Carl von Ossietzky
Universität
Oldenburg

FROM ACOUSTIC ENVIRONMENTS TO NEURAL REPRESENTATIONS: INVESTIGATING NATURALISTIC SOUNDSCAPE PERCEPTION WITH EEG

von der Fakultät VI - Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines

Doktor der Naturwissenschaften (Dr.-rer. Nat.)

angenommene Dissertation von

THORGE HAUPT

geboren am 03.07.1997 in Delmenhorst

Prof. Dr. Martin G. Bleichner

Prof. Dr. Frederic Dehais

PD. Dr. Stephanie Rosemann

Tag der Disputation: 11.03.2026

ACKNOWLEDGMENTS

"What is a friend?
A single soul dwelling in two bodies."

Aristotle in (Diogenes Laertius and Yonge, 2006)

For many reasons, the last year was one of the most difficult years of my life. Apart from having to complete the biggest project I have ever created, aka this thesis, I had to deal with many personal changes. It was a year full of reflection, challenges, and new experiences. Above all, it was a year of personal growth. Like so many challenges we face in life, the most important lessons are only realized much later, after growth, unbeknownst to us, has already occurred. However, one lesson that I learned directly is the importance of having the right group of people around you. Without them, I would be nowhere near where I am now, and for that, I am beyond thankful.

To Martin, my thesis supervisor. I thank you for having the patience and guidance to navigate my rabbit hole tendencies and side quest explorations, to always lead me back on track. You have given me the freedom to explore my scientific curiosity, leading me into the depths of science that have undoubtedly shaped me as the scientist I am today. To Manu, the best office mate I have ever had. Always there to answer my questions about anything, inspiring me with your own curiosity. I thank you for always having an open ear to my rants about life, my singing, and coffee-fueled bouts of energy. To Stefan and the whole of the neuropsychology department. Thank you for teaching me to strive for and become a more critical and better scientist. For all the snacks that have tremendously contributed to the completion of this thesis, and most importantly, all the laughs for the notoriously bad puns this group has conjured up over the past four years. To Julius, Mareike, and Jules for providing much needed feedback on several drafts of this thesis. To my thesis reviewers, I thank you for taking the time that you have spent on reading this long work.

To my family. I am not sure I can even express in words the gratitude and love I feel towards you. Throughout my life, your presence was the only consistency in the chaos that is life, I could rely on. You are the most important thing in my life, and I cannot

thank you enough for supporting me throughout all the endeavors and adventures that I have gone on. Without you, none of this would have been possible, and for that, you have my eternal gratitude.

To my other family, though not related by blood, all my friends. You were there for me through all the tears of life. Those that were the product of experiencing the darkest moments, when life seemed bleak, to those rivers of joy that left me with a sore core from laughing so hard till breathing became my only purpose in life. The conversations that we have shared have taken me into the depths of the purpose of life and the universe, to the mundane, where we spent hours discussing whether pasta or rice is the best carbohydrate, and back. From sharing emotional parachutes, exploring life, and everything else, I cannot wait to see where we will go and what awaits us next. Without you, I would not be the person I am today, and for that, you have my thanks.

So, what I am trying to say is that I merely wrote down the words of this thesis, but you all are the reason it exists. This thesis belongs to you as much as it belongs to me. Unless, of course, there are errors, these are only my doing, for you all are the only thing that is right, for certain.

ABSTRACT

Neuroscientific investigations of auditory perception have traditionally been conducted in controlled laboratory settings, where participants are instructed to minimize movement and respond to artificial stimuli such as click trains or isolated tones. While this approach allows for precise experimental control, it limits the ecological validity of the findings, raising questions about how well they generalize to real-world listening scenarios. Everyday hearing involves continuously changing acoustic environments that contain overlapping sources and dynamically evolving sound events. Understanding auditory perception in such naturalistic soundscapes requires approaches that balance both experimental rigor and the use of naturalistic stimuli.

This thesis advances the study of naturalistic soundscape perception by developing and evaluating methods for analyzing neural responses to complex auditory environments using Electroencephalography (EEG). Across three studies, we investigated how acoustic and perceptual features, temporal context, and real-world recording conditions influence the neural representation of soundscapes.

