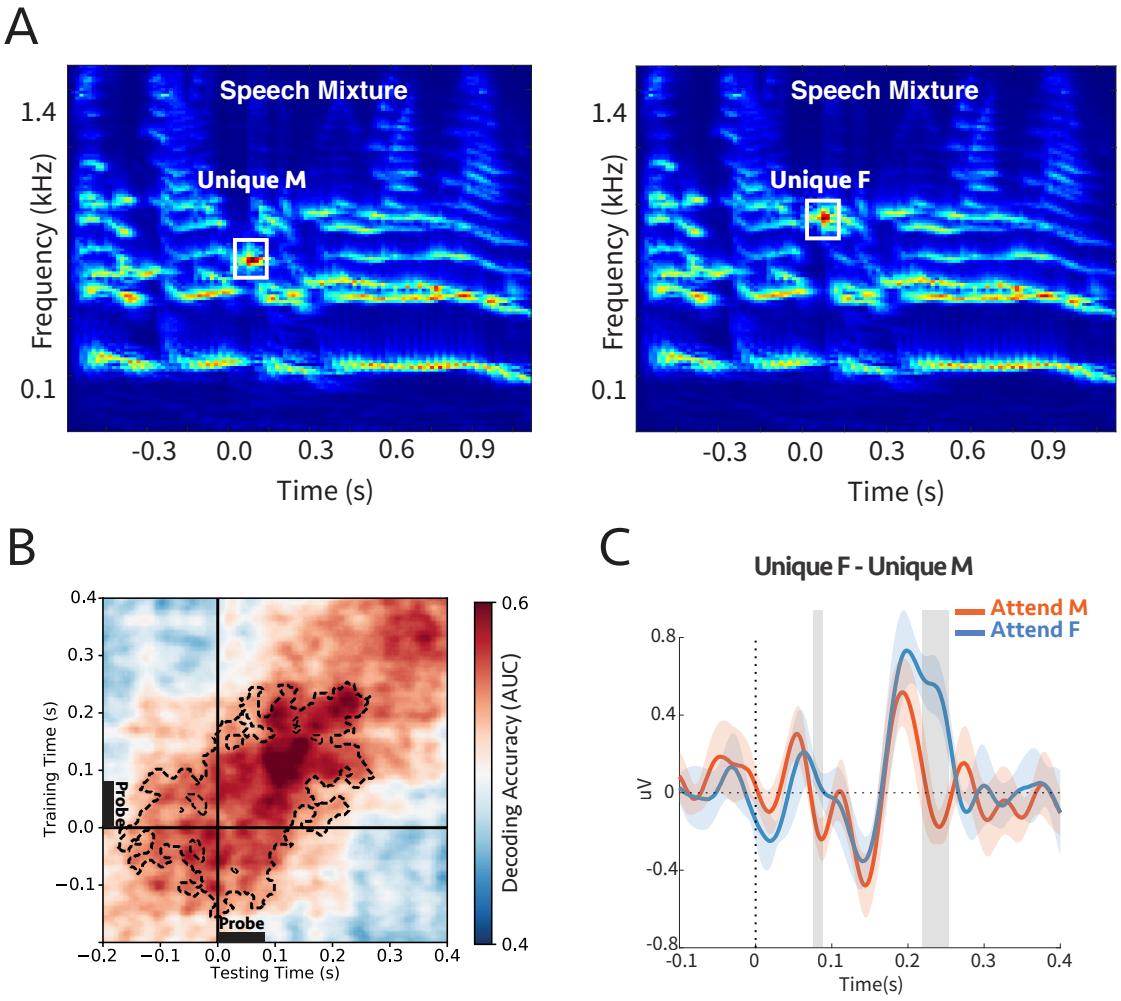

Figure 3.7: **(A)** Cochlear spectrogram of a sample stimulus mixture for experiment 4b. It consisted of two (male and female) voices. The participants were instructed to report the color or the number of the speaker who uttered the target call-sign. The probe's frequency was aligned with the 2^{nd} or 3^{rd} harmonic of the female or male, respectively. Therefore, the probe-tone was either unique to male (unique M) or unique to female (unique F). **(B)** Decoding performance for the probe-tones. Classifiers trained and tested separately at each time in a 600 ms time window of the probe-tone (-200ms to 400ms). Cluster-corrected significance was contoured with a dashed line. The classifiers could decode attention significantly above chance for up to 280 ms after the probe-tone onset ($p = 0.015$). **(C)** Comparison between the difference in evoked responses to unique F and unique M probe-tones for two attentional conditions at *Cz* channel (placed on the center of the mid-line sagittal plane) (Figure 3.21; see section 3.2.7). There are significant differences for shaded areas (larger than), suggesting the opposite effect of attention on coherent and incoherent tones. **(n=7)**.

3.4 Discussion

This chapter explored the dynamics and role of temporal coherence and attention in the binding of features that belong to one auditory source and its segregation from background sources in a cluttered auditory scene. The temporal coherence hypothesis predicts that acoustic features that induce coherently-modulated neural responses should bind together perceptually to form a single stream. One piece of evidence for this perceptual stream formation is taken to be the physiological enhancement of all the coherent responses. It is also postulated that an essential component of this process is attention directed to one or more of the coherent sets of acoustic features, which then initiate mutual interactions, and hence binding of all coherent features. Previous studies have shown that responses to an attended (pure-tone) stream are enhanced relative to the remaining background in a mixture [70, 71]. However, it has remained unclear whether the neural responses to the individual constituents of a complex stream are also similarly modulated by attention to only one of its elements and how this is related to its perceptual formation.

In the series of EEG experiments reported here, we demonstrated that when listeners attended to one attribute of a complex sound sequence, other temporally coherent responses were similarly modulated; incoherent responses were relatively suppressed (or *oppositely* modulated) while leaving shared elements unchanged. This was found to be true over a wide range of attributes, be it the pitch of a sequence of harmonic complexes (Experiment 1), the timbre of inharmonic complexes (Experiment 2), noise burst sequence (Experiment 3), or the call-sign by a single speaker in a mixture (Experiment 4). Of crucial importance, this was the case even for those features in the sound sequence that were not directly accessible to the listener. For example, when subjects attended to the pitch of a harmonic complex or the timbre of an inharmonic

complex, they rarely reported being able to (or spontaneously) listen to the individual constituent tones, yet these components became modulated as if they were directly attended to. In fact, in experiment 3, subjects completely ignored the accompanying complex tones while attending only to the noise bursts, yet response modulations still occurred for the unique coherent tones, i.e., they acted as part of the foreground noise stream.

To access the responses of the individual components of a stream (despite the poor spatial resolution of the EEG), we investigated the responses to probe tones and probe complexes that relied on the persistent effects of attention in the midst or just following the end of the streams when there were no interfering signals from other sounds. The effects of attention on the probe responses, however, were not always easy to interpret because the array of the 64 EEG electrodes pick up complex mixed signals deriving from many regions of the brain. Thus, specifying and interpreting an EEG response to use for the measurement requires combining (and not simply averaging) the recordings from all these electrodes. Therefore, the term ”response enhancements” used in our original hypothesis does not always literally mean an increased response amplitude or power, but rather a response-modification that is robust and repeatable when attentional conditions are identically manipulated. While these changes are often detected as enhancements in the power of the response DSS component (particularly when using simple pure-tone streams instead of the complex multi-component streams here; [59, 60]), we focused instead on a more flexible measurement approach that detects these changes through linear estimators. Specifically, a set of classifiers were trained to decode the attended/unattended responses near the probe-tone time window and were then tested at other times (such as generalizing the estimators to the speech beginning) to demonstrate that the response patterns induced by attention persisted during the probe. The patterns of generalized decoders (panel B in Figures 3.3 - 3.7) may reflect the un-

derlying brain processing and the dynamic of the responses to the probe tones. For instance, as described in King and Dehaene 2014 [51], the squared pattern of the temporal generalization matrices in experiments 1, 2, and 3 (Figures 3.3B, 3.4B, 3.5B) suggests a sustained brain process encompassing the probe tone. Moreover, the patterns in Figure 3.7B showed a temporally jittered activity due to the subtle variation in the probe tone onset relative to the call-sign timing. Finally, Figure 3.6B showed a more complex and slowly increasing activity. Clearly, furthermore targeted experiments are needed to investigate specifically the origin of the different patterns of dynamics of the temporal coherence process, and the details of the response buildup during auditory streaming.

It should be noted that the responses to the probes placed at the end of the sequences do not simply reflect some kind of an overall attentional change in the offset responses of the preceding complexes, because if they did, then they would have behaved similarly under attentional switching regardless of whether they were aligned to unique or shared frequency channels. Instead, probe responses are in fact sensitive to whether the frequency channel to which they are aligned is unique or shared, even under the same overall attentional conditions that would have left the offset responses of the complexes identical.

To summarize, the overall findings from this study are consistent with the temporal coherence hypothesis where correlated responses become bound as a single stream that integrates the elements of the sequences regardless of: (1) the temporal regularity of the sequences, e.g., uniform or irregular (Figure 3.3 and Figure 3.4) (2) stimulus types across the sequences, e.g., noise or tones-complexes (Figure 3.5); (3) whether the tones are harmonic or inharmonic complexes (Figures 3.3, 3.4, and 3.5); (4) whether the sequences are spectrally near or far apart (Figures 3.3, 3.4, 3.5, and 3.6); and crucially, (5) whether the sequence parameters (e.g., pitch and timbre) are

stationary or dynamically slowly evolving as in speech (Figure 3.6 and Figure 3.7).

Temporal coherence is essentially an associative process likely enabled by rapidly formed and modulated connectivity among coherently responsive neurons. This process is analogous to the well-known Hebb's rule of "fire together, wire together", except that it occurs at a much faster pace (within hundreds of milliseconds, as evidenced by the rapid build-up following the call-signs in Figure 3.7). It is also promoted and controlled by "attention", a notion that is difficult to define precisely. However, experiments in animals and human subjects have demonstrated that without the engagement and attentional focus on the task, or selective attention to specific features of the stimuli, these rapid modulations of connectivity which are manifested as perceptual binding, and hence stream formation, become far weaker or absent [28, 35]. The underlying biological foundations of this process remain largely unknown but are currently the target of numerous ongoing studies.

We end by observing that the concept of temporal coherence likely applies in a similar way in other sensory modalities such as vision. In a dynamic visual scene, features of a visual object, such as its pixels, move together coherently in the same direction and speed, inducing highly correlated neural responses. Conversely, pixels of independent objects move with different relative motion and can thus be segregated easily from those of other objects based on this idea (Lee and Blake 1999). Also, multi-modal integration, such as enhanced comprehension of speech in an audio-visual scenario (lip-reading), may well be explained by temporal coherence, i.e., the temporal coincidence between the visual representation of the lip motion and the acoustic features of the syllables can strongly bind these sensory features, and hence improve the intelligibility of speech [72–74].

3.5 Supplementary Figures

This section contains supplementary results for all the experiments described in section 3.3. We included results for behavioral performances of the subjects participated in all experiments, as well as other supporting data for the main EEG results. These supporting materials are not essential for inclusion in the full text of the previous sections, but would nevertheless benefit the reader.

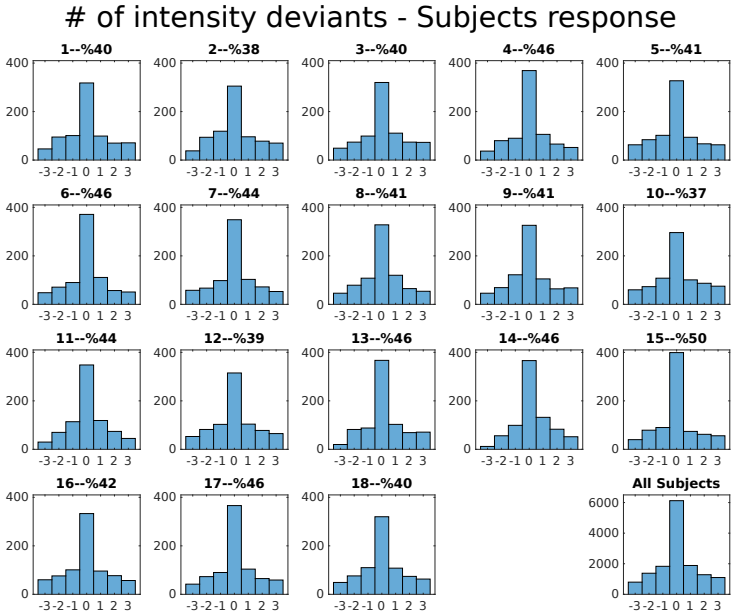

Figure 3.8: *Behavioral results for 3.3.1:* In this experiment, listeners were instructed to count the number of deviants in the target (attended) sequence, which was uniformly distributed between 0-3 (four choices) across trials, and hence, the chance level was at %25. Each subplot shows the histogram of the true number of deviants minus the subject's response. Therefore, in these subplots, "0" means the correct response (hit), positive numbers mean that listeners missed one or some of the deviants, and negatives mean response was larger than the actual number of deviants. Each subplot's title includes the subject's number followed by their percentage of correct answers (hit rate). All the subjects performed above the chance level.

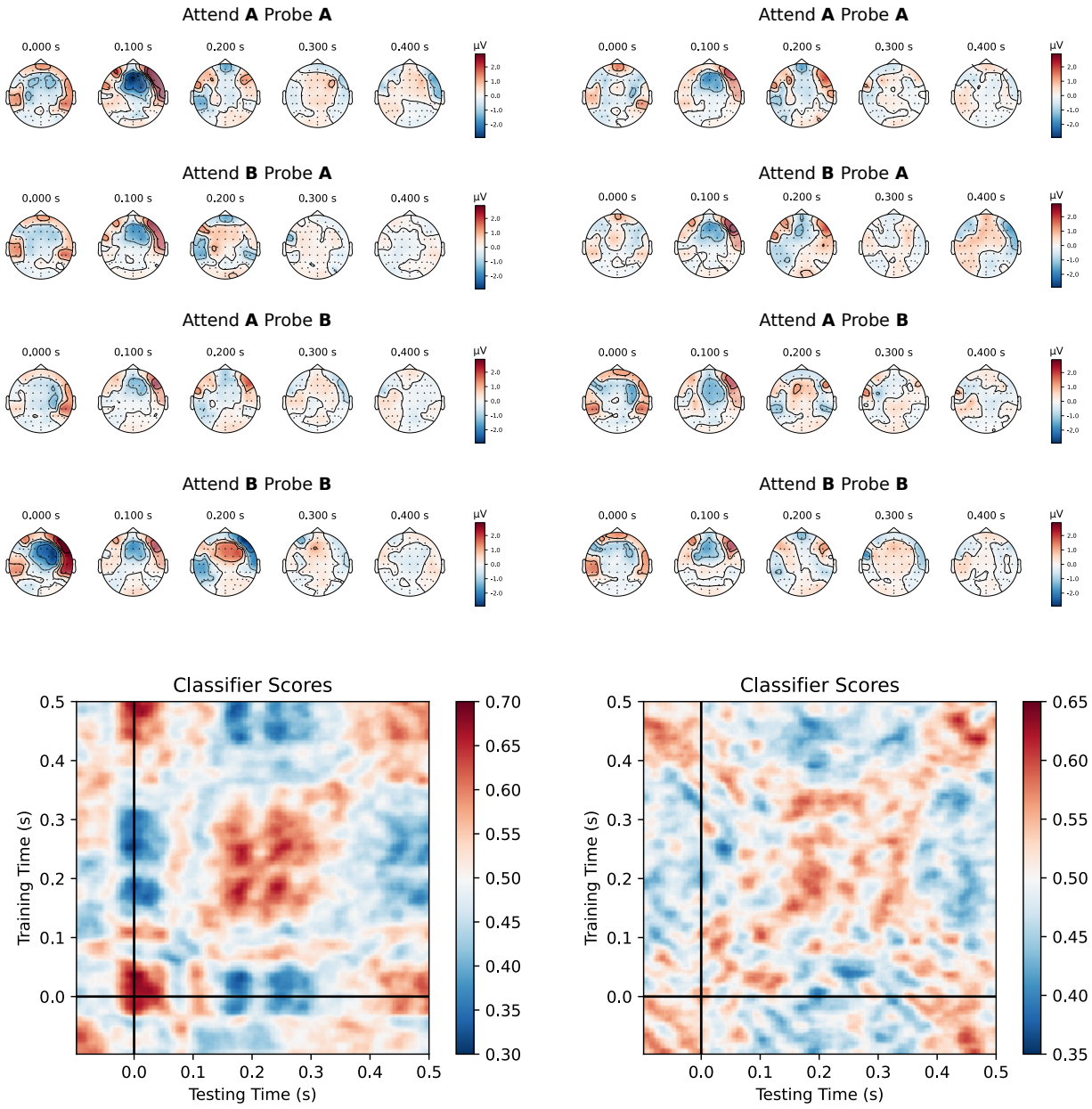

Figure 3.9: *The relation between the classifier scores and EEG topomaps (a subject example in 3.3.1).* Topomaps of probe A and probe B is plotted for different attentional conditions for **(A)** *Expected* and **(B)** *Unexpected* cases. Linear classifiers were trained at each time point on the responses from all 64 channels (topomaps) in order to decode the focus of attention. At the subject level, the trained classifier could capture the differences in the topomap patterns caused by the attentional changes, e.g., the differences between the topomaps of *Attend A probe A* and *Attend B Probe A*. The classifier scores showed the robustness of the effect for a given subject across all trials; in the second-level test (depicted in Figure 3.3B), we showed the robustness of the effect sizes across all subjects.