In the first study, I examined how different categories of features, acoustic properties and perceptual meta-information, explain neural variability. Combining detailed acoustic features with sound identity (sound labels) information improved model predictions, demonstrating that incorporating perceptually relevant information enhances the ability to capture neural responses to complex auditory scenes.

The second study expanded this approach by integrating temporal dynamics, modeling Inter Onset Interval (IOI) intervals to investigate neural adaptation. These findings extend adaptation effects to complex naturalistic soundscapes, showing that sensitivity to temporal regularities emerges even without explicit perceptual labeling. The responses measured with EEG reflect large-scale cortical dynamics consistent with adaptation as a fundamental property of auditory processing.

Where the first two studies explored various aspects of the soundscape to explain neural variability, the third study evaluated whether minimal and mobile EEG hardware can be used to record neural activity Beyond the Lab (BTL). Participants performed an auditory attention task under both controlled laboratory and real-world conditions (sitting and walking). The results demonstrated reliable decoding of attention in stationary outside the lab compared to laboratory settings. In the BTL walking

condition, however, an artifactual response was detected in the walking condition. This highlights the need for refined preprocessing and artifact detection methods when measuring BTL situations.

Together, these studies provide a methodological framework for investigating how the brain processes complex acoustic environments. The findings show that EEG captures large-scale neural dynamics consistent with auditory cortical processing, where acoustic, contextual, and temporal features jointly shape neural responses. Ultimately, this work bridges controlled laboratory research and real-world neuroscience, offering insights into how auditory information is represented in the brain and informing future applications in neurotechnology and cognitive hearing science.

ZUSAMMENFASSUNG

Neurowissenschaftliche Untersuchungen der auditiven Wahrnehmung werden traditionell in kontrollierten Laborumgebungen durchgeführt, in denen die Teilnehmer angewiesen werden, ihre Bewegungen zu minimieren und auf künstliche Reize wie Klickfolgen oder isolierte Töne zu reagieren. Dieser Ansatz ermöglicht zwar eine präzise experimentelle Kontrolle, schränkt jedoch die ökologische Validität der Ergebnisse ein und wirft Fragen darüber auf, inwieweit sie sich auf reale Hörsituationen übertragen lassen. Das alltägliche Hören findet nämlich in sich ständig verändernden akustischen Umgebungen statt, in denen sich verschiedene Schallquellen überlagern und Klangereignisse dynamisch entwickeln. Um die auditive Wahrnehmung in solchen naturalistischen Klanglandschaften zu verstehen, sind Ansätze erforderlich, die sowohl experimentelle Genauigkeit als auch die Verwendung naturalistischer Reize berücksichtigen.

Diese Arbeit treibt die Erforschung der Wahrnehmung naturalistischer Klanglandschaften voran, indem sie Methoden zur Analyse neuronaler Reaktionen auf komplexe auditive Umgebungen unter Verwendung von Elektroenzephalographie (EEG) entwickelt und evaluiert. In drei Studien untersuchten wir, wie akustische und wahrnehmungsbezogene Merkmale, der zeitliche Kontext und reale Aufnahmebedingungen die neuronale Repräsentation von Klanglandschaften beeinflussen.

In der ersten Studie haben wir untersucht, wie verschiedene Kategorien von Merkmalen, akustische Eigenschaften und wahrnehmungsbezogene Metainformationen, neuronale Variabilität erklären. Die Kombination detaillierter akustischer Merkmale mit Informationen zur Klangidentität (Klangbezeichnungen) verbesserte die Modellvorhersagen und zeigte, dass die Einbeziehung wahrnehmungsrelevanter Informationen die Modelle verbessert, neuronale Reaktionen auf komplexe Hörszenen zu erfassen.

Die zweite Studie erweiterte diesen Ansatz durch die Integration zeitlicher Dynamiken und die Modellierung von dem Interval zwischen zwei Tönen, um die neuronale Anpassung an den akustischen Kontext zu untersuchen. Diese Ergebnisse erweitern bekannte Anpassungseffekte aus Laborstudien auf komplexe naturalistische Klanglandschaften und zeigen, dass die Empfindlichkeit für zeitliche Regelmäßigkeiten auch ohne explizite Wahrnehmungskennzeichnung auftritt. Die mit EEG gemessenen

Reaktionen spiegeln kortikale Dynamiken wider, die mit der Anpassung als grundlegender Eigenschaft der auditiven Verarbeitung übereinstimmen.