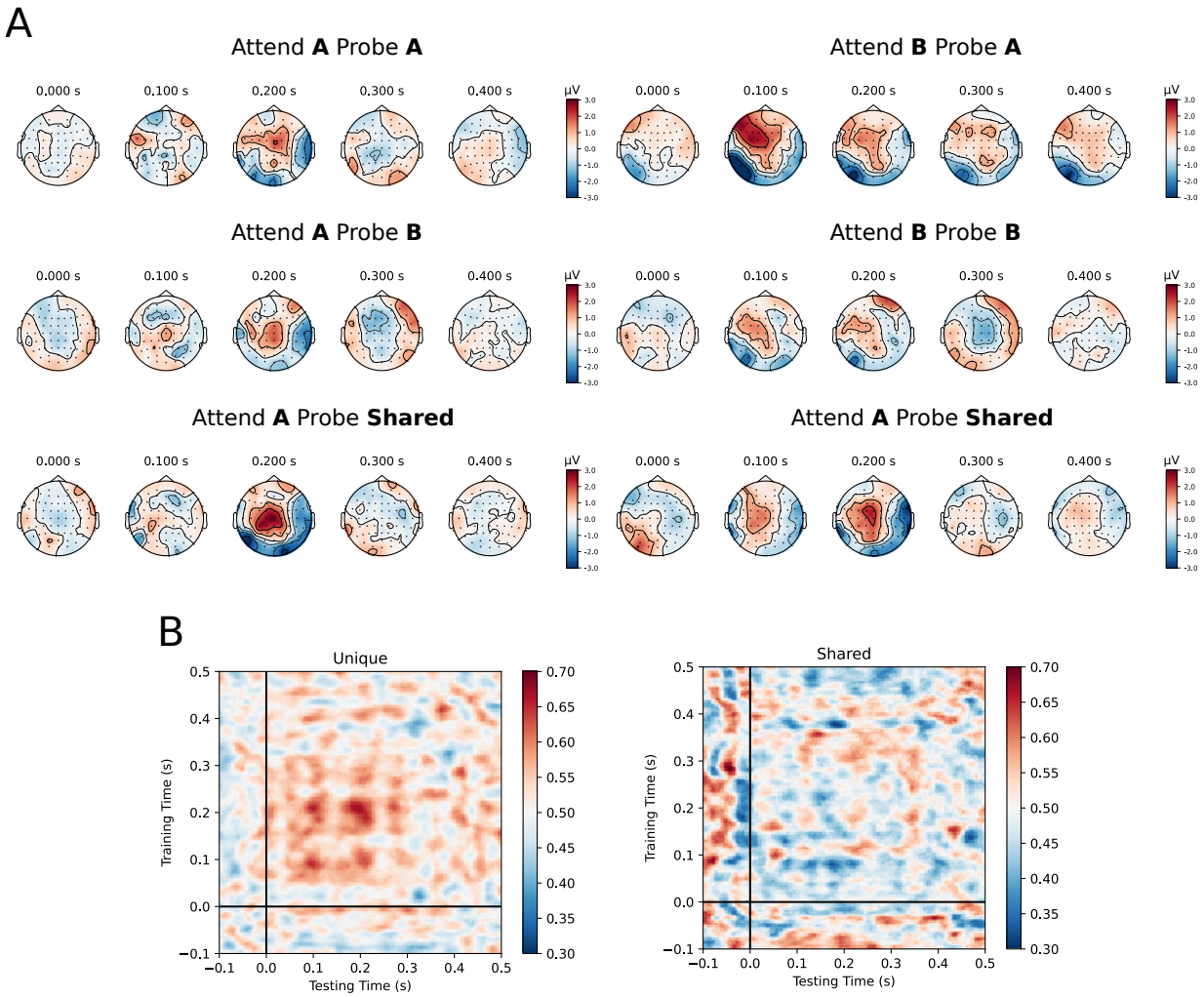

Figure 3.10: *The relation between the classifier scores and EEG topomaps (one subject example in 3.3.2).* **(A)** Average topomaps of probe A, probe B, and the shared probe were plotted for different attentional conditions. Linear classifiers were trained at each time on the signals from all 64 channels (topomaps) in order to decode the focus of attention. At the subject level, the trained classifier tried to capture the differences in the topomap patterns caused by the attention. **(B)** The classifier scores demonstrated the robustness of the effect for the given subject across all trials. For the unique probe (left), the performance of the classifiers was above chance, which means that there was a consistent difference, i.e., a difference between *"Attend A Probe A"* and *"Attend B Probe A"* in topomap patterns across all trials. Conversely, the shared probe scores suggest that the difference between attentional conditions was not robust since it was not linearly separable.

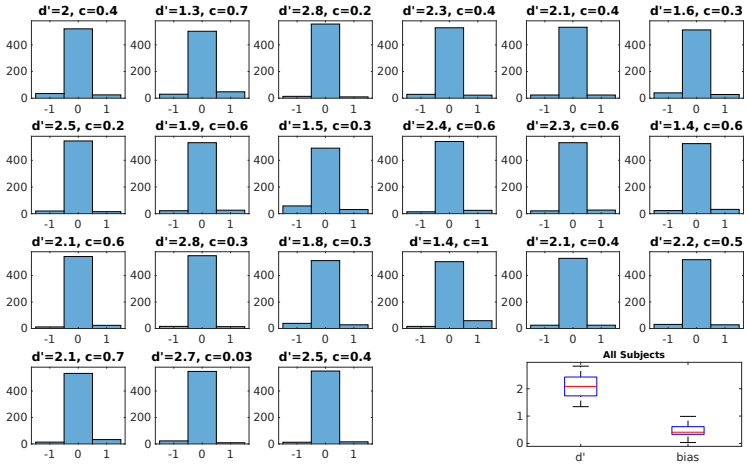

Figure 3.11: *Behavioral results for 3.3.2:* In this experiment, listeners were instructed to detect a deviant in the target (attended) sequence. Each subplot shows the histogram of a deviant's presence (0 or 1) minus the subject's response. Therefore, in these subplots, "0" means the correct response (hit), +1 means listeners missed the deviant, and -1 reflects the false alarms. The title of each subplot includes the subject's d'prime followed by their bias.

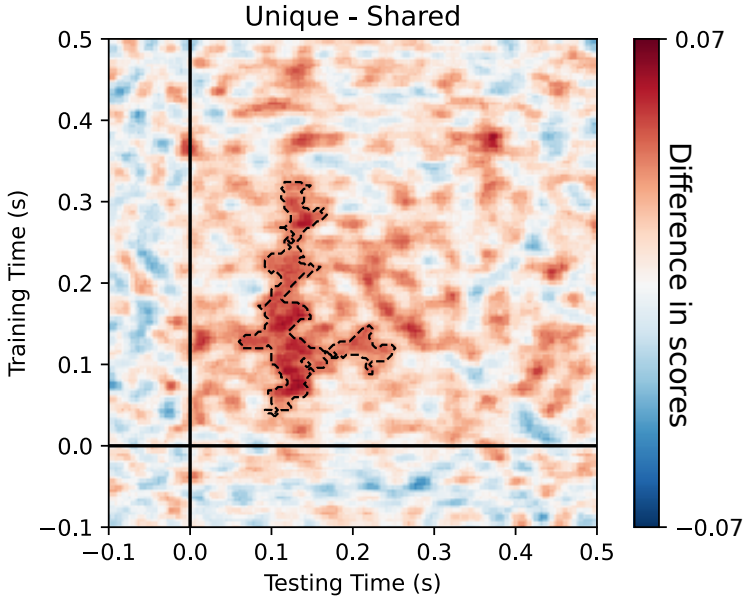

Figure 3.12: *Comparison between unique and shared decoder scores for 3.3.2.* Scores show the average difference between decoding scores of Unique and Shared probe-tone for all subjects. The difference is significant for the time region contoured between the dashed lines ($p = 0.009$).

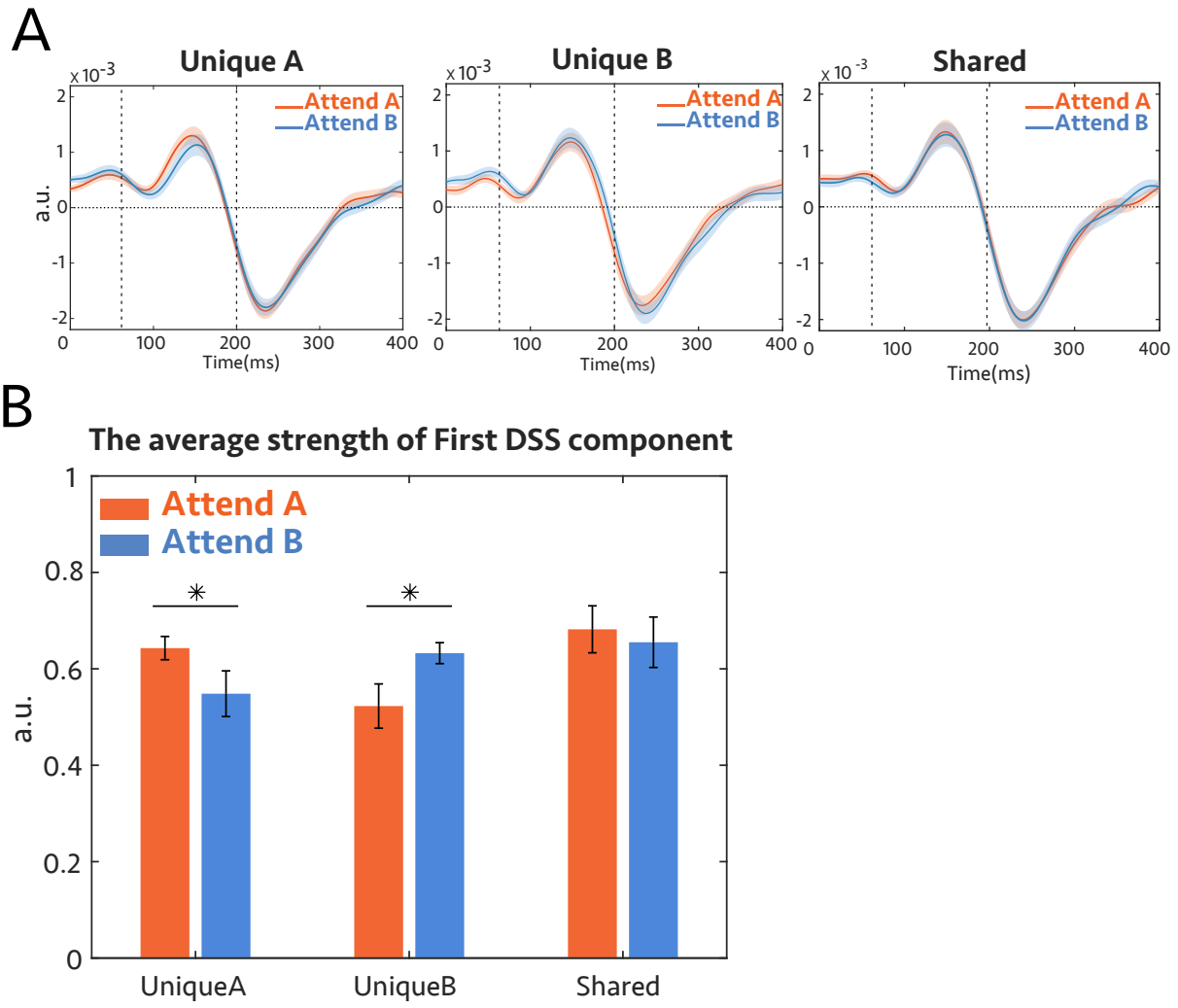

Figure 3.13: *DSS evoked response for 3.3.2.* Data were submitted to DSS analysis (see section 3.2.7) using the average across trials as a bias filter. The aim was to isolate the most repeatable auditory component by applying a spatial filter. **(A)** Grand average of the most repeatable EEG response to the probe-tone extracted by DSS for each subject; onset of the probe tone is at 0. Left: The response when the probe is at the frequency unique to complex A, middle: when the probe tone is a unique component of complex B, right: The response when the probe tone is a shared component, for attention to tone complex A (orange), and attention to tone complex B (blue). In Figure 3.4, we subtracted the orange and blue curves in unique A from the same colors in unique B. **(B)** The *average* amplitude of the neural response from 60 ms to 200 ms after the probe-tone onset. For the unique frequency channels, the attended condition has significantly higher power than the unattended condition ($p = 0.03$ for unique A and $p = 0.01$ for unique B), while the *average* of the shared channel does not show any modulation with attention ($p = 0.6$).

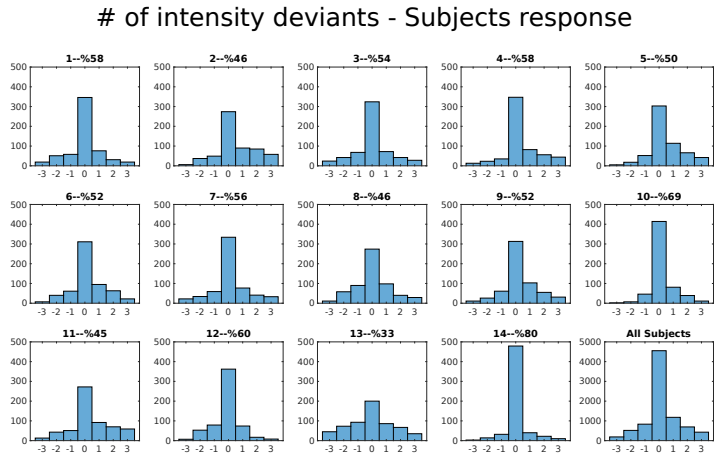

Figure 3.14: *Behavioral Results for 3.3.3:* In this experiment, listeners were instructed to count the number of deviants in the target (attended) noise sequence, which was uniformly distributed between 0-3 (four choices) across trials, and hence, the chance level was at %25. Each subplot shows the histogram of the true number of deviants minus the subject's response. Therefore, in these subplots, "0" means the correct response (hit), positive numbers mean that listeners missed one or some of the deviants, and negatives mean response was larger than the actual number of deviants. Each subplot's title includes the subject's number followed by their percentage of correct answers (hit rate). All the subjects performed above the chance level (chance level = %25).

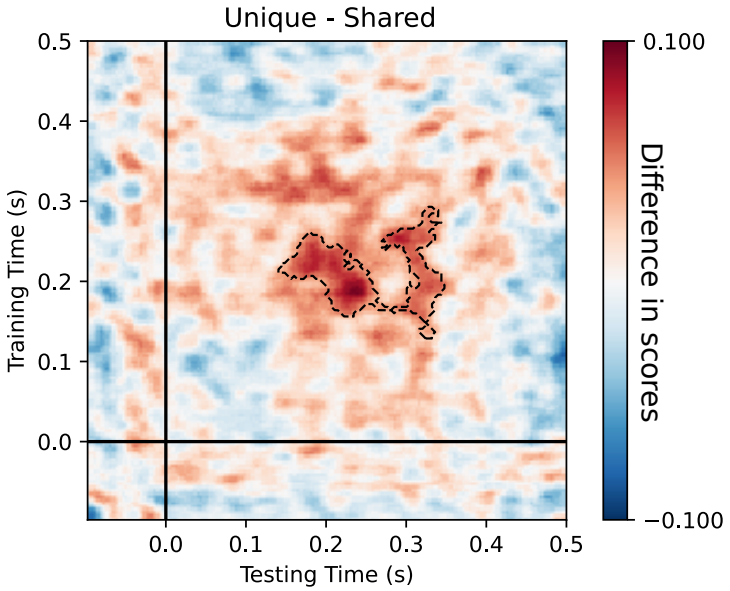

Figure 3.15: *Comparison between unique and shared decoder scores for 3.3.3.* Scores show the average difference between decoding scores of Unique and Shared probe-tone for all subjects. The difference was significant for the time region contoured between the dashed lines ($p = 0.004$).

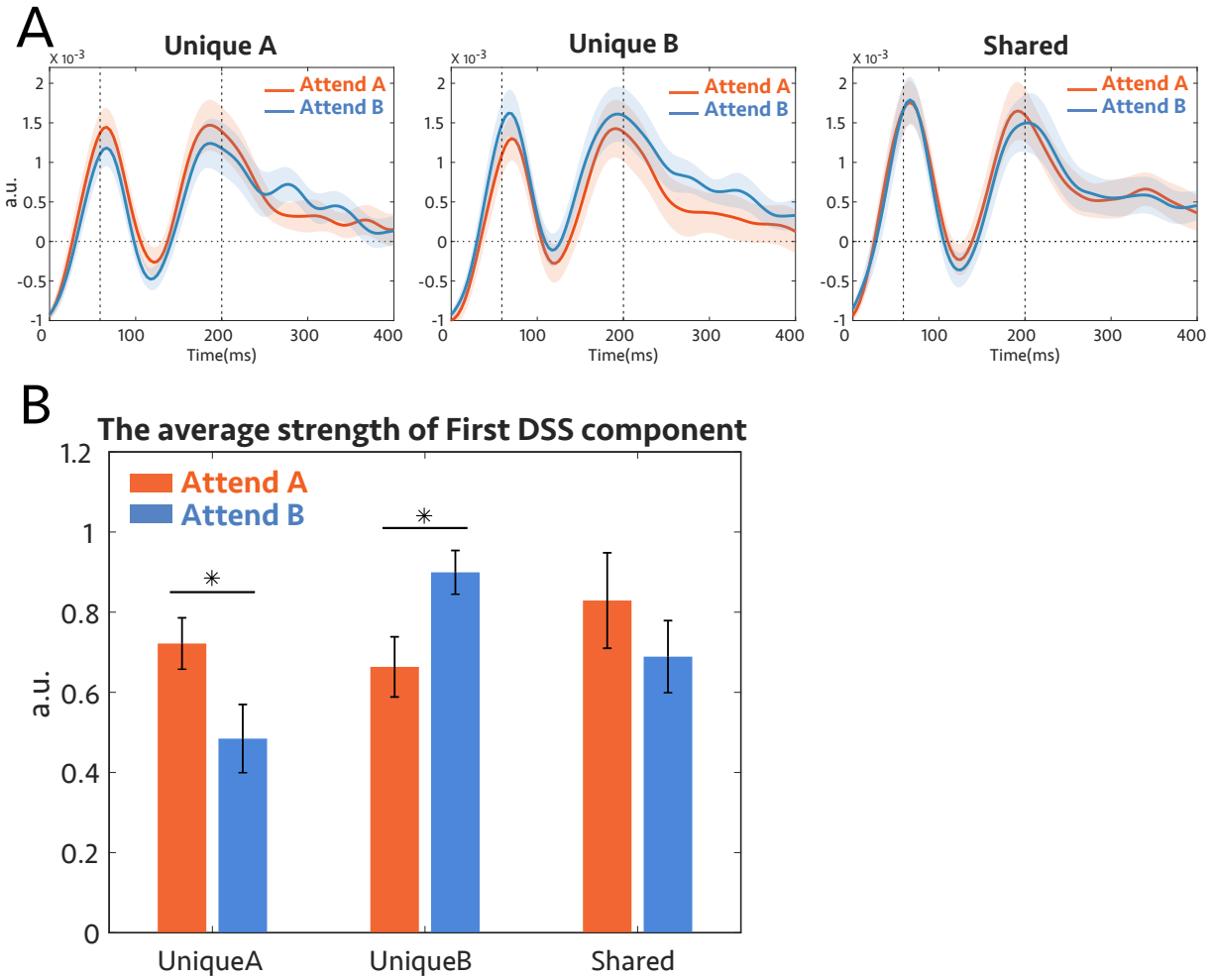

Figure 3.16: *DSS evoked response for 3.3.3.* Data were submitted to the DSS using the average across trials as a bias filter. The aim was to isolate the most repeatable auditory component by applying a spatial filter. **(A)** Grand Average of the most repeatable EEG response to the probe-tone extracted by DSS for each subject; onset of the probe tone is at 0. Left: The response when the probe was centered at the unique A frequency channels. Middle: The response to the probe-tone unique to complex B. Right: The response when the probe tone was a shared component, under attend to tone complex A (orange) and attend to tone complex B (blue), the curves are comparable. In Figure 3.5C of the main text, we subtracted the orange and blue curves in unique A from the same colors in unique B. **(B)** The average strength of the neural response of the first DSS component from 60 ms to 200 ms after the probe-tone onset. For the unique frequency channels, the attended condition had significantly higher power than the unattended condition ($p = 0.04$ for unique A and $p = 0.01$ for unique B), while the *mean* of the shared channel did not show any modulation with attention ($p = 0.24$).

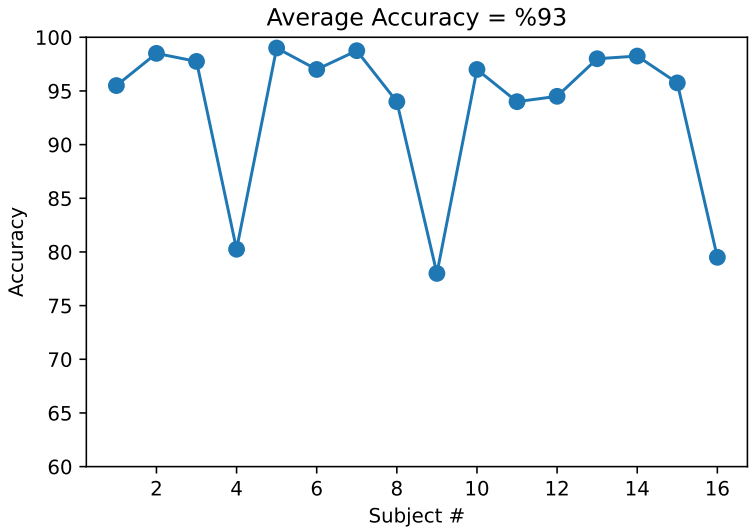

Figure 3.17: *Behavioral Results for 3.3.4a:* In this experiment, listeners were instructed to report the number or the color of the attended speaker. Each point shows the accuracy for each subject.

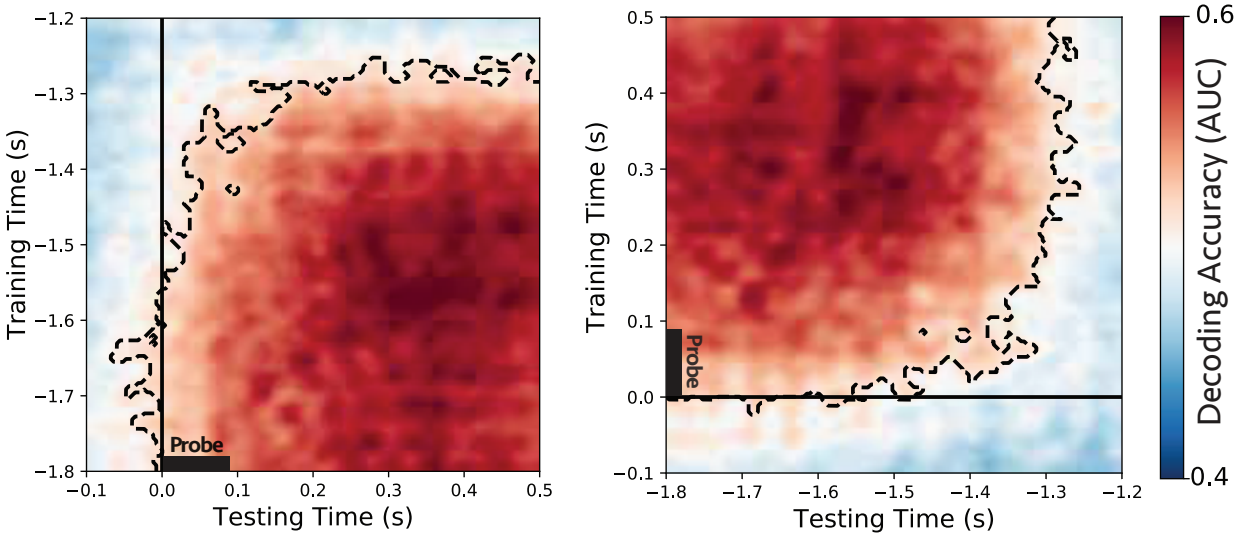

Figure 3.18: *Generalizing decoders across time.* Classifiers trained and tested separately at each time instant in two other 600 ms time windows. *Left:* trained at the beginning of the speech (-1.8 sec to -1.2 sec) and tested during the probe-tone (-100ms to 500ms). *Middle:* trained during the probe time window (-0.1 sec to 0.5 sec) and tested during the beginning of the speech (-1.8 sec to -1.2 sec). These results suggest that the modulatory effect of attention is generalizable across times during speech and probe-tone.

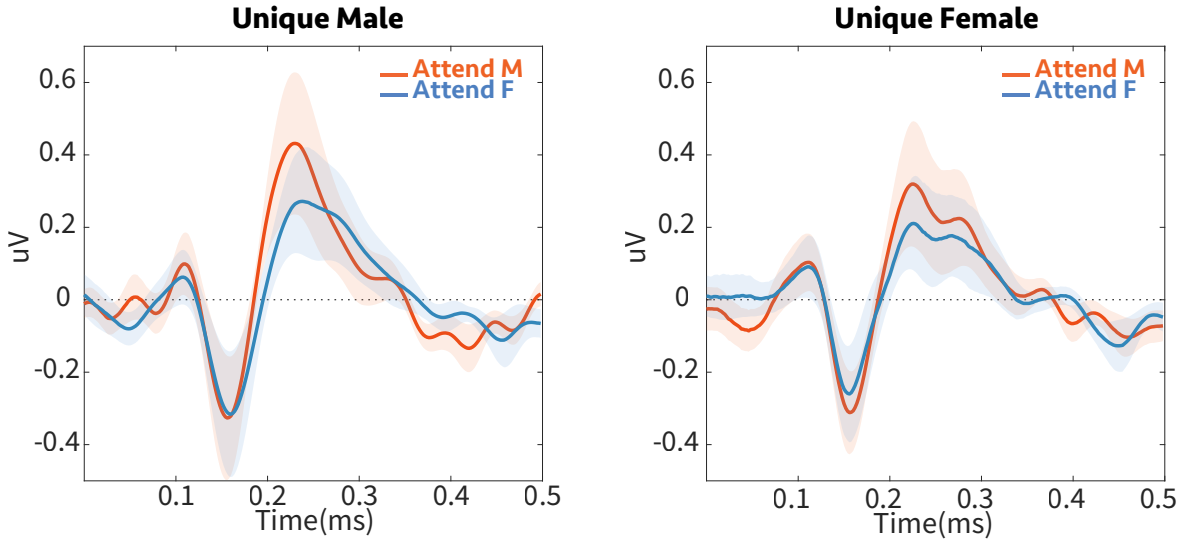

Figure 3.19: *Evoked responses at channel Cz for 3.3.4a.* Average EEG Evoked response to the probe tone at channel 'Cz', computed after denoising data and projecting back the first 5 DSS components to sensor space (see 3.2.7). The onset of the probe is at time 0.- The difference between *unique female* and *unique male* probe tone was computed for the orange (attend male) and blue (attend female) curves for each subject, and it was depicted in Figure 3.6C of the main text.

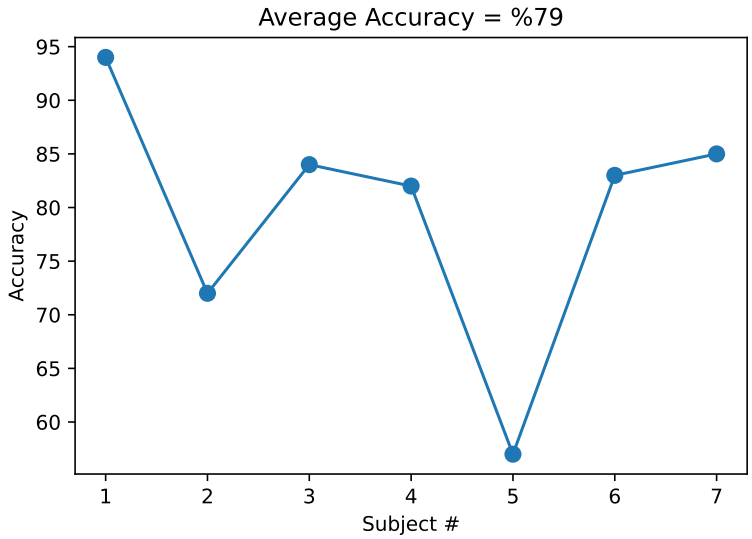

Figure 3.20: *Behavioral Results for 3.3.4b:* In this experiment, listeners were instructed to report the number or the color of the speaker who uttered the callsign. Each point shows the accuracy for each subject.

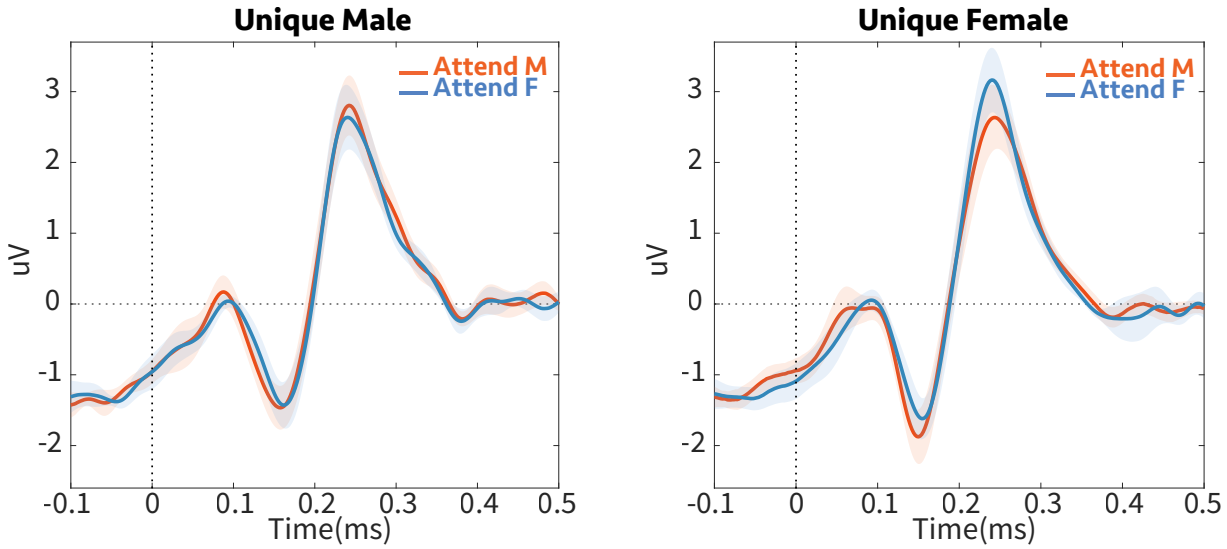

Figure 3.21: *Evoked responses at channel Cz for 3.3.4b.* Average EEG Evoked response to the probe tone at channel 'Cz', computed after denoising data and projecting back the first 5 DSS components to sensor space (see 3.2.7). The onset of the probe is at time 0. The difference between *unique female* and *unique male* probe tone was computed for the orange (attend male) and blue (attend female) curves for each subject, and it was depicted in Figure 3.7C of the main text.

Chapter 4: **Cortical Encoding of Statistical Learning in the Context of Musical Scales**

4.1 Introduction

To face the challenge of being immersed in constantly changing environments, human brains encode the statistics of everyday stimuli through mere passive exposure and implicit learning. Implicit statistical learning involves the extraction of structure from input and is generally thought to occur incidentally, through exposure to positive examples, without instruction, and without intention to learn [75–77]. Music offers a great opportunity to investigate statistical learning in an ecologically-valid setting, as converging evidence suggests that implicit learning underlies music acquisition for musicians and non-musicians. However, the question of which specific musical features support this type of learning while passively listening to music remains unexplored.

Music is produced and enjoyed in every human culture and displays incredible variability in its structure [78–80]. Yet, a few recurrent features suggest that cognitive factors constrain certain structural aspects of music [81]. Among them, musical scales (i.e., the collection of discrete pitches or notes used to compose melodies of a given musical style or culture) were found to exhibit common structural features across the vast majority of musical traditions. Almost

all scales in the world display non-uniform intervallic structure: the collection of pitches are separated by intervals of different sizes [82–84]. Recent behavioral data provided evidence that specific scale structures enhance the learning effects [85]. In particular, non-uniform scales were shown to enhance the learning of grammar (i.e., statistical relations) within the context of familiar (12-TET) and unfamiliar (14-TET and 16-TET) musical systems. The benefit originates from an enhanced internal representation of the tonal space carried by the scale, in which relations between tones are better defined in a non-uniform scale. The present chapter reports findings that reveal the neural bases of the facilitatory effects of non-uniform scales.

In this chapter, we provide evidence that facilitatory effects of statistical learning are supported by non-uniform scales, which may explain their prevalence across musical systems [85]. Melodies are generated using a fixed artificial grammar based on a first-order Markov chain and derived from either a uniform or non-uniform scale. After an exposure phase, listeners had to report incorrect transitions embedded in a new set of test melodies while the EEG signal was recorded. Behavioral results demonstrated enhanced performance in processing syntactic regularities in melodies associated with the non-uniform scales. EEG recordings mirrored these findings: Evoked Response Potentials (ERPs) associated with distinct probabilities of note transitions suggested that neural responses better tracked the melodies derived from the asymmetric scale. In addition, a set of linear decoders trained on classifying neural processing of grammatical structure exhibited enhanced accuracy in the context of the non-uniform scale. These findings strengthen the hypothesis that the ubiquity of non-uniform scales across cultures results from cognitive constraints that facilitate learning syntactic regularities in melodies.

4.2 Experimental Design and Procedure

4.2.1 Participants

Sixteen adult participants with self-reported normal hearing participated in this study, conducted at the University of Maryland. One participant was removed from the analysis for not keeping the earphones in place during the experimental procedure. Among the fifteen remaining participants (9 females, mean age = 25 years, $SD = 8$), two had five or more years of formal musical training, and all were still engaged in daily musical practice. All participants were given course credits or monetary compensation for their participation. The experimental procedures were approved by the University of Maryland Institutional Review boards. Written, informed consent was obtained from each subject before the experiment.

4.2.1.1 Musical Scales

Participants were presented with melodies generated from hexatonic scales in the two following structure conditions, uniform and non-uniform. Each scale was composed of six tones in 12-TET. The two scale conditions were obtained by positioning tones in the 12-TET in a manner that conformed to the different intervallic structural properties, as illustrated in Figure 4.1A. The uniform scale was composed of intervals (i.e., space between the pitch of subsequent tones) of equal sizes. In contrast, the non-uniform scale was composed of intervals of different sizes so that each tone had a unique set of intervallic relations with all other tones when moving from one tone to another in the same direction (clockwise/counter-clockwise or up/down) across the octave span.

4.2.1.2 Grammar

Melodies were composed of the tones within a given scale, and their construction was determined by a first-order Markov chain inspired by Rohrmeier et al. [86]. Since two scale types were used (hexatonic, 12-TET), two different grammars were used, each including all the tones in the scales. However, the complexity of both grammars was kept constant concerning the transition probabilities that were used to generate the melodies. Schematic grammar representations are shown in Figure 4.1B. Each node corresponds to a tone in the scale. Notably, the correspondence between nodes and notes was randomized for each participant and for each structure condition. Arrows connecting nodes determine the permissible transitions between notes, along with the probability of transition. The "reference" version of the grammar was determined for each listener prior to the exposure phase. The "alternative" version of the grammar was obtained by switching nodes 3-4 and 5-6, which introduced ten possible wrong transitions. Melodies generated with the alternative grammar contained a set of three transitions between tones that were never part of the melodies generated with the reference grammar.

4.2.1.3 Melodies

All melodies were composed of 500 ms sine tones to which a tapered-cosine (Tukey) window was applied. Tones were not separated by a silence interval. During the exposure phase, 100 melodies were generated in real-time using the grammar structure and the pitches of tones defined by each scale. During the exposure phase, melodies were produced using the current structure condition (uniform or non-uniform) and the reference version of the grammar. During the test phase, half of the melodies were produced the same way using the reference grammar, and half

of the melodies were produced the same way but using the alternative grammar. 40 reference and 40 alternative melodies were presented in random order during the test phase. All melodies were constrained so that they did not exceed 15 tones and had to reach the final note, as defined by the grammar.

4.2.2 Procedure

The experiment was divided into two parts, each corresponding to a structure condition in which the order of testing was randomized across participants. During each part, listeners had to first complete an exposure phase during which they listened to 100 melodies. During this phase, melodies with the designated scale and grammar were generated in real-time. Only the correct grammar version was used to generate the exposure melodies. Throughout this phase, listeners had to simply click a mouse to play the next melody. Immediately following the exposure phase, participants completed a test phase during which 80 melodies were generated on the fly; half of them were generated using the reference version of the grammar and the other half with the alternative version. After each melody, participants had to report whether this melody sounded familiar or unfamiliar, with respect to what they just were exposed to in the previous phase. Participants were tested individually in an EEG testing booth. Audio files of the stimuli were encoded at 16-bit resolution and 44.1 kHz sampling rate and presented via Etymotics Research ER-2 earphones. The stimuli were presented at a comfortable loudness level above 60 dB SPL (A-weighted). Instructions were displayed on a computer screen, and participants' responses were collected with a keyboard and mouse. Informal debriefing with participants indicated that both scales were perceived as equally unfamiliar, and no formal familiarity ratings were collected

after each session.

4.2.3 Data Acquisition

Electroencephalogram (EEG) data were recorded using a 64-channel system (ActiCap, BrainProducts) at a sampling rate of 500 Hz with one ground electrode and re-referenced to the average. We used a default fabric head-cap that holds the electrodes (EasyCap, Equidistant layout).

4.2.4 EEG Prepossessing

EEG data were first mean-centered to perform zero-order detrending. We detected bad channels as exhibiting amplitude above three standard deviations from the channel average. Selected bad channels were then interpolated using a weighted sum of neighboring channels' signals. To avoid artifacts caused by low-pass filtering, we subtracted its slow varying trend from each channel by robust-fitting a $30th$-order polynomial [47]. We then applied a low-pass filter with a 40 Hz cut-off and downsampled the resulting data to 100 Hz. Using a time-shift PCA, eyeblink artifacts were isolated and projected out using data collected by the HEoG, and VEoG channels [48]. Finally, the EEG data were re-referenced again by subtracting the robust mean (i.e., as defined in [47]) before it was epoched using the triggers sent at the beginning of each trial. Bad epochs were selected based on an amplitude above three standard deviations and discarded for the analysis.

4.2.5 Temporal Response Function

To evaluate the different topographical mapping in melodic encoding between the two scale structures, we used a brain decoding method based on Temporal Response Function (TRF) [87]. The TRF is based on a class of linear time-invariant models that describes the linear transformation of stimuli features to the neural signal (EEG) by its impulse response after ridge regression. Unlike Evoked Response Potentials (ERPs), the response function obtained reflects a modeled neural response to a *specific* set of features (when an ERP represents the grand average to the whole stimulus). More precisely, the TRF optimally describes the mapping between a given set of features of a sensory input $s(t)$ and the neural response $r(t)$ collected from each channel n of the neural signal, such as defined in Eq. 4.1:

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

Where τ is the specific range of lags for which the response at time t is described (here, [-100 - 500] ms) and $\varepsilon(t)$ is the residual error at each channel n not explained by the model. The TRF $w(\tau,n)$ is estimated by minimizing the mean-squared error between the actual neural response and the one predicted by the convolution $w(\tau,n) * s(t-\tau)$. The model is optimized using ridge regression and assuming a certain degree of regularization to prevent over-fitting. This regularization parameter is optimized in the $[10^{-3}, 10^{3}]$ interval, using logarithmic steps, and for each data set. Cross-validation via leave-one-out evaluation using Pearson's correlations between the predicted and actual neural responses is conducted to evaluate the model's performance. The resulting topographical map indicates the strength of stimulus feature encoding at each EEG

channel.

4.2.6 Denoised ERPs

For the ERP analysis, a specific denoising algorithm called Denoised Source Separation (DSS) was applied (for a detailed explanation, see [52]). In a nutshell, DSS isolates components of signals that are mostly repeated across repetitions of trials to keep the relevant signal (e.g., one that reflects stimuli properties) and remove the signal resulting from noise. In the present study, the Denoised Source Separation (DSS) filter's output was the weighted sum of the signals from the 64 EEG electrodes, in which the weights were optimized to extract the repeated neural activities across trials. This transformation yielded 64 uncorrelated brain source activities (e.g., DSS components), which were ordered by a repeatability score. Since the trials were not precisely identical between repetitions, we selected only the first five most repeatable DSS components and projected them back into the sensor space to obtain cleaned signals. Finally, we used the obtained denoised Cz electrode (placed on the mid-line sagittal plane center) for the ERPs analysis.

4.2.7 Decoding

To evaluate the separability of neural traces elicited by melodies from the two scales, we trained a set of logistic regression classifiers on the preprocessed EEG data (e.g., not DSS-denoised). At each time point t we used the matrix of observations $X_t \in R^{N\times 64}$, for N samples of all 64 electrodes to predict the labels $y_{t'} \in \{0,1\}^N$. Here, the labels correspond to the two grammar conditions (alternative versus reference melodies) or the state of transitions (correct versus incorrect), or the probability of transitions (low probability versus high probability transitions).