Während die ersten beiden Studien verschiedene Aspekte der Klanglandschaft untersuchten, um neuronale Variabilität zu erklären, wurde in der dritten Studie getestet, ob minimale und mobile EEG-Hardware zur Aufzeichnung neuronaler Aktivität außerhalb des Labors verwendet werden kann. Die Teilnehmer führten eine auditive Aufmerksamkeitsaufgabe sowohl unter kontrollierten Laborbedingungen als auch unter realen Bedingungen (sitzend und gehend) durch. Die Ergebnisse zeigten eine zuverlässige Dekodierung der Aufmerksamkeit im stationären Zustand außerhalb und im Labor. Beim Gehen, außerhalb des Labors, wurde jedoch eine artefaktische Reaktion festgestellt. Dies unterstreicht die Notwendigkeit verfeinerter Vorverarbeitungs- und Artefaktdetektionsmethoden bei der Messung von BTL-Situationen.

Zusammen bieten diese Studien einen methodischen Rahmen für die Untersuchung, wie das Gehirn komplexe akustische Umgebungen verarbeitet. Die Ergebnisse zeigen, dass das EEG großräumige neuronale Dynamiken erfasst, die mit der auditorischen kortikalen Verarbeitung übereinstimmen, bei der akustische, kontextuelle und zeitliche Merkmale gemeinsam neuronale Reaktionen formen. Letztendlich schlägt diese Arbeit eine Brücke zwischen kontrollierter Laborforschung und realer Neurowissenschaft und bietet Einblicke in die Darstellung auditorischer Informationen im Gehirn sowie Informationen für zukünftige Anwendungen in der Neurotechnologie und kognitiven Hörwissenschaft.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

A1	primary Auditory Cortex
AAD	Auditory Attention Decoding
AC	Acoustic Features
AEP	Auditory Evoked Potential
BCI	Brain-Computer Interfaces
BTL	Beyond the Lab
CNIC	Central Nucleus Inferior Colliculus
CP	Cognitive Priors
DCN	dorsal Cochlear Nucleus
dB	Dezibel
DNN	Deep Neural Net
ECoG	Electrocorticography
EEG	Electroencephalography
EPSP	Excitatory Postsynaptic Potential
ERP	Event-Related-Potential
FDR	False Discovery Rate
fMRI	functional Magnetic Resonance Imaging
fNIRS	functional Near-Infrared Spectroscopy
ICA	Independent Component Analysis
IC	Inferior Colliculus
iEEG	invasive EEG
IOI	Inter Onset Interval
IPSP	Inhibitory Postsynaptic Potential
ISI	Inter-Stimulus-Interval
LFP	Local Field Potential

LSL	Lab Streaming Layer
LTI	Linear Time-Invariant
MGB	Medial Geniculate Body
MGd	dorsal Medial Geniculate Body
MGm	medial Medial Geniculate Body
MGv	ventral Medial Geniculate Body
MEG	Magnetic Encephalography
MMN	Mismatched Negativity
MSE	Mean Squared Error
PCA	Principal Component Analysis
PHL	Portable Hearing Lab
PSP	Postsynaptic Potentials
RMS	Root-Mean-Squared
SI	Sound Identity
SNR	Signal-to-Noise Ratio
SOC	Superior Olivary Complex
SSA	Stimulus-Specific Adaptation
STG	Super Temporal Gyrus
TRF	Temporal Response Functions
VCN	ventral Cochlear Nucleus
VR	Virtual Reality

Part I

INTRODUCTION

This part introduces the theoretical foundations that have guided the development of my thesis. In addition to these conceptual considerations, I have included quotes at the beginning of each chapter. Drawn from both scientific and literary sources that have accompanied me over the past four years, these quotes serve as reflections of inspiration both scientifically and personally.

1

MOTIVATION

"How often do you hear a single sound by itself?
Only when doing psychoacoustic experiments
in a sound-proof booth!"

Darwin (2005)

I remember the exact moment when I got hooked on Neuroscience. It was while I was reading a chapter on visual processing for the course Brain and Behavior at Tilburg University during my Psychology Bachelor's. When learning how light falls onto the differently distributed rods and cones on the retina, and their receptive fields get stimulated, collective neural activity triggers more neural activity distributed around the brain to somehow create my visual perception. What struck me was the absurdity of thinking about how the very processes I was reading about were actively enabling me to read about those very processes. My fascination, which started with vision, was quickly extrapolated to the other senses, higher-order cognitive functions, emotions, and decision making. The core question that drove my interest in neuroscience is to understand how the brain constructs from all the sensory impressions a continuous and coherent everyday life experience.