This was repeated for every time point t' of each epoch. This analysis was conducted two times: on the epochs collected from the uniform scale conditions and the non-uniform scale condition. We trained the decoders on EEG signals for each subject at each time point of the melodies (from onset to 6 s after onset). Therefore, the decoder at each time point learns to predict the grammar conditions (i.e., alternative versus reference) using the topography of the EEG samples for this time point. Additionally, temporal generalization analysis was conducted to capture the dynamics of topographical patterns of EEG signal over time (for more details on that, see [51]). To achieve that, we systematically evaluated each classifier from each time point to all other time points. Concretely, this means that a classifier trained to separate labels at a given time point is then used to predict the labels at all other time points.

To validate the classifier's performances, we used 5-fold cross-validation. This means that for each individual data set, the trained classifier was used to predict labels on the 5th portion of unseen data over five iterations. The area under the receiver operating characteristic curve (AUC) was used to quantify the classifiers' performance. We implemented this decoding analysis using sci-kitlearn [49] and MNE [50] libraries in python 3.6.

4.2.8 Statistical Analysis

ERP analysis. To compare the averaged ERPs between conditions, we performed bootstrap resampling in order to estimate the standard deviation (SD) of the difference between the state of transitions. Significance levels were set for the difference between the two signals above $2\times$estimated SD. Error bars represent ±SEM (standard error of the mean).

Temporal Response Functional. To evaluate the significance of topographical differences

between the two maps, an FDR-corrected paired t-test was conducted on the distribution of Pearson's r-values resulting from the TRF analysis ($\alpha = .05$).

Classifiers. Statistical analysis for the classifiers was performed with a one-sample t-test with random-effect Monte-Carlo cluster statistics for multiple comparison correction using the default parameters of the MNE spatio_temporal_cluster_1samp_test function [53]. Error bars in all figures represent ±SEM (standard error of the mean).

4.3 Results and Discussion

Just like language, music is ubiquitous, found in all known human cultures [88–90] but displays varying structural norms across cultures [78, 79, 91, 92]. Because playing and listening to music reinforce social bonding around everyday rewarding experiences [93], learning music can be seen as an essential ability for society survival [94–99]. Thus, individuals may seek musical exposure and implicitly learn the structural regularities underlying the music corpus corresponding to their culture.

Despite the overwhelming variability in structural features across musical cultures, some few features actually display significant prevalence. Among them, the non-uniform structure of musical scales is present in a vast majority of musical cultures [100], motivating the hypothesis that it could benefit learning tonal hierarchies and melodic regularities [83, 101]. Pelofi & Farbood conducted a behavioral study showing enhanced learning of artificial grammars for melodies generated from asymmetric scales [85].

Building up on the same design, the present study seeks to establish that these facilitatory effects originate from the enhanced neural encoding of melodies derived from asymmetric scales.

Specifically, the encoding of the statistical structures (manifested as grammatical features) was investigated in the context of a uniform and non-uniform scale, schematically represented in Figure 4.1A. An artificial first-order grammar (Figure 4.1B) was used to generate melodies. First, in an exposure phase, listeners were presented with a corpus of 100 *reference* melodies generated from the original grammar. In the following test phase, half of the presented melodies were from the original grammar (i.e., *reference* melodies). In contrast, the other half contained wrong transitions that were inserted so as to induce syntactic violations in the melodies (i.e., *alternative* melodies). Listeners had to report whether each of the melodies sounded familiar with respect to what they heard in the exposure phase. This procedure was repeated two times, one for each scale condition. Details of the experimental setup, subjects, and stimuli are provided in section 4.2.

4.3.1 Behavioral Results

Following the exposure phase, listeners were presented with either correct or incorrect (i.e., containing incorrect transitions) melodies and had to report whether the melody in each trial sounded familiar. A mean d' values for both scales are shown in Figure 4.1C. The performance was significantly higher when melodies were generated with the non-uniform scale (paired t-test, $p = 0.023$). Over the course of the test phase, listeners were presented with more and more incorrect transitions in the melodies, which could potentially alter the representation of the grammar they acquired throughout the exposure. To account for this expected drift in performance over time, mean d' values were averaged across listeners for different evenly divided sets of trials over time (first set: trials 1-27, middle set: trials 28-55, and last set: trials 56-80, Figure 4.1D). For

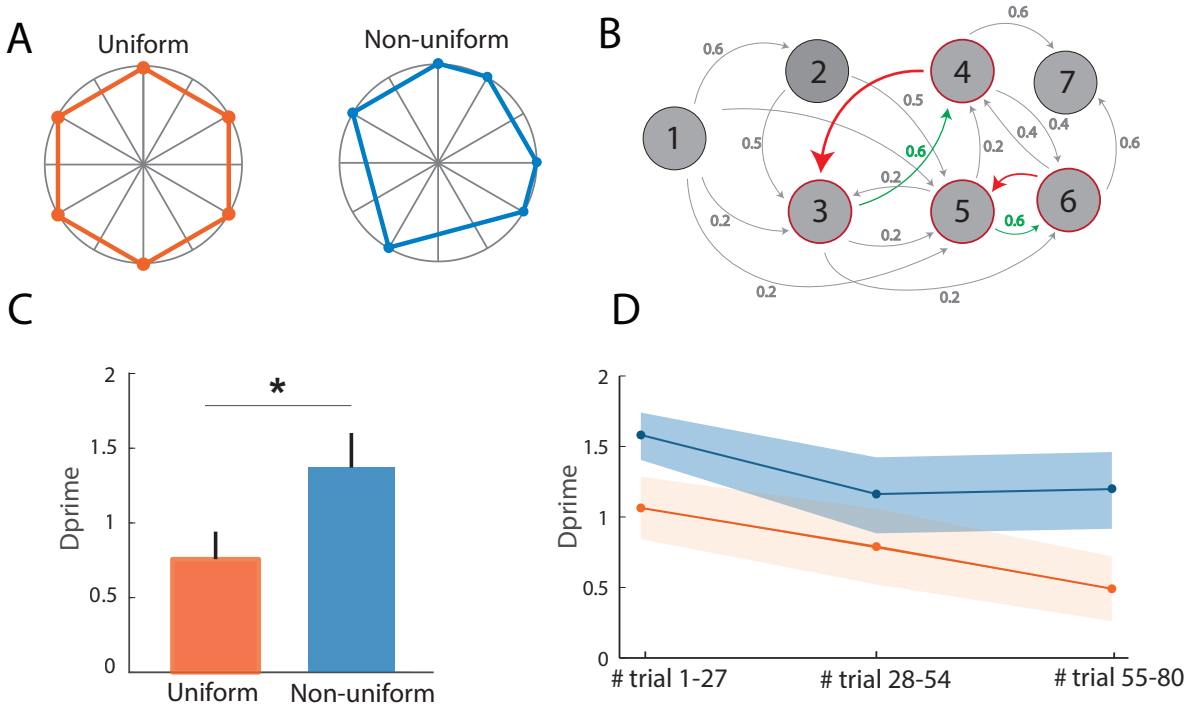

Figure 4.1: **Method and behavioral results**. **(A)** Schematic representation of the uniform (red) and non-uniform (blue) scales as circular diagrams. **(B)** The first-order Markov-chain grammar used to generate melodies from the uniform and non-uniform scales. Nodes represent scale notes; gray and green arrows represent the transitions between nodes used to generate exposure and reference melodies; and red arrows represent two possible examples of "incorrect" transitions used to generate half of the test melodies (i.e. the alternative melodies). **(C)** d' values averaged across participants by symmetry condition: uniform (red) and non-uniform (blue). Error bars correspond to standard error. **(D)** To account for the drift in performance during the test session, d' values are averaged across participants for three different trial groups: trials 1-27, trials 28-54 and trials 55-80.

both scale, a significant drift in performance over time was observed (two-way Repeated Measure ANOVA: $F(14, 2) = 4.75$, $p = 0.017$).

4.3.2 Neural Encoding of Melodies

To investigate the difference in grammar encoding for uniform and non-uniform scales, we trained a set of linear classifiers to discriminate between reference (i.e., played with the same grammar as during the exposure phase) and alternative (i.e., those containing syntactic violations)

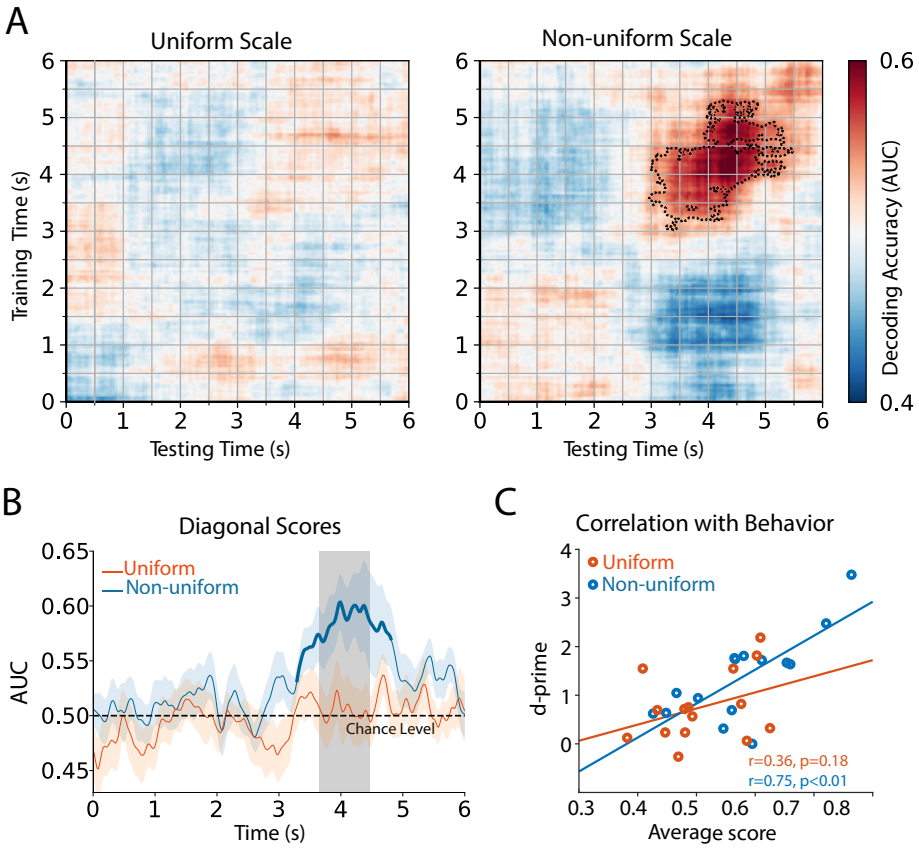

Figure 4.2: **Decoding alternative versus reference melodies**. We trained a set of linear classifiers in order to decode reference from alternative melodies. **(A)** Decoding performance is depicted here for *uniform* (left) and *asymmetric* (right) scales. Classifiers trained and tested separately at each time in a 6 seconds time window following the onset of melodies. Cluster-corrected significance is contoured with a dashed line. The classifier scores were significantly above the chance level only for the asymmetric scale ($p < 0.05$). **(B)** Decoding performance for the same training and testing time points, which is equal to the diagonal scores in part **A**. The bold curve marks time points where predictions were significantly above the chance level ($p < 0.05$). The difference between the scores in uniform and non-uniform scales was significant for the gray bar. **(C)** The average decoding scores for the significant times (Bold curves in part B) were correlated with the corresponding behavioral performance, across subjects. The decoding scores and d' are significantly correlated only for the non-uniform scale. These results suggest that neural processing encoded the melodies in asymmetric scale significantly better.

melodies. We used the signal collected over the entire melody so as to probe at what time point in the melody the decoding accuracy diverged between the two scale conditions. The classifiers were a set of independent logistic regressors trained to discriminate EEG signals using data from

all 64 sensors as detailed in section 4.2. In other words, the trained classifiers sought to linearly separate the two conditions based on the differences within the EEG topographic maps due to the reference and alternative melodies. To probe for time generalization dynamics [51], the classifiers were trained on the EEG response of the 6 seconds of melodies at each time point t and tested at time t', where t and t' were within the melody time window (0 s to 6 s).

Figure 4.2A illustrates the time generalization dynamics of decoders' performance for the melodies generated from the uniform (left) and non-uniform (right) scales. The classifiers' scores were significantly above the chance level for the non-uniform scale ($p_{min} < 0.05$) between 3 and 5 seconds after onset. This demonstrates that listeners could learn the unfamiliar and artificial musical grammar from melodies generated in this scale. The pattern of the significance region (right) suggests a temporally jittered activity due to the subtle variations in the emergence of the effects across subjects [51]. Conversely, the classifiers could not discriminate between the reference and alternative melodies when melodies were generated using the uniform scales ($p_{min} > 0.05$). To simplify the visualization, we directly compared the diagonal scores in non-uniform and uniform scales (i.e., the scores for which testing and training data were synchronized); we observed a statistical difference between the classifier performances in which the area under the receiver operating characteristic curve (AUC) was significantly larger for non-uniform scale at around [3-5] seconds following the melodies onset (Figure 4.2B).

Finally, we correlated the average decoding scores of each subject for the duration of the region of significant behavioral performance (d'), with Pearson correlation as the test statistic. The behavioral performance was significantly correlated with the average decoding scores only for the non-uniform scale ($p < 0.01$, correlation in significant regions $r = 0.75$) as seen in Figure 4.2C.

4.3.3 Topography of Scale Effect

To visualize the effect of scale in EEG sensor space, a Temporal Response Function (TRF) analysis [87] was conducted on the data collected for the uniform and non-uniform scales. The TRF is a decoding technique used to account for the neural encoding of continuous signal features, such as envelope, semantics, or phoneme for speech [102–105] and envelope and syntax for music [106, 107]. Here, we used the probability of note transitions (i.e., the grammatical structure) between the notes of each melody as the regressor feature. This could give an index of how well the grammatical structure of the melodies was represented in the neural data for each scale. This technique fits a kernel, the TRF, that describes the linear mapping of the stimulus into the neural signal using ridge regression. The kernel is fit to minimize the difference (in terms of mean-squared error) between the actual neural response and the predicted neural response. The encoding index is then assessed using a cross-validated evaluation using r Pearson's correlations between the predicted and actual neural responses. Details of the TRF analysis are provided in section 4.2.

Figure 4.3A shows r-values that indicate the topography of the effect of scale. Two TRFs were computed on all the test melodies for the uniform and non-uniform scales. The "real model" is used as a regressor for the predicted surprise for each note, defined as the -log(Probability). The "null model" used a 100 permutation using a shuffled version of the surprise as a regressor. For each scale, the baseline r-values from the null model was subtracted from the real model. Then r-values (corrected with baseline) from the uniform scale were subtracted from the non-uniform scale and plotted in Figure 4.3A. The topographies reveal that significant scale effects were observed on the centro-lateral electrodes (Figure 4.3A, left), consistent with previous studies

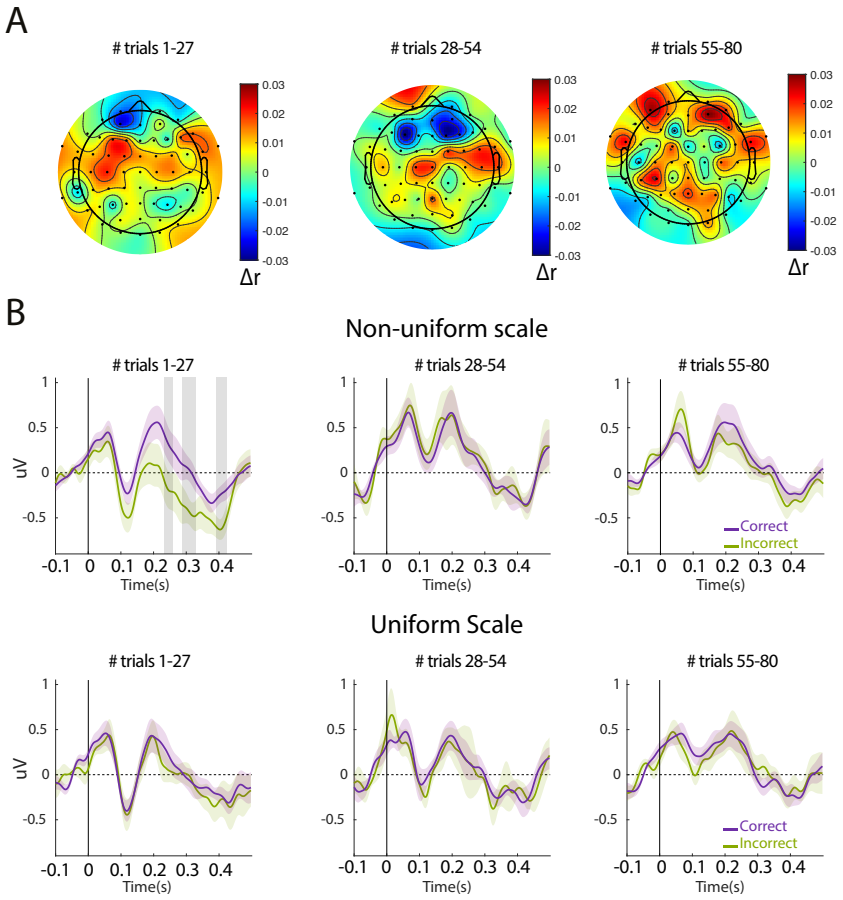

Figure 4.3: Evoked Responses to the Tone Transitions. **A** Topography of difference in grammar encoding. Non-uniform - Uniform r-values obtained from the TRF of probability of notes for all melodies (reference and alternative) are first subtracted from a null model (100 permutation). **B** Evoked responses at channel Cz for *correct* and *incorrect* transitions. Comparison between the evoked responses due to the *correct* transitions with *incorrect* transitions for three separate sets of trials. Time windows 1, 2, and 3 referred to the first, second, and third portions of the trials. There were significant differences between the *correct* and *incorrect* transitions only in *time window 1* for the asymmetric scales.

investigating the encoding of musical grammar in EEG signals [107].

4.3.4 Neural encoding of note transitions

To further analyze the neural processing of melodies, we compared the evoked responses to the notes for incorrect transitions (i.e., inconsistent transitions with the reference melodies)

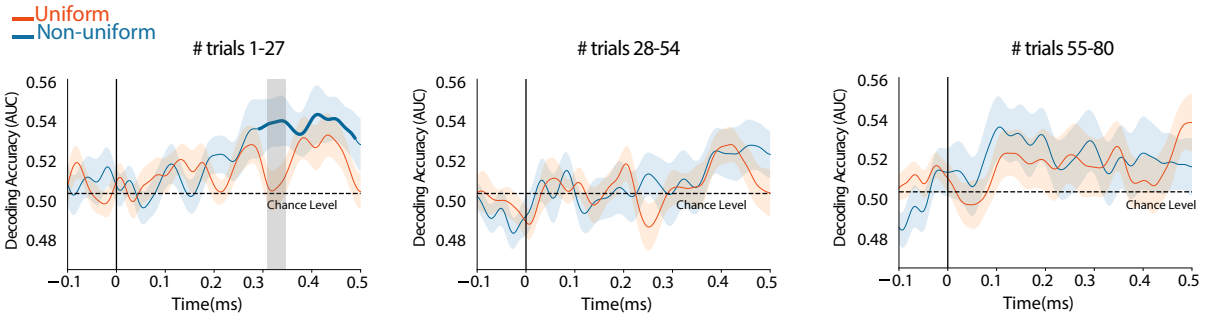

Figure 4.4: Decoding the *correct* versus *incorrect* transitions. Decoding performance for *uniform* (orange) and *asymmetric* (blue) scales in three separate time windows. Time windows 1, 2, and 3 referred to the first, second, and third portions of the trials. Cluster-corrected significance is marked with a bold line. The classifier scores were significantly above the chance level only for the asymmetric scale ($p < 0.05$) in the the first time window. The difference between the scores in uniform and asymmetric scales was significant for the gray bar.

against the notes' evoked responses in reference melodies. We obtained denoised ERPs by applying the denoising method based on spatial filtering (see section 4.2 for details). As before, we divided trials into three separate time sets to account for the alteration in response to incorrect transitions (first set: trials 1-27, middle set: trials 28-55, and last set: trials 56-80). We observed a significant difference between correct and incorrect transitions only for the non-uniform scale and in the first time window Figure 4.3B.

Finally, we computed and applied a set of linear classifiers to decode the correct versus incorrect transitions in those three sets of trials over time. The decoding scores were significantly above the chance level only for the non-uniform scale, starting at 350 ms after the onset of the tones ($p < 0.05$). The classifiers, however, failed to decode the transitions for the uniform scale, suggesting that neural processing of melodies was sensitive even for a smaller scale like the tone transitions. Therefore, EEG responses better encode the statistical structure of these melodies when they are driven from the non-uniform scales.

4.4 Conclusion

Processing and learning music stands as a fundamental aspect of human cognition, as most human engage spontaneously and regularly into musical activities. Although musical cultures drastically vary from one society to another, some recurrent and quasi-universal features suggest that certain cognitive principles may constrain the way musical cultures are shaped [100]. In particular, almost all known musical scales display intervals of different sizes (whole tones and semitones in the Western system) which are positioned around the octave in a way that maximizes *uniqueness* through intervallic, non-uniform structures [108]. By contrast, uniform scales are seen very rarely, but are famously featured in Gamelan music of Java [84].

Recent findings suggest that the pervasiveness of non-uniform scales pertains to the cognitive benefits they provide for learning melodic regularities [85]. Here, we present further evidence demonstrating differences in the neural encoding of melodic structural regularities derived from the uniform and non-uniform scales. A one-order Markov chain was used to generate melodies that were either derived from a uniform or non-uniform scale. After a short exposure presenting melodies respecting the grammar, a test phase presented either similar melodies (i.e., reference melodies) or test melodies that were generated using a slightly different version of the grammar (i.e., alternative melodies). Behavioral as well as neural data were collected during the test phase. The analysis revealed that the behavioral effect of enhanced learning for the non-uniform condition was paralleled by an enhanced neural encoding of the grammar for these melodies. Using a linear classifier, we demonstrated that the two types of melodies (i.e., alternative *versus* reference) were better separated in the neural data when derived from the non-uniform scale. This separability was correlated with behavioral performances at the individual level. Therefore, we

observed that implicit statistical learning is enabled only in the context of the non-uniform scales and it is diminished for the uniform scales. Thus, this contributes to the universality of non-uniform scale features in music perception. These findings provide evidence for the interaction of neural processing constraints and the associated cultural influences.

One other example of how innate, sensory principles shape musical features concerns how the smallest intervals found between structural tones (*e.g.*, semitones) [82] reflect typical frequency discrimination abilities [109]. Recently, Jacoby and McDermott [110] found that listeners from different cultural backgrounds produce complex rhythms converging on integer-ratio temporal intervals. However, this line of research also demonstrates the opposite—that seemingly pervasive elements of musical structure fail to point to universal aspects of perception and production [111, 112].

Overall, these results shed further light on the challenging question of how primary cognitive and sensory constraints intersect with cultural influences and result in universal features in human musical perception and production [113, 114]. From a cross-domain perspective, Chomsky's assertion [115] that innate principles underlie the many manifestations of language (the concept of a Universal Grammar) provided a framework for such a question and, in the process, ignited much research in the field of human cognition. The discovery of structural universals in language, such as the architecture of phonological structures [116], has tremendously impacted our understanding of the human mind [117]. In the domain of music cognition, such efforts have been far more limited but will surely benefit from the recent development of corpus-based statistical analyses [80, 100].

Chapter 5: **Moving to Imagination Land: Decoding the Neural Responses During Auditory Imagery Using the Listening Brain Signals**

5.1 Introduction

Auditory imagery is an endogenous neural process during perception when elicited voluntarily without any auditory sensory input [118]. Imagery tasks are an effective way to probe the internal computations at the foundation of auditory processing in the human cortex (e.g., sensory-motor theory). Mental imagery, in general, may also be a window to investigate *consciousness*, where some studies reported it is associated with intermediate states of consciousness [119]. Studying auditory imagery can also serve numerous applications in BCI, such as speech recognition during a speech imagery task ranging from decoding phonemes and words to synthesizing complex sentences from neural processing of speech imagery.

Several studies argue that perception emerges as an interaction between sensory input and endogenous neural activity as an internal model [120–123]. From this perspective, the endogenous neural activities during imagery share underlying processes with auditory listening, where perception in imagery and listening tasks elicit overlapping patterns of activity. Specifically, auditory imagery tasks share several brain regions with listening conditions, such as temporal lobes related to the auditory perception [124–130], Wernicke's areas which are associated with

speech perception [131], Broca's area (inferior frontal gyrus), which is related to speech production [132–134], and regions associated with higher levels of cognition like Prefrontal Cortex (PFC) [126, 135, 136] for speech and music imagery tasks.

Although fMRI studies reveal the brain regions involved during imagery, they do not provide any information regarding its temporal and often spectral neural processing. Instead, the temporally finer techniques such as EEG, MEG, or ECoG indicate that there are common patterns of neural activities between listened and imagined responses. This shared neural processing can be seen when the neural signals encode the melodic expectations during music imagery and listening [126, 129, 131], and crucially when the neural processing during a music imagery task preserves the temporal dynamics of the acoustic stimuli such as note transitions and silences' timing [137, 138]. These findings were also consistent with an ECoG study, where acoustic features of imagined speech could be reconstructed based on the computational model built from listened speech [139]. Some studies also have shown an early-stage interaction between the top-down generation of speech mental imagery and bottom-up stimulus-driven perception by using the imagery-perception repetition paradigm [140, 141]

In a light of predictive coding and aforementioned evidences, Di Liberto et al [138], proposed a computational model for music imagery in which the endogenous and exogenous neural activity during perception are two distinct additive components, and the stimulus expectations (e.g., predictions) influence both neural signals evoked by listening and imagery musics. This model suggest that a direct relationship exists between listened and auditory imagery, and that one we can find a mapping to predict imagery neural signals from listening responses.

The timing issue with imagery tasks makes decoding the imagery neural signals very challenging; these difficulties arise since one does not know *exactly* when a subject starts an imagery

task and does not have an estimate of the actual neural processes of auditory imagery for each subject. One may overcome these challenges if one has access to an estimate of the neural responses during an imagery task and use the estimate to decode the content of mental processing during an actual imagery task. Here, we attempt to find this relationship in an EEG experiment where collected musical imagery and listening responses from highly trained musicians. We then trained a non-linear neural network to relate the two sets of responses, i.e. to use the neural responses of listened music to reconstruct the imagery neural signals. We trained the encoder-decoder neural network architecture to show that there is a generalizable mapping that can reliably predict the imagery signals from listening. Moreover, in an additional MEG experiment, we found that such a mapping can be independent of the nature of the stimuli, in that we could transfer a trained mapping on music to make it work equally well on the speech stimuli without losing the prediction power. This suggests that for this specific task, there is an overlapping neural substrate and processes for imagery and listening conditions. Furthermore, these processes have similar underlying mechanisms for different acoustic stimuli (be it speech or music).

5.2 Materials and Methods

5.2.1 Data Acquisition and Experimental Procedure

Experiment 1 (EEG): Twenty-one healthy and highly trained musicians (6 females between 17 and 35 years old, median = 25) participated in the EEG experiment. Each subject is provided written informed consent and was paid for their participation. The study was undertaken under the Declaration of Helsinki and was approved by the Health Research Ethics Evaluation Board of Paris Descartes University (CERES 2013-11). EEG data were recorded from 64 scalp electrodes

in an elastic cap using a BioSemi ActiveTwo (BioSemi Instrumentation). EEG signals were acquired at a sampling rate of 2048 Hz and referenced to the average. Participants were instructed to minimize motor activities while performing the task. To control for muscle movements three additional electrodes were placed on the upper midline of the participants' neck, jaw, and right wrist. The stimuli were presented in python at a sampling rate of 44100 Hz using a Genelec 8010 loudspeaker. Testing was conducted at École Normale Supérieure in a dimmed room.

The stimuli experiment consisted of 4 melodies, each 35 seconds long, from a monophonic MIDI corpus of Bach chorales (BWV 349, BWV 291, BWV 354, BWV 271). All chorales use similar compositional principles. There were 88 trials in which participants were asked to either listen or perform mental imagery, with each melody presented 11 times per condition (listening and imagery) throughout the experiment. The presentation order of the resulting 88 trials was randomized. Participants were asked to read the music scores placed at the center of the desk during listening and imagery conditions. In order to maximize the imagery performance, participants were asked to become familiar with the melodies using the provided scores. A tactile metronome (Peterson Body Beat Vibe Clip) marking the start of 100 bpm bars (each 2.4 s) was placed on the left ankle of all participants to allow them to perform the mental imagery task with high temporal precision. A constant lag of 35 ms was determined during the pilot experiments based on the subjective report of the participants, who reported that the metronome with a lag of 0 ms was not in sync with the music. That correction was applied to all participants with the same lag value. Neural data from 0 to 500 ms after each metronome onset were excluded from the main analyses to avoid contamination due to tactile responses. Note that the EEG response to the metronome reflects a mixture of tactile and auditory responses in the listening condition. This data for the *experiment 1* has been published in [137, 138].

Experiment 2 (MEG): Seven self-reported normal hearing subjects (3 female, median = 24) were participated in the MEG experiment. All subjects were highly trained musician. The experimental procedures were approved by the University of Maryland Institutional Review Boards, and all participants were given course credits or monetary compensation for their participation.. Written, informed consent was obtained from each subject before the experiment. MEG data were acquired using whole head KIT (Kanazawa Institute of Technology) system, with A 157 axial gradiometer. We collected data at 1 kHz sampling rate with an online 500 Hz low pass filter, and a 60 Hz notch filter. Subjects rested in the supine position in a magnetically shielded room (VAC), while the MEG data were recorded. The experiment consisted of 4 different stimuli – two melodies and two speech snippets – each 27 seconds long. The melodies were derived from a monophonic MIDI corpus of Bach chorales (BWV 263, BWV 354). One of the melodies was identical to the *experiment 1*, and both melodies use similar compositional principals. The speech stimuli consisted of two distinct part of a poem ("A Visit from St. Nicholas," Moore or Livingston, 1823), and the two parts (poem 1 and poem 2; Table 5.1) were recorded by a professional voice actor. We used Audacity software for the audio editing. A noise reduction, and loudness normalization was performed to match the loudness of the chorale audio files. To ensure that the melodies and the poems had the same tempo, we split the speech audio into separate lines

Poem 1	*Poem 2*
When out on the lawn – **there** arose such a clatter	**He** was dressed all in fur – **from** his head to his foot
I sprang from my bed – **to** see what was the matter	**And** his clothes were all tarnished – **with** ashes and soot;
Away to the window – **I** flew like a flash	**A** bundle of toys – **he** had flung on his back
Tore open the shutters – **and** threw up the sash	**And** he looked like a peddler – **just** opening his pack
The moon on the breast – **of** the new-fallen snow	**His** eyes, how they twinkled! – **his** dimples, how merry!
Gave a lustre of midday – **to** objects below	**His** cheeks were like roses – **his** nose like a cherry!

Table 5.1: Speech stimuli consisted of two separate parts of "A Visit from St. Nicholas" poem. The onset of the bold words are in sync with the downbeat of the metronome.

and manually moved each chunk so the onset of each line aligned with a metronome downbeat at 120 bpm. Throughout the experiment there were 10 listening and 10 imagery trials per stimuli. Therefore, the listening and imagery conditions had 40 trials each. We presented the resulting 80 trials with randomized order to each subject. We provided the participants with the stimuli, the written poems, and the music scores a few days before the experiment, and tasked them to become familiar with the stimuli. A separate training session was conducted with individual participant to maximize their imagery performance. During the training, each participant practiced the listening and imagery tasks, as well as singing the melodies and reciting the poems to check for accuracy. During the MEG experiment, the stimuli audio was delivered to the subjects via Etymotics Research ER-2 insert earphones at a comfortable loudness level (70 dB). To ensure that participants performed the mental imagery task with high temporal precision, a visual metronome (A clock shape with a moving hand, see Figure 5.2A) marking the start of 120 bpm bars (each 2 s) was presented on the screen. In the speech stimuli, the downbeats of the metronome was synchronized with the first and a middle words in each line of the poem (see the bold words in table 5.1).

5.2.2 EEG Preprocessing

Preprocessing of the EEG data were performed offline using MATLAB software (MathWorks). EEG signals were mean-centered. The bad channels which exceeded a threshold criterion (standard deviation of channels amplitude) were detected and were interpolated based on the data from the neighbor channels using spherical spline interpolation. EEG signals were bandpass filtered between .1 Hz to 30 Hz with Butterworth zero-phase window of order 2 using 'filt-

filt' function in MATLAB. Channels were then re-referenced to the average of the 64 channels. All subsequent analyses were performed in Python 3.8, mne-python 0.24.1 [50], and eelbrain 0.33 [142]. The MEG data were filtered from 0.1 to 30Hz using an FIR filter (mne-python 0.24.1 default settings), and downsampled to 100 Hz.

5.2.3 MEG Preprocessing

We used Python 3.8 to preprocess and analyse the MEG data offline. Saturating channels were excluded (approximately two channels on average) and the data were denoised using time-shift principal component analysis [48] to remove external noise, and sensor noise suppression [52] to suppress channel artifacts. All subsequent analyses were performed in mne-python 0.24.1 [50] and eelbrain 0.33 [142]. The MEG data were filtered from 0.1 to 30Hz using an FIR filter (mne-python 0.24.1 default settings), and downsampled to 100 Hz.

5.2.4 Encoder-Decoder Neural Network

We developed an end-to-end Encoder-Decoder Convolutional Neural Network (CNN) architecture in PyTorch 1.10 to predict the brain signals during the imagery tasks from the brain signal during the listening condition (See Figure 5.1). The encoder part consisted of two convolutional layers. The layers convolve over time with ReLU non-linearities. Therefore the Encoder transforms the input sensor data into a latent space. The Decoder block, which includes two transposed convolutional layers (transpose convolution over time), transformed back the latent space signals into the sensor space. All hyperparameters for this network –including the number of neurons, kernel and stride size, learning rate, batch size, and the number of epochs – were

optimized through cross-validation, grid search, and trial and error. In the final network, we used a latent space of dimension 32, a kernel size 4, and a stride of size 2. We initialized the parameters using PyTorch default initialization. Since we had a sufficiently large number of subjects for experiment 1, we only used this network on EEG data. Pearson correlation was used to evaluate the model. Before, training the data, we standardized all the data to have zero-mean and bounded amplitude between -1 and 1.

To train the network, we left out all data from a subject (test subject) and used them for the model evaluation. We trained the network on the EEG data from the remaining subjects in two steps: 1) We pre-trained the architecture as an auto-encoder, in which the network learned to reconstruct the brain signals from the same input. We train the network for listening and imagery conditions simultaneously for all subjects (in the train sets). This way, the trained network learns a common latent space for listening and imagery across all subjects. Subsequently, when the auto-encoder is fully trained, 2) We fixed the encoder weights and continued training the decoder block. We only used the listened EEG in this step to reconstruct the imagined EEG signal.

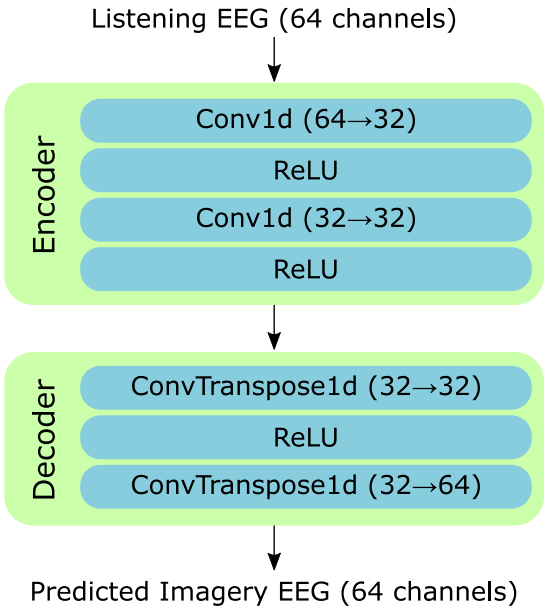

Figure 5.1: The flow of the input data in the final encoder-decoder network used to reconstruct the imagery EEG. The encoder transform the 64-channel EEG input with convolution and ReLU non-linearities. All layers convolve over time dimension. $Conv1d(d_{in} \rightarrow d_{out})$, and $ConvTranspose1d(d_{in} \rightarrow d_{out})$ mean a one-dimensional convolution and transpose convolution, respectively, with d_{in}, and d_{out} number of input, and output channels.

We evaluate the model and check whether it is generalizable to unseen data from the test subject. The listening EEG data from the test subject were used as an input, and the model

predicted the imagery EEG signals. Finally, we evaluate the performance by computing the correlation between the *true imagery* signals (ground truth) and the *reconstructed imagery* signals. The significance of the correlation was assessed by comparing the results against the noise and the null model, where we used scrambled phase and shuffled trials, respectively. For the null model we shuffled the order of the trials, where the results does not produce a matching EEG pairs. We repeated the above steps to include EEG data for all twenty-one subjects as in the test subject.

5.2.5 Temporal Response Functions (TRFs)

A representation of the stimulus was used along with the reconstructed and true brain responses to estimate the TRFs (equation 5.1).

$$r(t,n) = \sum_{\tau} w(\tau,n)s(t-\tau) + \varepsilon(t,n) \tag{5.1}$$

Here we used the stimulus representation (predictor) $s(t-\tau)$ and fit the filter $w(\tau,n)$, to predict the reconstructed and true brain signals $r(t,n)$ at each time point t. $w(\tau,n)$ is the TRF value at lag τ and for n^{th} sensor, and $\epsilon(t,n)$ is the residual noise. We used a total of 1200 ms for the lag in TRFs (-300 ms to 900 ms). For this analysis, tone onset was used as a predictor to estimate TRFs at each EEG channel using the boosting algorithm [143], as implemented in eelbrain [142]. The boosting algorithm may result in overly sparse TRFs, and hence an overlapping basis of 30-ms Hamming windows (with 10-ms spacing) was used to allow smoothly varying responses. We finally compared the estimated TRFs for reconstructed imagery with the TRFs for the true imagery.

5.2.6 Linear Mapping for MEG data

Since we do not *yet* have a sufficiently large MEG data set in experiment 2, we used a linear mapping for each subject separately to reconstruct the imagery MEG signals from listening MEG. A multivariate TRF (mTRF; [87]) analysis was used to find a linear mapping between listening and imagery MEG signals. Here we estimated the mTRFs and used MEG data in the listening condition as a predictor, we then predict the imagery MEG signals. We evaluate the predicted signals using the correlation, similar to the non-linear model described above (section 5.2.4). We fit the mTRFs on all the MEG data including the music and poem condition, and also we performed a cross-condition analysis by separately fitting the mTRFs on MEG signals during music (speech) imagery and tested it on MEG data during speech (music).