What became apparent rather quickly, however, was that much of neuroscience, though precise and informative, studies the brain under conditions far removed from daily life. Controlled laboratory experiments often use stripped-down stimuli in carefully isolated environments, raising the question: do these neural responses generalize to the richly layered experiences of real-world perception?

My first experience with taking neuroscience out of the laboratory was during my master's in neuroeconomics, where I worked with Brain-Computer Interfaces (BCI). BCIs are designed to utilize brain activity to interact with the world by obtaining measurable brain states that can be distinguished (answering questions with yes and no, by imaging raising the left or right arm). Here, the core goal of BCI research is not to investigate those brain processes, but rather to find discernible brain states.

Thus, I joined the Everyday Life neurophysiology group, which focused on investigating auditory perception in natural environments, BTL, using minimal, wearable hardware. One question that encapsulated this line of research was: What happens in the brain when a seemingly irrelevant detail suddenly dominates perception? For instance, a classmate's casual remark, "Are you also annoyed by how often the professor says 'um'?", can shift your entire auditory experience of a lecture. Previously unnoticed sounds become impossible to ignore. The interesting question then becomes, what changed in the brain at that moment, and whether such a transition can be tracked in situ, with mobile neuroimaging?

Admittedly, audition wasn't my initial area of focus. However, I became increasingly intrigued by its selective and context-dependent nature, where certain sounds are filtered and do not reach conscious perception. What occurs in the brain when sounds enter awareness? Why am I not perceiving names being called at the cafe, but mine directs my attention instantly to the counter? Why do I sometimes have a tune randomly in my head, to then realize it is being played very softly somewhere? Is the solution to answer those questions to measure BTL? These are the questions that motivate this thesis. My aim is to contribute to bridging the gap between controlled laboratory settings and real-world perception by studying auditory processing in complex, dynamic soundscapes using mobile EEG. I hope to lay foundational steps towards neural measurement of auditory perception of everyday life.

2

AUDITORY PATHWAY

"Not a single one of the cells that compose you knows who you are, or cares."

Dennett (2006)

Key Takeaways

- **Pathway**: Cochlea - Brainstem - Inferior Colliculus - Thalamus - Auditory Cortex.
- **Increasing selectivity**: from basic frequency and timing to object-level representations, neural selectivity increases along the auditory pathway.
- **Dual pathway**: lemniscal (high fidelity), non-lemniscal (multimodal).
- **Multi-timescale adaptation**: adaptation to sound statistics occurs in every region along the pathway.

What you will learn:
How sound is decomposed and transformed into neural code, from outside the head to cortex-level processing.

This chapter provides a foundational overview of the anatomical and functional organization of the auditory pathway to support the interpretation of non-invasive neural recordings in later chapters. Understanding how sound information travels from the cochlea to cortical regions is essential for linking observed EEG responses to underlying neural processes. By detailing the structure and function of key auditory processing centers, from the brainstem to the auditory cortex, this chapter contextualizes how naturalistic, dynamic soundscapes are transformed into neural representations. The overview also highlights processing mechanisms such as tonotopy, temporal precision, auditory object formation, and adaptation to dynamic acoustic environments. Together, these insights establish a foundation to investigate the neural basis of real-world auditory perception.

2.1 ANATOMY

2.1.1 *Cochlea*

What we intuitively refer to as sound is not a physical object, but a perceptual experience constructed by the brain. This distinction is captured by the classic thought experiment: "Does a tree falling in a forest with no living organism around produce a sound?". The answer depends on how we define sound. If sound is the subjective experience of hearing, it only exists in the presence of a perceiver. In contrast, physical sound refers to the mechanical propagation of pressure waves through a medium, such as air. While it is difficult to determine at what point in time and space perception emerges, its precursor undoubtedly begins with these alternating compressions and rarefactions of air. These eventually reach our ears and trigger the process of auditory transduction (Plack, 2023).