5.2.7 Linear Classifiers

Decoding analysis was performed using sci-kitlearn [49] and MNE [50] libraries in python 3.8. We trained a set of linear decoders to classify the reconstructed and true imagery signals into each stimulus category (chorales 1-4 in experiment 1 and chorales 1-2, poems 1-2 in experiment 2). For the duration of a trial, T, at each time point $t \in [0, T]$, we trained a classifier using the matrix of observations $X_t \in R^{T \times M}$ (true or reconstructed), for M electrodes in T samples, to predict the vector of labels $y_t \in \{0, 1, 3, 4\}^T$. The labels are corresponding to {*chorale 1, chorale 2, chorale 3, chorale 4*} in experiment 1 , and {*chorale 1, chorale 2, poem1, poem 2*} in experiment 2. Therefore, the classifiers' objective was to predict one of the four stimuli a participant imagined for a given time t. To predict the imagery content for the entire trial's duration, we took the majority vote of the T classifiers. Furthermore, we trained classifiers on

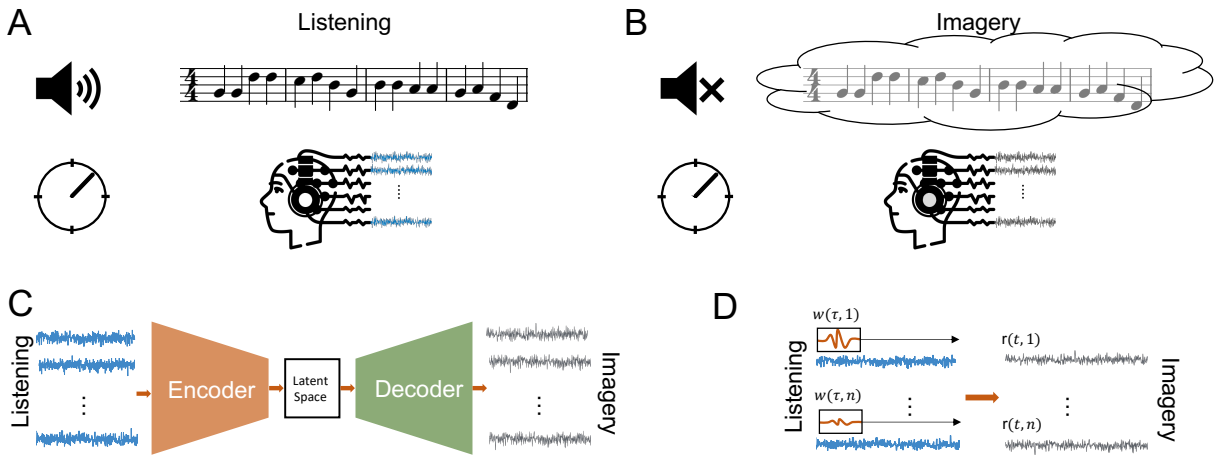

Figure 5.2: Summary of the experiments. **(A)**, **(B)**, EEG (experiment 1) or MEG (experiment 2) were recorded, while the participants imagined or listened to the audio stimuli. In the experiment 1, the stimuli consisted of 4 different monophonic melodies by Bach. During the first experiment, a vibro-tactile metronome was attached to the left ankle helped the precise execution of the imagery task. In the second experiment (experiment 2), the stimuli consisted of two monophonic melodies and two speech stimuli. A visual metronome in the form of a clock with a moving hand was provided to the participants for the precision in the imagery task (for more details in both experiments see section 5.2).**(C)** The recorded listening and imagery EEG were used to train a non-linear mapping between the listening and imagery signals of $N-1$ subjects. During the training the encoder learns to transform the input into a latent space that is shared between listening and imagery EEG. The fully trained model, then, was tested on unseen data-sets of an unseen subject (see section 5.2.4). **(D)** For the MEG data (experiment 2) we used a linear multivariate Temporal Response Functions (mTRFs) model to reconstruct the imagery MEG signals from the listening MEG signal. The mTRFs were trained and evaluated separately on each subjects' dataset.

the reconstructed imagery signals to the true imagery EEG signals.

Logistic regression classifiers were used, with 5-fold cross-validation, within-subject for all the trials. We used the accuracy and the area under the receiver operating characteristic curve (AUC) to quantify the classifiers' performance. We summarized the classifiers' scores in the confusion matrices.

5.2.8 Statistical Analysis

The one-tail pairwise Wilcoxon signed-rank test [54] was used to assess significance of the 1) Pearson's correlation coefficients against the Null model distribution and 2) The classifiers performance against the theoretical chance level. We checked the significance at the subjects and group levels. Correction for multiple comparisons was applied where necessary via the false discovery rate (FDR) approach.

5.3 Results

5.3.1 Mapping the Listening EEG to Imagery EEG

In the first experiment, we used EEG data recorded from 21 highly trained musicians while they were preforming listening and imagery tasks. The stimuli were consisted of four different monophonic melodies from Bach chorales. We attempted to find a mapping to predict the imagery EEG signals from those signals during the listening task. We specifically aimed for a mapping that performs robustly on an unseen subject data. A non-linear encoder-decoder architecture was used to transform the listening EEG and reconstruct

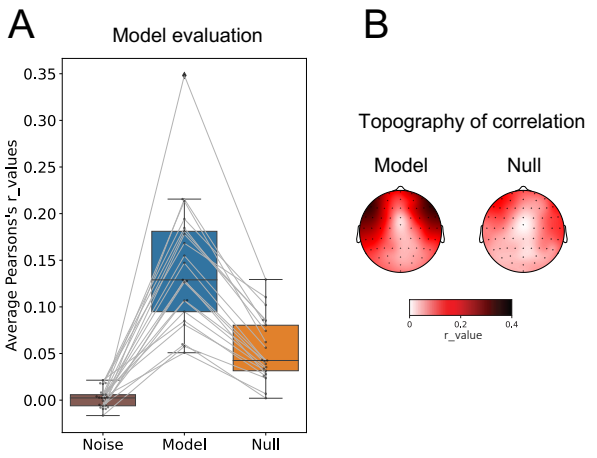

Figure 5.3: Evaluating the model performance. the non-linear model mapped the listening to and reconstructed the EEG imagery signals. **(A)** A model prediction correlation (Pearson's r-value) compared with a null and a noise model. Each gray dot on the box plot represent a r-value for a subject. The correlation was significant for all subjects ($p < 0.05$) and at the group level ($p << 0.05$) after the FDR multiple comparison correction. **(B)** Topography of the model prediction correlation compared with the null model.

the imagery EEG signals (see Figure 5.2). We

first train this model on N-1 subjects and evaluate its performance on the left-out participant (section 5.2.4 for more details).

Figure 5.3A summarizes the model performance compared with the noise and the randomly shuffled (null) models. The Pearson's correlation coefficient of the model output and the true imagery was significantly larger than the null model for the group level ($p << 0.05$) and for all subjects ($p < 0.05$). The robustness of the model indicated that the model transformation found common patterns of activity across subjects and shared neural processing between listening and imagery music. Interestingly, the topography of the model prediction correlation in 5.3B was similar to the topography of a linear model in the melody expectation encoding, which was already reported in [137], for both listening and imagery conditions. These results suggested that the mapping learned to preserve the musicality-related neural processing drive by listening and imagery, aka music perception.

5.3.2 Classification

We further probed whether the reconstructed imagery brain signals preserved the stimulus-related information by investigating whether the reconstructed imagery signals are linearly separable. Therefore, a set of linear classifiers were trained on the reconstructed imagery EEG signals to classify them into one of the four melodies. We also trained similar classifiers on the true imagery and compared the classifiers' performances between the two conditions. To decode the entire trial, we took the majority vote of the classifiers for each trial. Remarkably the decoding accuracy for the reconstructed imagery was significantly larger ($accuracy = 0.67 \pm 0.3$ for recon-

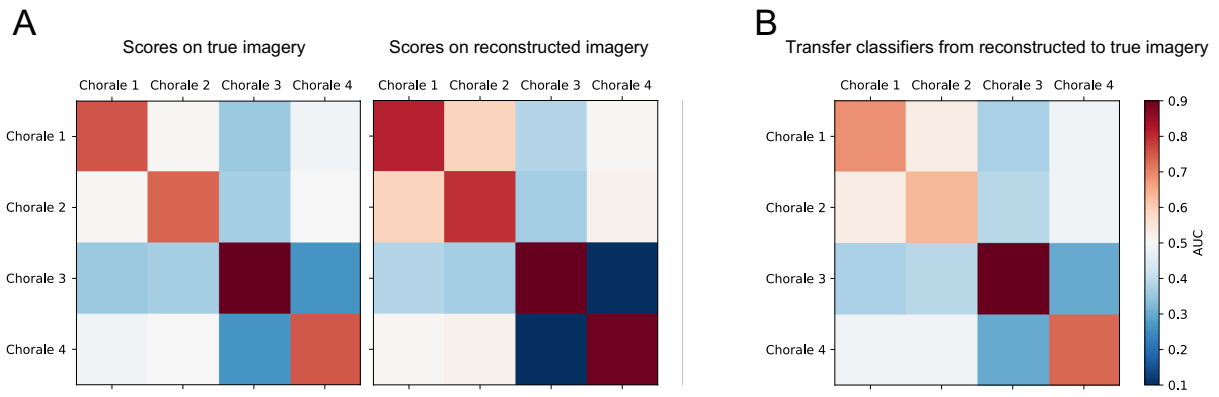

Figure 5.4: Training classifiers on reconstructed and true imagery signals. We trained a set of decoders at each time point to classify the signals into the four melodies. **(A)** The classifiers confusion matrices for the true (left panel) and the reconstructed (right panel) signals. The decoder scores were significantly larger for the reconstructed imagery ($p << 0.05$), suggesting the trained encoder-decoder network preserved and enhanced the neural processing encoded stimulus-related. **(B)** The classifiers were trained on the reconstructed and tested on true imagery signals. The decoder performance were significantly above the chance level ($accuracy = 0.54, p << 0.05$). This is a proof of concept where it is feasible to decode the imagery signals using a transformation from auditory neural signals.

structed than for true imagery, $accuracy = 0.59 \pm 0.4$, $p << 0.05$). Furthermore, the confusion matrices indicated that the reconstructed signals were more linearly separable than in the case of the true imagery (Figure 5.4). These results provide further evidence that the non-linear mapping preserved the musically related neural processes, and that therefore the reconstructed imagery is more denoised.

Furthermore, we trained classifiers on the reconstructed imagery signals at each time point t and tested on the same time point t of the true imagery EEG signals. Figure 5.4B shows a similar confusion matrix to the one for the true imagery, with an $accuracy = 0.54 \pm 0.4$ which was significantly above the chance level ($p << 0.05$). Thereby, we showed that not only there is a reliable mapping that reconstruct the imagery signals from the listening data, but also we could use that reconstructed signal to decode the contents of the mental processing during the music

imagery, even though that the non-linear mapping did not have access to the true imagery signals.

5.3.3 Comparing the TRFs of True and Reconstructed Imagined

We performed a TRF analysis that were obtained using the responses to tone onsets in both the true and reconstructed imagery. By comparing the two sets of TRFs, we were able to assess how similar were the true and reconstructed responses. Figure 5.5 shows that the TRFs fitted for the reconstructed imagery (right panel) were similar to, but smoother than the true imagery TRFs (left panel). Again, as mentioned before, these TRFs suggest that the reconstructed imagery responses are a smoother version of the true imagery, where it preserved the stimulus-related signals that are shared across several subjects.

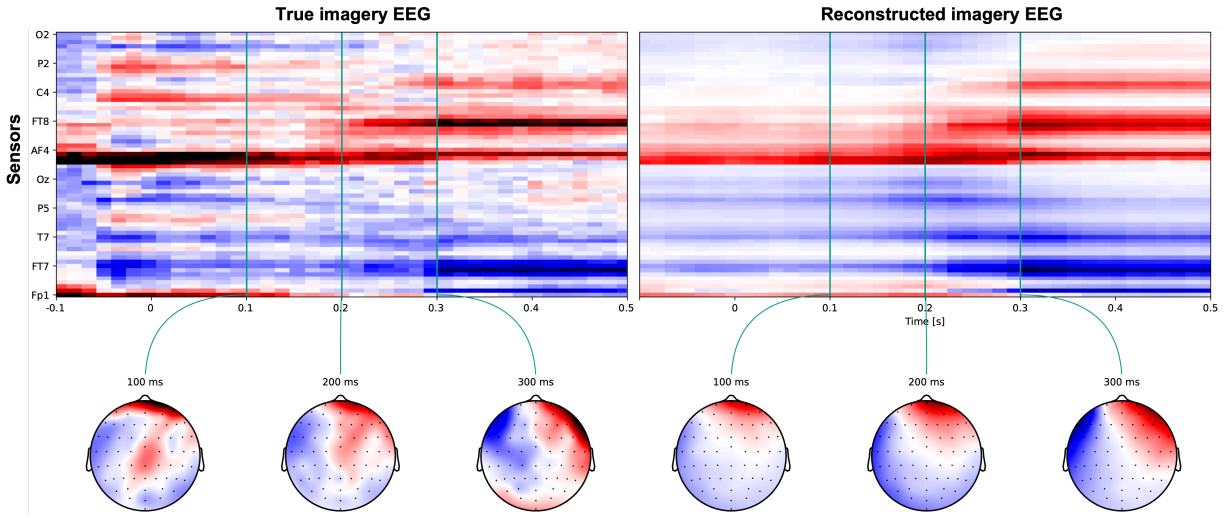

Figure 5.5: Comparing the onset TRFs of true imagery and reconstructed imagery EEG signals. Tone onsets, along with the reconstructed and true imagery signals, were used to estimate the TRF for each EEG channel. We compared the TRFs between the true (left panel) and reconstructed responses. We observed that the TRFs were similar for both conditions, with the reconstructed signals yielding smoother filters.

5.3.4 Mapping the Listening to Imagery responses in MEG

We extend the results in two directions in the second experiment by adding speech stimuli. First, we checked whether a similar mapping exists between imagined and listened speech by training a linear model on the MEG signals. Second, we explored whether we can transfer a model trained on listened and imagery music to speech and vice versa. The second point is crucial since it provides an insight into how similar or distinctive speech and music imagery are.

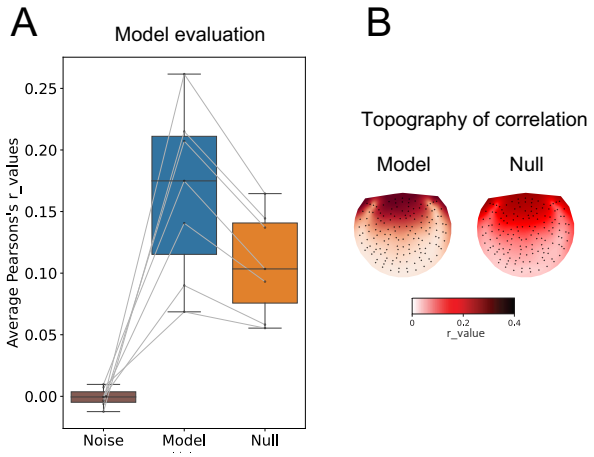

Figure 5.6: Evaluating the model performance. A linear model mapped the listening to and reconstructed the MEG imagery signals. **(A)** A model prediction correlation (Pearson's r-value) compared with a null and a noise model. Each gray dot on the box plot represent a r-value for a subject. The correlation was significant for all subjects ($p < 0.05$) and at the group level ($p << 0.05$) after the FDR multiple comparison correction. **(B)** Topography of the model prediction correlation compared with the null model.

Seven musically trained subjects participated in imagery and listening tasks while we recorded their neural activities using MEG signals. In this experiment, we presented the participants with two music and two speech stimuli. Similar to experiment 1, the music stimuli included two monophonic chorales by Bach. For the speech stimuli, we recorded two distinct parts of *"A Visit from St. Nicholas"* poem (see section 5.2). The speech stimuli were constructed, so the onset of each poem line is aligned with a downbeat of a metronome at 120 bpm (see section 5.2). In 80 randomly ordered trials, participants were asked to imagine or listen to one of the stimuli. A visual metronome was provided in both imagined and listening tasks to help the participants synchronize their performance in the imagery task.

We trained a linear model (see 5.2D) to map the listened data and reconstruct the imagery MEG signals, using the data from both speech and music stimuli. Similar to experiment 1, the model was evaluated by measuring the Pearson's correlation coefficient of the model output (reconstructed imagery) and the true imagery and compared it with the output of a null and a noise model (see section 5.2 for details). The average r values were significantly larger in the model for the group level ($p << 0.05$) and for all the subjects ($p < 0.05$; Figure 5.6A).

It is important to note that we used the linear mapping in experiment 2 since we did not have a sufficiently large data set to train a more complex non-linear model. Although the linear model could not be generalized to unseen subject data, the linear model's performance within the subjects was robust enough to suggest that an encoder-decoder model will perform as strongly, given a more extensive data set.

We performed one more analysis similar to the previous experiment, where a set of linear classifiers were trained to determine the label of the stimuli for true and reconstructed imagery. Figure 5.7A summarizes the decoding scores for true (left panel) and reconstructed imagery (right panel) imagery signals. As in experiment 1, the decoding accuracy for the reconstructed imagery was significantly larger than for true imagery ($accuracy = 0.7 \pm 0.4$ for reconstructed compared to $accuracy = 0.61 \pm 0.3$, $p << 0.05$). Furthermore, the confusion matrices indicate the reconstructed signals are more linearly separable than the true imagery, particularly for the MEG signals associated with the poems.

To summarize, these results suggest there is a mapping

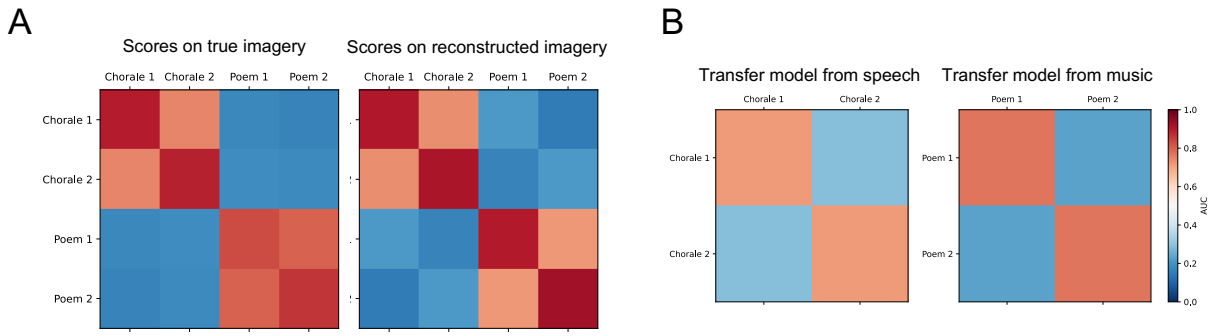

Figure 5.7: Training classifiers on reconstructed and true imagery signals – experiment 2. We trained a set of linear estimators at each time point to classify the signals into the four stimuli (two chorales and two poems). **(A)** The classifiers confusion matrices for the true (left panel) and the reconstructed (right panel) signals. The decoder scores were significantly higher for the reconstructed imagery ($p << 0.05$), suggesting the linear mapping preserved and enhanced the neural processing encoded stimulus-related. **(B)** *Left panel:* The linear mapping was trained on the neural responses associated with the poems (listened and imagery), and the models were transferred to reconstruct the imagery chorales from the listened. The linear decoders, then, were used to classify the melodies. The decoder scores were significantly above the chance level ($accuracy = 0.65, p << 0.05$). *Right panel:* Likewise, a linear model was transferred from the neural processing of the poems to the chorales. The decoder scores were significant on the model output ($accuracy = 0.66, p << 0.5$).

5.3.5 Transfer of Mapping between Music and Speech

We finally investigated similarities between music and speech imagery by transferring mappings across the two conditions. We first trained a linear mapping using only the music data and reconstructed the imagery speech using the trained mapping and listened speech MEG. We also reversed this by training a mapping on speech and transferring it to the music MEG signals. To evaluate the performance of the transferred mappings, we performed linear classifications to determine the label of the stimuli for reconstructed imagery signals. Figure 5.7B showed the decoding scores for reconstructed speech imagery using the music mapping (left panel) and reconstructed music imagery using the speech mapping (right panel). The decoding accuracy for speech and music imagery was significantly above chance level ($accuracy = 0.66 \pm 0.2$ for re-

constructed speech and $accuracy = 0.65 \pm 0.3$ for reconstructed music imagery, $p << 0.05$), suggesting that the mapping between listening and auditory imagery is independent of the stimulus content.

5.4 Discussion

Decoding imagery signals is challenging due to timing problems and the weak EEG/MEG responses to auditory imagery, which are embedded in considerable noise. In addition, the dynamics and the shapes of the non-invasive neural measures such as EEG and MEG are poorly undrestood, which elevates the challenge in decoding the imagery tasks. Nevertheless, if a reliable mapping exists, one can predict imagery neural signals from speech imagery and decode the content of the mental processing during an actual imagery task. In the present study, we investigated the feasibility of reconstructing the auditory imagery neural signals using a model to transform the neural responses during a listening task. We hypothesized that the imagery and listening conditions have similar neural activity patterns so that we could find such a mapping. Thereby, one could decode the true imagery signals and classify the mental content with these reconstructed neural signals.

Here, in an EEG experiment, we reconstructed music imagery signals for 21 subjects using a non-linear mapping between listening and the ground truth imagining neural activities. To transform the listening EEG responses, the model utilized the neural signals only and did not require explicit knowledge of the stimuli dynamics such as melody timing and identity of the notes. We demonstrated that by using the output of this mapping, we could train classifiers to decode the true imagery signals reliably. Of crucial importance, the model was generalizable to unseen sub-

ject data without having access to the subject's ground truth imagery signals. i.e., for a given test subject, we were able to reconstruct imagery from listening signals, train a set of linear decoders on the reconstructed signals, and predict the labels for the melody the subject was imagining. This indicates that the model was robust enough to predict the neural components shared across subjects, sensitive to individual events in a continuous melody, with a remarkable signal-to-noise ratio. Moreover, we observed a higher predicting power for the reconstructed imagery signals than the true imagery or even listening neural signals, which suggests the model preserved and enhanced the stimuli-related neural signals and the model's output is more denoised.

In a MEG experiment, we extended the findings from music imagery, showing the plausibility of reconstructing the speech imagery by transforming the listened speech. We used a linear mapping for each subject and replicated the results discussed above for the EEG experiments. Furthermore, we attempt to transfer a trained model from neural processing of one type of stimuli like music to another type like speech. We showed that linear mapping could be trained regardless of the stimuli' type, and we could still reconstruct imagery signals for a stimulus that was never used in the model training. These results imply that the relation between neural mechanisms of auditory imagery and listening is independent of the stimuli nature and that one could train a general mapping based on any stimulus during listening tasks and reconstruct imagery neural signals for a different type of stimuli.

To summarize, the overall findings from these experiments suggest that given precisely aligned imagery and listening signals for sufficiently large data-set, there exist a trainable universal mapping that is able to predict the imagery signals from any listening tasks. This framework enables future studies where one may use such a mapping along with an speech recognizer to reconstruct stimuli from imagery signals.

Chapter 6: **Conclusion and Future Directions**

6.1 What is Next?

The human brain is a highly complex system, consisting of 10^{11} number of neurons (units of computing processors) interacting with each other via 10^{15} connections [144]. The brain regions are highly interconnected, with each region receiving feedforward and a lot more feedback inputs from other regions. Addressing any question regarding this complicated object requires lots of observations as well as *useful models* – as *"all models are wrong, but some are useful"*. This thesis mainly focused on the problems in the human auditory system. The perception of sound has been extensively explored in humans and other animals over the last few decades; however, there is still a lot to discover in this field of research. In the present thesis, we attempted to decode the brain signals collected non-invasively through a series of EEG and MEG experiments and push our understanding regarding the perception of complex auditory objects in complex auditory environments. We investigated top-down modulations on the perception of sounds, such as the effect of attention on perceiving sound streams, the effect of memory on perceiving the patterns in melodies, and finally, the effect of memory and internal models on imagining (and thus perceiving) sound sequences like music and speech.

In chapter 3, we introduced a novel experimental paradigm in which we exploited the effect of attention on the brain response to a single-frequency tone to gain a view of the responses to fre-

quency components within complex mixtures. Specifically, we monitored the modulatory effects of ongoing and persistent attention on the responses to a probe tone to measure the enhancement and suppression of the complex components. At the same time, subjects were instructed to selectively attend to various target streams. With this technique, we showed that the response modulations are consistent with the predictions of temporal coherence and thus provide evidence for how the brain can perceptually segregate complex mixtures online. Also, using this paradigm, one might study the effect of attention on a different sound attribute or explore the binding of individual frequency components in a different context. For example, speech consists of both voiced and unvoiced sounds, and less is known about whether and how the unvoiced portions (such as fricatives) are segregated. Some recent evidence observed that listeners use both the consonant and vowel portions for speech segregation, suggesting that frequency components of the fricatives are also grouped and bind perceptually with other voiced components of speech [145–147]. Using the probe-tone paradigm, explained in chapter 3, we measured the MEG responses of 19 participants to two frequency components aligned with the peak fricative frequency of male ($6000Hz$) and female ($4000Hz$) voices. Our preliminary results suggest a strong effect of attention on both frequencies, indicating binding the frequency components of fricatives with speech (Figure 6.1).

The brain tracks auditory streams by forming expectations through learning patterns and statistical structures. At the finer temporal resolutions, this is indeed the principle of temporal coherence. For example, if the statistics of the stimuli change abruptly, the brain can no longer track and thus cannot perceive it as a single stream. On the slower time scales, statistical learning translates to storing the slower patterns in memory, such as learning the probability of transitions between notes in western music [148]. In chapter 4, we explored the brain's ability to learn

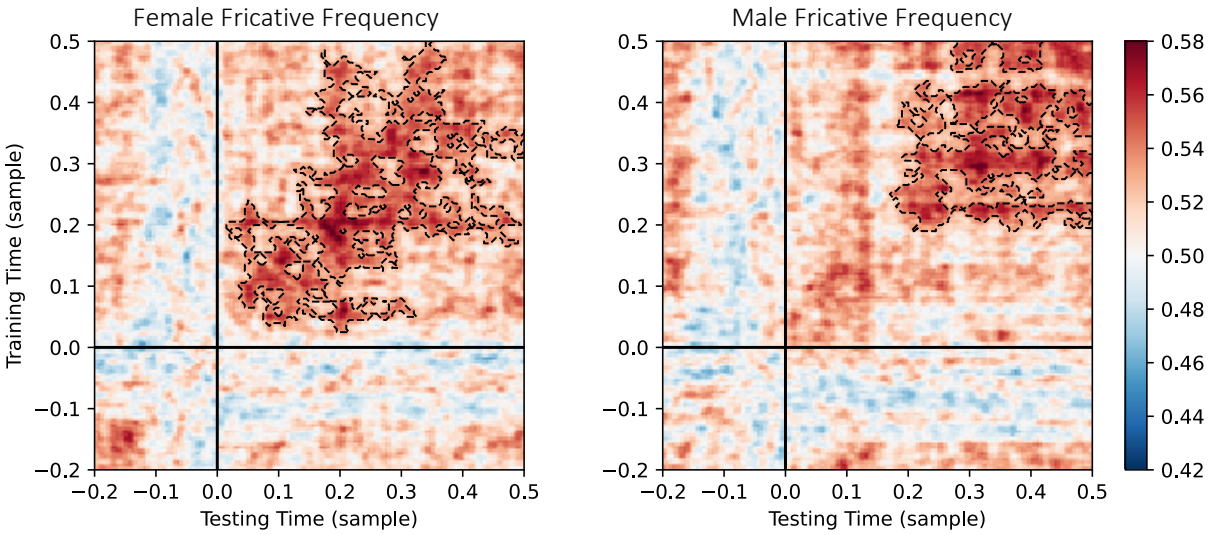

Figure 6.1: *Binding the frequency components of fricatives with speech.* The stimulus consisted of alternating male and female voices; each uttering random syllables includes a fricative and a vowel. The participants were asked to pay attention to one of the streams and detect a repetition in the target sound. We collected MEG signals from 19 subjects and performed the attention decoding analysis on the MEG responses to the probe tones presented 160 ms after the main sequence offset. Decoding performance for female (left; at 4000 Hz) and male (right; at 6000 Hz) probe tones. Classifiers were trained and tested separately at each time in a 700 ms time window of the probe-tone (-200ms to 500ms). The significant cluster is contoured with a dashed line. The classifiers could decode attention for both frequencies.

statistical patterns in different contexts using music stimuli. We generated melodies using a fixed artificial grammar derived from either a uniform or non-uniform scale. After an exposure phase, we showed that the brain learns the grammar only in the context of the non-uniform scales; conversely, we did not observe any evidence where the brain learns statistics of an identical grammar derived from a uniform scale. Therefore, this provides confirmation for why almost all known musical scales display intervals of different sizes (whole tones and semitones in the Western system), which are positioned within the octave in a way that maximizes *uniqueness* through non-uniform structures. A fundamental component of the rhythmic structure of music is its probabilistic relation between the events, which can be modeled as expectation strengths. The

regularities of music prompt our brain to build such expectations, which are accurately estimated by computational models of the musical structure [149], allowing us to assess the precise neural encoding of music expectations. In line with probabilistic learning of musical grammar, one might model and study diversity in individuals' musical enculturation, reflecting cultural effects on music perceptions. For example, one might model the statistical structures of Chinese and European music [150], and use these models to predict the brain responses of a listener novice to Chinese music. Presumably, the prediction power of the Chinese model should be poorer than the European model; however, the brain learns the statistical patterns as the listener is more exposed to Chines music, and as a result, the brain's responses to the Chinese music are predictable from the Chinese model, and it should get closer to the European model.

Music is an elaborate symbolic system conveyed via acoustic signals, and its appreciation involves several hierarchical levels of processing. This hierarchy starts with processing primary perceptual attributes, such as pitch, timbre, intensity, and location, which are encoded at or before the primary auditory cortex [151, 152]. Consistent with our findings in chapter 4, higher-order rules of grammar and engagement are then supposedly extracted and represented in secondary auditory areas, and other associative regions [153,154]. As we discussed above, forming expectations from this grammar, presumably, is how the listeners interact and engage with music. Recent evidence suggests that melodic expectations play a role in music imagery [137]. In chapter 5, we build on this assumption where melodic expectations (and other higher-level music processing) have shared neural representations during music listening and imagery. We developed a mapping between the listening and imagery neural signals (collected with EEG) and showed that using the reconstructed imagery signals, we could decode which melody was presented in the mind of the participants during the imagery tasks. Like music, studies showed that probabilistic relations

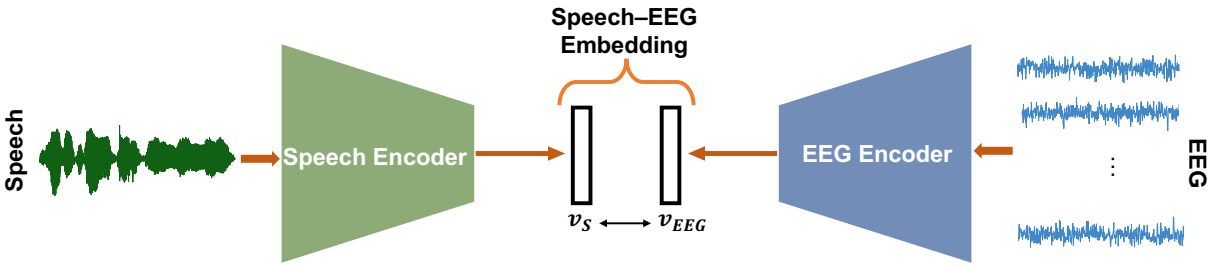

Figure 6.2: *Building a Speech Recognizer for Speech Imagery.* The speech encoder transforms the speech waveforms into a semantic space (one might use any off-the-shelf speech recognizer like [157]). The EEG encoder network does a similar transformation for the predicted EEG signals. Therefore, the trained network has a set of weights, with which the EEG signals are projected onto some semantic space. One might use this projection to decode the imagined speech.

between speech's elements are also encoded in secondary auditory areas and other associative regions [155, 156]. Therefore, the discussion regarding music perception we made above might also apply to speech, where the higher processing stage of speech, such as representations of the statistical structures of phonemes and words, are common during speech listening and imagery. This indeed, may be why we could find a similar mapping for the speech in chapter 5.

One intriguing future prospect for this study is to apply the proposed mapping we developed earlier in order to predict the imagery EEG (or MEG) signals – note that this can be done on the collected listened EEG alone. Then, one might develop a speech recognition model on these predicted signals, using a large corpus of speech and neural data, and read the *inner speech* using the brain signals of imagery. We propose, for example, a framework depicted in Figure 6.2, in which two separate network is trained. The speech and EEG encoders transform the speech and the predicted EEG data, respectively, into two embedding vectors in a Speech-EEG embedding space. The objective of this network is to find a set of weights where the two embedding vectors have the smallest possible distance for the aligned EEG and speech data. Therefore, one might

use this speech recognizer to decode the EEG imagery signals for a much larger set of words and sentences.

Bibliography

[1] Rodolfo Llinás and Urs Ribary. Consciousness and the brain: The thalamocortical dialogue in health and disease. In *Annals of the New York Academy of Sciences*, volume 929, pages 166–175. John Wiley & Sons, Ltd, jan 2001.

[2] J.O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, 1982.

[3] E. Kandel, J. Schwartz, J.H. Jessell, S. Mack, T. Jessell, and J. Dodd. *Principles of Neural Science, Fourth Edition*. McGraw-Hill Companies,Incorporated, 2000.

[4] Bradley R. Postle. *Essentials of Cognitive Neuroscience*. Wiley, 2016.

[5] J. Donald Hams. Bases of hearing science. *Ear and Hearing*, 6(2), 1985.

[6] Steven SenYantis. *Sensation and Perception*. New York, NY: Worth Publishers., 2014.

[7] Taishih Chi, Powen Ru, and Shihab Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118:887–906, 09 2005.

[8] Klaas E Stephan, Claus C Hilgetag, Gully A.P.C. Burns, Marc A O'Neill, Malcolm P Young, and Rolf Kötter. Computational analysis of functional connectivity between areas of primate cerebral cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1393):111–126, jan 2000.

[9] R. Meddis, E. Lopez-Poveda, R.R. Fay, and A.N. Popper. *Computational Models of the Auditory System*. Springer Handbook of Auditory Research. Springer US, 2010.

[10] D.L. Wang, G.S. Brown, and G.J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006.

[11] L F Haas. Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(1):9–9, 2003.

[12] Dale Purves and Elizabeth M Brannon. *Principles of Cognitive Neuroscience*. Sunderland, Sinauer Associates, Publishers, 2013.

[13] S.J. Luck. *An Introduction to the Event-Related Potential Technique, second edition*. A Bradford Book. MIT Press, 2014.

[14] Terence Picton, S.A. Hillyard, H.I. Krausz, and R Galambos. Human auditory evoked potentials. i: Evaluation of components. *Electroencephalography and clinical neurophysiology*, 36:179–90, 03 1974.

[15] Risto Näätänen and Terence Picton. The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology*, 24(4):375–425, jul 1987.

[16] Timothy Roberts, Ferrari Paul, Steven Stufflebeam, and David Poeppel. Latency of the auditory evoked neuromagnetic field components: Stimulus dependence and insights toward perception. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 17:114–29, 04 2000.

[17] L. McEvoy, J.P. Mäkelä, M. Hämäläinen, and R. Hari. Effect of interaural time differences on middle-latency and late auditory evoked magnetic fields. *Hearing Research*, 78(2):249–257, aug 1994.

[18] Juha Pekka Vasama and Jyrki P. Mäkelä. Auditory pathway plasticity in adult humans after unilateral idiopathic sudden sensorineural hearing loss. *Hearing Research*, 87(1-2):132–140, jul 1995.

[19] Bernd Lütkenhöner and Olaf Steinsträter. High-precision neuromagnetic study of the functional organization of the human auditory cortex. *Audiology and Neuro-Otology*, 3(2-3):191–213, 1998.

[20] Janis E Oram Cardy, Paul Ferrari, Elissa J Flagg, Wendy Roberts, and Timothy P.L. Roberts. Prominence of M50 auditory evoked response over M100 in childhood and autism. *NeuroReport*, 15(12):1867–1870, aug 2004.

[21] Janis E Oram Cardy, Elissa J Flagg, Wendy Roberts, and Timothy P.L. Roberts. Auditory evoked fields predict language ability and impairment in children. *International Journal of Psychophysiology*, 68(2):170–175, may 2008.

[22] Albert Bregman. Auditory scene analysis: The perceptual organization of sound. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, 95:250, 01 1990.

[23] Brian Moore and Hedwig Gockel. Factors influencing sequential stream segregation. *Acta Acustica united with Acustica*, 88:320–333, 05 2002.

[24] John C Middlebrooks, Jonathan Z Simon, Arthur N Popper, and Richard R Fay Editors. *The Auditory System at the Cocktail Party*, volume 60. Springer Handbook of Auditory research, 2017.

[25] Douglas S. Brungart, Brian D. Simpson, Mark A. Ericson, and Kimberly R. Scott. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, 110(5):2527–2538, 2001.

[26] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J. Oxenham, and Shihab A. Shamma. Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, 61(2):317–329, 2009.

[27] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3):114–123, 2011.

[28] Kai Lu, Yanbo Xu, Pingbo Yin, Andrew J. Oxenham, Jonathan B. Fritz, and Shihab A. Shamma. Temporal coherence structure rapidly shapes neuronal interactions. *Nature Communications*, 8:13900, 2017.

[29] Christophe Micheyl, Heather Kreft, Shihab Shamma, and Andrew J. Oxenham. Temporal coherence versus harmonicity in auditory stream formation. *The Journal of the Acoustical Society of America*, 133(3):EL188–EL194, 2013.

[30] Christophe Micheyl, Coral Hanson, Laurent Demany, Andrew J. Oxenham, and Shihab Shamma. Auditory stream segregation for alternating and synchronous tones. *Journal of Experimental Psychology: Human Perception and Performance*, 39(6):1568–1580, 2013.

[31] Albert S. Bregman and Jeffrey Campbell. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89(2):244–249, 1971.

[32] Soqajan Zera. Detectisoqahng temporal onset and offset asynchrony in multicomponent complexes. *Journal of the Acoustical Society of America*, 93(2):1038–1052, 1993.

[33] Jan Zera and David M. Green. Effect of signal component phase on asynchrony discrimination. *The Journal of the Acoustical Society of America*, 98(2):817–827, 1995.

[34] Sundeep Teki, Maria Chait, Sukhbinder Kumar, Shihab Shamma, and Timothy D Griffiths. Segregation of complex acoustic scenes based on temporal coherence. *eLife*, 2013(2), jul 2013.

[35] James A. O'Sullivan, Shihab A. Shamma, and Edmund C. Lalor. Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *Journal of Neuroscience*, 35(18):7256–7263, 2015.

[36] Sundeep Teki, Nicolas Barascud, Samuel Picard, Christopher Payne, Timothy D. Griffiths, and Maria Chait. Neural Correlates of Auditory Figure-Ground Segregation Based on Temporal Coherence. *Cerebral Cortex (New York, NY)*, 26(9):3669, 2016.

[37] James O'Sullivan, Jose Herrero, Elliot Smith, Catherine Schevon, Guy M. McKhann, Sameer A Sheth, Ashesh D Mehta, and Nima Mesgarani. Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron*, 104(6):1195–1209.e3, 2019.

[38] Lakshmi Krishnan, Mounya Elhilali, and Shihab Shamma. Segregating Complex Sound Sources through Temporal Coherence. *PLoS Computational Biology*, 10(12), 2014.