The journey of these mechanical vibrations begins in the external ear, where the auricle captures and funnels pressure fluctuations into the external auditory canal. These fluctuations cause the tympanic membrane (eardrum) to vibrate, transmitting mechanical energy via the ossicles of the middle ear: the malleus, incus, and stapes. The stapes is connected to the oval window of the cochlea, the first stage of the inner ear, where mechanical energy is converted into electrical signals (Figure 1A) (Kandel et al., 2021, p. 599).

The cochlea is a spiral-shaped, fluid-filled structure organized tonotopically: different regions respond to different frequencies. High-frequency sounds stimulate the base of the cochlea, while low-frequency sounds travel further toward the apex. Complex sounds that reach the cochlea are thus decomposed into their frequency content, exciting the basilar membrane at different locations. This spatial gradient covers a frequency range from approximately 20 Hz to 20 kHz, with a logarithmic distribution of frequency mapping along the membrane (Figure 1B) (Greenwood, 1990; Kandel et al., 2021, pp. 601–602).

The mechanical excitation of the basilar membrane at the frequency-tuned location moves the inner hair cells of the organ of Corti, which sits atop the basilar membrane, against the tectorial membrane. This shearing motion opens mechanosensitive ion channels in the hair cells, resulting in depolarization (Fettiplace and Kim, 2014; Hudspeth, 1985). Each hair cell synapses onto several spiral ganglion neurons, which form the cochlear division of the vestibulocochlear nerve (cranial nerve VIII), also referred to as the auditory nerve.

Each hair cell and its associated ganglion cells have a characteristic frequency, the frequency at which they are most sensitive. The frequency tuning curves of these neurons are not strictly exclusive to a single frequency; they also respond to nearby frequencies. The responsiveness to other frequencies depends on stimulus intensity, meaning that with increasing loudness, tuning curves broaden (Robles and Ruggero, 2001). Some neurons have low response thresholds and are sensitive to quiet sounds, while others only respond at higher intensity levels, up to 100 Dezibel (dB). However, the firing rate of any given neuron saturates at very high intensities due to physiological limits, typically around 500 spikes per second (Kandel et al., 2021, pp. 622, 623). To process and convey information exceeding the physiological limit necessitates population coding (Heil, 2004). Here, a distributed response across multiple neurons conveys information about both sound intensity and frequency.

Ultimately, the auditory nerve transmits a rich, multidimensional signal to the brain, conveying detailed information about the frequency content, temporal structure, and intensity of incoming sounds, information that forms the basis of the information processing chain, which ultimately triggers neural activity that can be measured non-invasively.

2.1.2 *Brain Stem*

The auditory nerve terminates in the cochlear nucleus, the first relay station of the auditory brainstem. The cochlear nucleus is subdivided into dorsal Cochlear Nucleus (DCN) and ventral Cochlear Nucleus (VCN) regions. This structure follows the tonotopically organization of the cochlea, with low-frequency inputs projecting ventrally and high-frequency inputs represented dorsally (Figure 1C).

The VCN transmits spectral and temporal features of the sound signal. This distinction is enabled by different types of neurons that code specifically for spectral properties, on- and offsets, and interaural time differences of the sound signal (Joris, Schreiner, and Rees, 2004). The cells sensitive to spectral and interaural time differences project to the SOC, the first major site of binaural integration, which has been related to spatial decoding of sound sources (Grothe, Pecka, and McAlpine, 2010). Beyond the SOC's primary role of processing spatial cues, it also contributes to efferent feedback mechanisms that modulate cochlear sensitivity. For instance, the SOC projects back to both inner and outer hair cells, regulating their responsiveness and helping shape the auditory nerve's encoding of sound intensity (Guinan, 2006). The neurons

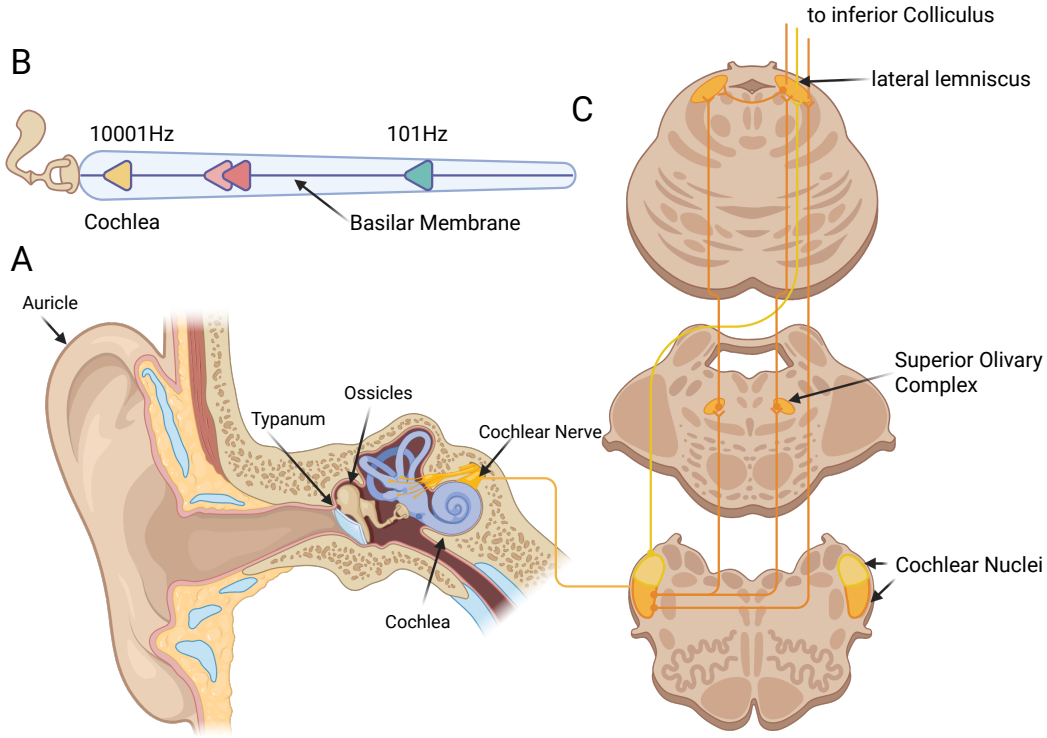

Figure 1: **A**: The human ear consists of the auricle, which channels sound into the ear canal, where the sound pressure sets the tympanic membrane into motion. The resulting mechanical energy is transmitted via the ossicles to the oval window. Here, the vibrations stimulate different sections of the cochlea depending on the spectral content. Mechanical energy is transduced into electrical signals, which are conveyed via the cochlear nerve to the cochlear nuclei. **B**: The cochlea maps the spectral content spatially. The base is sensitive to high- and the apex to low frequencies. Complex sounds are thus decomposed into their spectral components. **C**: Sections of the brain stem showing the afferent connections of the cochlear nuclei with the Superior Olivary Complex (SOC) and lateral lemniscus. Adapted from Kandel et al. (2021), with modifications. Created with BioRender.com.

then converge in the lateral lemniscus, which projects to the Inferior Colliculus (IC) (Figure 1C) (Antunes and Malmierca, 2021).

Going back to the DCN, it integrates, among others, spectral, vestibular, and somatosensory information. The DCN is involved in conveying information regarding sound localization to the IC via the lateral lemniscus. This allows for more accurate spatial mapping of sound sources and enables adaptive responses to changing acoustic environments. For instance, neurons in the dorsal cochlear nucleus can suppress responses to expected or self-generated sounds, reflecting an early form of predictive processing (Oertel and Young, 2004).

Together, these subdivisions of the cochlear nucleus initiate parallel processing streams that preserve and refine distinct aspects of the incoming sound signal for further integration in higher auditory centers.

2.1.3 *Inferior Colliculus*

The IC is a major midbrain hub where ascending auditory pathways converge and auditory information is integrated and refined. Located in the tectum of the midbrain, the IC acts as a relay between brainstem structures and higher-order centers such as the thalamus and auditory cortex. It receives both ascending inputs from the lateral lemniscus and SOC, as well as descending projections from cortical areas, including feedback from layer V of the auditory cortex (Figure 2A) (Cant, 2005; Schofield, 2005).

Functionally, the IC supports two major auditory processing streams: the lemniscal and non-lemniscal pathways (Carbajal and Malmierca, 2018). The lemniscal pathway is primarily formed by the Central Nucleus Inferior Colliculus (CNIC) and is composed of tightly organized laminae of tonotopically aligned neurons. These so-called core cells convey detailed bottom-up information about the physical properties of sounds, such as frequency, intensity, and temporal structure. In the CNIC, the spectral tuning of neurons is broader compared to previous structures of the auditory pathway and is characterized by broad inhibitory circuits that sharpen the tuning of excitatory neurons. From the CNIC projections reach the ventral Medial Geniculate Body (MGv) in the auditory thalamus, preserving its high-fidelity auditory information for further processing in the cortex (Figure 2B) (Cant, 2005; Hackett, 2011).