[39] Jennifer K. Bizley, Ross K. Maddox, and Adrian K.C. Lee. Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. *Trends in Neurosciences*, 39(2):74–85, feb 2016.

[40] Huriye Atilgan, Stephen Town, Katherine Wood, Gareth Jones, Ross Maddox, Adrian KC Lee, and Jennifer Bizley. Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron*, 97:640–655, 01 2018.

[41] Huriye Atilgan and Jennifer K. Bizley. Training enhances the ability of listeners to exploit visual information for auditory scene analysis. *Cognition*, 208, mar 2021.

[42] Simon Krogholt Christiansen and Andrew J. Oxenham. Assessing the effects of temporal coherence on auditory stream formation through comodulation masking release. *The Journal of the Acoustical Society of America*, 135(6):3520–3529, 2014.

[43] Elyse S. Sussman. Auditory scene analysis: An attention perspective. *Journal of Speech, Language, and Hearing Research*, 60(10):2989–3000, 2017.

[44] David H Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436, 1997.

[45] Denis G Pelli. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10(4):437–442, 1997.

[46] Mario Kleiner, David H Brainard, Denis G Pelli, Christopher Broussard, Tobias Wolf, and Diederick Niehorster. What's new in Psychtoolbox-3? A free cross-platform toolkit for psychophysiscs with Matlab and GNU/Octave. Technical report, 2007.

[47] Alain de Cheveigné and Dorothée Arzounian. Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *NeuroImage*, 172:903–912, may 2018.

[48] Alain de Cheveigné and Jonathan Z. Simon. Denoising based on time-shift PCA. *Journal of Neuroscience Methods*, 165(2):297–305, sep 2007.

[49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[50] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7:267, 2013.

[51] J-R. King and S. Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210, apr 2014.

[52] Alain De Cheveigné and Jonathan Z. Simon. Denoising based on spatial filtering. *Journal of Neuroscience Methods*, 171(2):331–339, 2008.

[53] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg- and meg-data. *Journal of Neuroscience Methods*, 164(1):177 – 190, 2007.

[54] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[55] Robert Carlyon, Sarah Thompson, Antje Heinrich, Friedemann Pulvermüller, Matthew Davis, Yury Shtyrov, Rhodri Cusack, and Ingrid Johnsrude. *Objective Measures of Auditory Scene Analysis*, chapter 47, pages 507–519. Springer, New York, NY, 01 2010.

[56] Christophe Micheyl and Andrew J Oxenham. Objective and subjective psychophysical measures of auditory stream integration and segregation. *JARO - Journal of the Association for Research in Otolaryngology*, 11(4):709–724, 2010.

[57] Punita G. Singh. Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre. *Journal of the Acoustical Society of America*, 82(3):886–899, 1987.

[58] N Grimault, Christophe Micheyl, R.P. Carlyon, P Arthaud, and Lionel Collet. Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing. *British journal of audiology*, 35:173–82, 07 2001.

[59] Juanjuan Xiang, Jonathan Simon, and Mounya Elhilali. Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *Journal of Neuroscience*, 30(36):12084–12093, 2010.

[60] Alan Power, John Foxe, Emma-Jane Forde, Richard Reilly, and Edmund Lalor. At what time is the cocktail party? a late locus of selective attention to natural speech. *The European journal of neuroscience*, 35:1497–503, 03 2012.

[61] Inyong Choi, Siddharth Rajaram, Lenny Varghese, and Barbara Shinn-Cunningham. Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in human neuroscience*, 7:115, 04 2013.

[62] Mark Stokes, Michael Wolff, and Eelke Spaak. Decoding rich spatial information with high temporal resolution. *Trends in Cognitive Sciences*, 19:636–638, 11 2015.

[63] Michael Wolff, Janina Jochim, Elkan Akyürek, and Mark Stokes. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, 20, 04 2017.

[64] Jean-Rémi King, Niccolo Pescetelli, and Stanislas Dehaene. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron*, 92:1122–1134, 12 2016.

[65] Pedro Pinheiro-Chagas, Manuela Piazza, and Stanislas Dehaene. Decoding the processing stages of mental arithmetic with magnetoencephalography. *Cortex*, jul 2018.

[66] Alain De Cheveigné and Lucas C. Parra. Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage*, 98:487–505, 2014.

[67] Shihab Shamma and Mounya Elhilali. *Chapter: Binding, Scene Analysis and Higher Cortical Centers in The Senses: A Comprehensive Reference, 2nd Edition*, chapter 2. ELSEVIER ACADEMIC Press, 2020.

[68] Robert Bolia, W. Nelson, Mark Ericson, and Brian Simpson. A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, 107:1065–6, 03 2000.

[69] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–6, 2012.

[70] Joel Snyder, Claude Alain, and Terence Picton. Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of cognitive neuroscience*, 18:1–13, 02 2006.

[71] Nai Ding and Jonathan Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012.

[72] Lynne Bernstein, Edward Auer, and Sumiko Takayanagi. Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, 44:5–18, 10 2004.

[73] Michael J. Crosse, Giovanni M. Di Liberto, and Edmund C. Lalor. Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, 36(38):9888–9895, 2016.

[74] Aisling E. O'Sullivan, Michael J. Crosse, Giovanni M. Di Liberto, Alain de Cheveigné, and Edmund C. Lalor. Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. *Journal of Neuroscience*, 41(23):4991–5003, 2021.

[75] Pierre Perruchet and Sébastien Pacton. Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in cognitive sciences*, 10:233–8, 06 2006.

[76] Morten H. Christiansen. Implicit Statistical Learning: A Tale of Two Literatures. *Topics in Cognitive Science*, 11(3):468–481, jul 2019.

[77] P. Rebuschat. *Implicit and Explicit Learning of Languages*. Studies in Bilingualism. John Benjamins Publishing Company, 2015.

[78] Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, and Matthias Robine. Automatic timbre classification of ethnomusicological audio recordings. 2014.

[79] Rainer Polak, Nori Jacoby, Timo Fischinger, Daniel Goldberg, Andre Holzapfel, and Justin London. Rhythmic prototypes across cultures: a comparative study of tapping synchronization. *Music Perception: An Interdisciplinary Journal*, 36(1):1–23, 2018.

[80] Samuel A Mehr, Manvir Singh, Dean Knox, Daniel M Ketter, Daniel Pickens-Jones, S Atwood, Christopher Lucas, Nori Jacoby, Alena A Egner, Erin J Hopkins, et al. Universality and diversity in human song. *Science*, 366(6468), 2019.

[81] Patrick Savage, Steven Brown, Emi Sakai, and Thomas Currie. Statistical universal reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 06 2015.

[82] Alexander John Ellis. *On the musical scales of various nations*. Journal of the Society of Arts, 1885.

[83] Richmond Browne. Tonal implications of the diatonic set. *In Theory Only*, 5(6):3–21, 1981.

[84] Bruno Nettl. An ethnomusicologist contemplates universals in musical sound and musical culture. *The Origins of Music*, 3(2):463–472, 2000.

[85] Claire Pelofi and Morwaread M Farbood. Asymmetry in scales enhances learning of new musical structures. *Proceedings of the National Academy of Sciences*, 118(31), 2021.

[86] Martin Rohrmeier, Patrick Rebuschat, and Ian Cross. Incidental and online learning of melodic structure. *Consciousness and Cognition*, 20(2):214–222, 2011.

[87] Michael J Crosse, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor. The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, 10:604, 2016.

[88] David Benedict Reck. *Music of the whole earth*. Macmillan Reference USA, 1977.

[89] Catherine Stevens. Cross-cultural studies of musical pitch and time. *Acoustical science and technology*, 25(6):433–438, 2004.

[90] Bruno Nettl. *The study of ethnomusicology: Thirty-three discussions*. University of Illinois Press, 2015.

[91] Mary A Castellano, Jamshed J Bharucha, and Carol L Krumhansl. Tonal hierarchies in the music of north india. *Journal of Experimental Psychology: General*, 113(3):394, 1984.

[92] Carol L. Krumhansl, Pekka Toivanen, Tuomas Eerola, Petri Toiviainen, Topi Järvinen, and Jukka Louhivuori. Cross-cultural music cognition: cognitive methodology applied to north sami yoiks. *Cognition*, 76(1):13–58, 2000.

[93] Patrick E Savage, Psyche Loui, Bronwyn Tarr, Adena Schachner, Luke Glowacki, Steven Mithen, and W Tecumseh Fitch. Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, 44, 2021.

[94] Steven Brown. Evolutionary models of music: From sexual selection to group selection. In *Perspectives in ethology*, pages 231–281. Springer, 2000.

[95] Ian Cross and IRM Morley. The evolution of music: Theories, definitions and the nature of the evidence. 2010.

[96] Ellen Dissanayake. Root, leaf, blossom, or bole: Concerning the origin and adaptive function of music. *Communicative musicality: Exploring the basis of human companionship*, pages 17–30, 2009.

[97] Nathan Oesch. Music and language in social interaction: Synchrony, antiphony, and functional origins. *Frontiers in Psychology*, page 1514, 2019.

[98] Jay Schulkin and Greta B Raglan. The evolution of music and human social capability. *Frontiers in neuroscience*, 8:292, 2014.

[99] Sandra E Trehub, Judith Becker, and Iain Morley. Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140096, 2015.

[100] Patrick E Savage, Steven Brown, Emi Sakai, and Thomas E Currie. Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, 112(29):8987–8992, 2015.

[101] David Butler and Helen Brown. Tonal structure versus function: Studies of the recognition of harmonic motion. *Music Perception*, 2(1):6–24, 1984.

[102] Sahar Akram, Alessandro Presacco, Jonathan Z Simon, Shihab A Shamma, and Behtash Babadi. Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling. *NeuroImage*, 124:906–917, 2016.

[103] Daniel DE Wong, S Fuglsang, Jens Hjortkjær, Enea Ceolini, Malcolm Slaney, and Alain de Cheveigné. A comparison of temporal response function estimation methods for auditory attention decoding. *Biorxiv*, pages 1–22, 2018.

[104] Michael P Broderick, Andrew J Anderson, and Edmund C Lalor. Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience*, 39(38):7564–7575, 2019.

[105] Giovanni M Di Liberto, James A O'Sullivan, and Edmund C Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.

[106] Giovanni M Di Liberto, Claire Pelofi, Shihab Shamma, and Alain de Cheveigné. Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology*, 41(1):361–364, 2020.

[107] Giovanni M Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D Mehta, Jose L Herrero, Alain de Cheveigné, Shihab Shamma, and Nima Mesgarani. Cortical encoding of melodic expectations in human temporal cortex. *eLife*, 9:e51784, 2020.

[108] Gerald J Balzano. The pitch set as a level of description for studying musical pitch perception. In *Music, Mind, and Brain*, pages 321–351. Springer, 1982.

[109] Jean Mary Zarate, Caroline R Ritson, and David Poeppel. Pitch-interval discrimination and musical expertise: Is the semitone a perceptual boundary? *The Journal of the Acoustical Society of America*, 132(2):984–993, 2012.

[110] Nori Jacoby and Josh H McDermott. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3):359–370, 2017.

[111] Josh H McDermott, Alan F Schultz, Eduardo A Undurraga, and Ricardo A Godoy. Indifference to dissonance in native amazonians reveals cultural variation in music perception. *Nature*, 535(7613):547–550, 2016.

[112] Nori Jacoby, Eduardo A Undurraga, Malinda J McPherson, Joaquín Valdés, Tomás Ossandón, and Josh H McDermott. Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, 29(19):3229–3243, 2019.

[113] Grant Ramsey. The fundamental constraint on the evolution of culture. *Biology & Philosophy*, 22(3):401–414, 2007.

[114] Philip V Bohlman. Ontologies of music. *Rethinking Music*, pages 17–34, 1999.

[115] Noam Chomsky. Aspects of the theory of syntax (vol. 11). *MIT Press. doi*, 10:90008–5, 1965.

[116] Roman Jakobson. Implications of language universals for linguistics. *Universals of Language*, pages 263–278, 1963.

[117] Thomas G Bever. The cognitive basis for linguistic structures. *Cognition and the Development of Language*, 279(362):1–61, 1970.

[118] Stephen Kosslyn, Giorgio Ganis, and William Thompson. New foundations of imagery. *Nature reviews. Neuroscience*, 2:635–42, 10 2001.

[119] David F. Marks. Consciousness, mental imagery and action. *British Journal of Psychology*, 90(4):567–585, 1999.

[120] Hanneke E.M. Den Ouden, Peter Kok, and Floris P. de Lange. How prediction errors shape perception, attention, and motivation, 2012.

[121] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: Knowns and unknowns, sep 2013.

[122] David J. Heeger. Theory of cortical function. *Proceedings of the National Academy of Sciences of the United States of America*, 114(8):1773–1782, feb 2017.

[123] M W Spratling. A review of predictive coding algorithms. *Brain and cognition*, 112:92–97, 2017.

[124] P. K. McGuire, D. A. Silbersweig, R. M. Murray, A. S. David, R. S. J. Frackowiak, and C. D. Frith. Functional anatomy of inner speech and auditory verbal imagery. *Psychological Medicine*, 26(1):29–38, jan 1996.

[125] Robert J Zatorre, Andrea R Halpern, David W Perry, Ernst Meyer, and Alan C Evans. Hearing in the mind's ear: A PET investigation of musical imagery and perception. *Journal of Cognitive Neuroscience*, 8(1):29–46, 1996.

[126] Andrea R. Halpern and Robert J. Zatorre. When that tune runs through your head: A PET investigation of auditory imagery for familiar melodies. *Cerebral Cortex*, 9(7):697–704, oct 1999.

[127] S. S. Shergill, E. T. Bullmore, M. J. Brammer, S. C.R. Williams, R. M. Murray, and P. K. McGuire. A functional study of auditory verbal imagery. *Psychological Medicine*, 31(2):241–253, feb 2001.

[128] Andrea R Halpern, Robert J Zatorre, Marc Bouffard, and Jennifer A Johnson. Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, 42(9):1281–1292, 2004.

[129] David J.M. Kraemer, C. Neil Macrae, Adam E. Green, and William M. Kelley. Sound of silence activates auditory cortex. *Nature*, 434(7030):158, mar 2005.

[130] André Aleman, Elia Formisano, Heidi Koppenhagen, Peter Hagoort, Edward H.F. De Haan, and René S. Kahn. The functional neuroanatomy of metrical stress evaluation of perceived and imagined spoken words. *Cerebral Cortex*, 15(2):221–228, jul 2005.

[131] Yizhen Zhang, Gang Chen, Haiguang Wen, Kun Han Lu, and Zhongming Liu. Musical Imagery Involves Wernicke's Area in Bilateral and Anti-Correlated Network Interactions in Musicians. *Scientific Reports*, 7(1):17066, dec 2017.

[132] B. Kleber, N. Birbaumer, R. Veit, T. Trevorrow, and M. Lotze. Overt and imagined singing of an Italian aria. *NeuroImage*, 36(3):889–900, jul 2007.

[133] Marina Papoutsi, Jacco A De Zwart, J. Martijn Jansma, Martin J Pickering, James A Bednar, and Barry Horwitz. From phonemes to articulatory codes: An fMRI study of the role of broca's area in speech production. *Cerebral Cortex*, 19(9):2156–2165, 2009.

[134] Xing Tian, Jean Mary Zarate, and David Poeppel. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, 77:1–12, apr 2016.

[135] Sibylle C Herholz, Andrea R Halpern, and Robert J Zatorre. Neuronal correlates of perception, imagery, and memory for familiar tunes. *Journal of Cognitive Neuroscience*, 24(6):1382–1397, jun 2012.

[136] César Lima, Nadine Lavan, S Evans, Zarinah Agnew, Andrea Halpern, Pradheep Shanmugalingam, Sophie Meekings, D Boebinger, Markus Ostarek, Carolyn Mcgettigan, Jane Warren, and Sophie Scott. Feel the noise: Relating individual differences in auditory imagery to the structure and function of sensorimotor systems. *Cerebral Cortex*, 25, 11 2015.

[137] Guilhem Marion, Giovanni M. Di Liberto, and Shihab A Shamma. The music of silence. Part I: Responses to musical imagery encode melodic expectations and acoustics. *Journal of Neuroscience*, 41(35), 2021.

[138] Giovanni M. Di Liberto, Guilhem Marion, and Shihab A Shamma. The music of silence. Part II: Music listening induces imagery responses. *Journal of Neuroscience*, 41(35):7449–7460, 2021.

[139] Stéphanie Martin, Peter Brunner, Chris Holdgraf, Hans Jochen Heinze, Nathan E. Crone, Jochem Rieger, Gerwin Schalk, Robert T. Knight, and Brian N. Pasley. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7(MAY):14, may 2014.

[140] Sari Ylinen, Anni Nora, Alina Leminen, Tero Hakala, Minna Huotilainen, Yury Shtyrov, Jyrki P. Mäkelä, and Elisabet Service. Two distinct auditory-motor circuits for monitoring speech production as revealed by content-specific suppression of auditory cortex. *Cerebral Cortex*, 25(6):1576–1586, jun 2015.

[141] Xing Tian, Nai Ding, Xiangbin Teng, Fan Bai, and David Poeppel. Imagined speech influences perceived loudness of sound. *Nature Human Behaviour*, 2(3):225–234, mar 2018.

[142] Christian Brodbeck, Teon L Brooks, Proloy Das, Samir Reddigari, and jpkulasingham. christianbrodbeck/eelbrain: 0.37, April 2022.

[143] Stephen David, Nima Mesgarani, and Shihab Shamma. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, 18:191–212, 10 2007.

[144] Sarah DeWeerdt. How to map the brain. *Nature*, 571(7766):S6–S8, July 2019.

[145] Marion David, Mathieu Lavandier, Nicolas Grimault, and Andrew J. Oxenham. Discrimination and streaming of speech sounds based on differences in interaural and spectral cues. *The Journal of the Acoustical Society of America*, 142(3):1674–1685, sep 2017.

[146] Marion David, Alexis N. Tausend, Olaf Strelcyk, and Andrew J. Oxenham. Effect of age and hearing loss on auditory stream segregation of speech sounds. *Hearing Research*, 364:118–128, jul 2018.

[147] Marion David, Mathieu Lavandier, Nicolas Grimault, and Andrew J. Oxenham. Binding of speech syllables when segregation occurs. In *Proceedings of the International Congress on Acoustics*, volume 2019-Septe, pages 85–92, sep 2019.

[148] Laura J Batterink, Ken A Paller, and Paul J Reber. Understanding the Neural Bases of Implicit and Statistical Learning. *Topics in cognitive science*, 11(3):482–503, 2019.

[149] M T Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, 2005.

[150] Marcus T. Pearce. Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, pages 1–18, 2018.

[151] Stefan Koelsch and Walter A. Siebel. Towards a neural basis of music perception. *Trends in Cognitive Sciences*, 9(12):578–584, 2005.

[152] Petr Janata. Neural basis of music perception. In *Handbook of Clinical Neurology*, volume 129, pages 187–205. Handb Clin Neurol, 2015.

[153] Robert J. Zatorrea and Valorie N. Salimpoor. From perception to pleasure: Music and its neural substrates, jun 2013.

[154] Giovanni M Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D Mehta, Jose L Herrero, Alain de Cheveigné, Shihab Shamma, and Nima Mesgarani. Cortical encoding of melodic expectations in human temporal cortex. *eLife*, 9, 2020.

[155] Christian Brodbeck, L. Elliot Hong, and Jonathan Z Simon. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, 28(24):3976–3983.e5, 2018.

[156] Christian Brodbeck, Shohini Bhattasali, Aura Cruz Heredia, Philip Resnik, Jonathan Z. Simon, and Ellen Lau. Parallel processing in speech perception: Local and global representations of linguistic context. *bioRxiv*, page 2021.07.03.450698, jul 2021.

[157] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, jun 2020.