In contrast, the non-lemniscal pathway arises from the dorsal, lateral, and rostral regions of the IC and surrounds the lemniscal core (CNIC). It is composed of matrix cells that are more broadly tuned, integrate multisensory input, and receive top-down feedback from cortical areas (Malmierca, 2003). These neurons exhibit prediction error activity, responding selectively when there is a mismatch between expected and actual auditory input, supporting their role in novelty detection and predictive coding (Carbajal and Malmierca, 2018; Parras et al., 2017). The non-lemniscal neurons project to other non-lemniscal areas in the thalamus, mainly the dorsal and medial geniculate division (Figure 2B).

Together, these pathways and subdivisions highlight the IC as not just a passive relay, but a dynamic processor involved in feature extraction, spatial localization, and integrative auditory cognition (Du et al., 2025), linking early brainstem processing with perception and expectation in the cortex.

2.1.4 *Thalamus*

The thalamus is a bilateral structure located in the diencephalon of the brain, serving as a central integrative hub with widespread projections across the central nervous system. Traditionally viewed as a passive relay station, contemporary research, however, has identified at least four overlapping canonical functions: (1) focal modulation of functional cortical areas, (2) interregional coupling to facilitate information transfer, (3) serving as a connector hub within large-scale brain networks, and (4) regulating the flow of information through dynamic gating mechanisms (Shine et al., 2023).

Within the auditory domain, the thalamus receives ascending input from the IC, which projects into the Medial Geniculate Body (MGB), the principal auditory thalamic nucleus (Figure 2A). The MGB is anatomically and functionally subdivided into three distinct regions: the ventral (MGv), dorsal Medial Geniculate Body (MGd), and medial Medial Geniculate Body (MGm) divisions (Carbajal and Malmierca, 2018). Mirroring the dual auditory pathways established in the IC, the lemniscal pathway, originating from the CNIC, terminates in the MGv. This division contains core cells that are sharply tuned, tonotopically organized, and transmit high-fidelity auditory information to the granular layer (III) of the primary auditory cortex (Figure 2B)(Antunes and Malmierca, 2021; Carbajal and Malmierca, 2018). In contrast, the non-lemniscal pathway, which arises from the lateral, rostral, and dorsal cortices of the IC, projects to the MGd and MGm. These divisions contain matrix cells that are more broadly tuned, integrate multimodal information, and project diffusely to supragranular (I, III), granular (IV), and infragranular (V) layers of the secondary auditory cortical areas. These nonlemniscal regions of the thalamus are in turn densely innervated by descending corticothalamic projections, which outnumber thalamocortical projections. This feedback loop has been suggested to play a prominent role in predictive coding and top-down modulation of sensory processing (Figure 2B) (Antunes and Malmierca, 2014, 2021; Somervail et al., 2025).

Collectively, this thalamocortical architecture underscores the role of the thalamus not merely as a conduit for auditory signals but as a dynamic regulatory node that contributes to perceptual salience, adaptive behavior, and contextual modulation of sound processing in everyday life. It is the last station before the auditory signals reach the auditory cortex.

2.1.5 *Auditory Cortex*

The auditory cortex is situated on the temporal lobe, encompassing the superior temporal plane and the superior temporal gyrus. It is composed of three major regions: the core, belt, and parabelt areas. The core region, including the primary Auditory Cortex (A1), lies posteromedially on Heschl's gyrus and receives input predominantly from the MGv of the thalamus via the lemniscal pathway (Bartlett, 2013). Surrounding the core are the belt areas, which are innervated by both the MGm and MGd as well as the auditory core. These, in turn, are surrounded by the parabelt, which receives projections from the primary auditory cortex as well as inputs from adjacent cortical regions (Figure 2A,B) (Celesia and Hickok, 2015). Importantly, the parabelt regions of the auditory cortex receive thalamic projections from the non-lemniscal pathway (Bartlett, 2013; Scott et al., 2017).

A1 follows the tonotopic organization of lower areas, with low frequencies represented rostrally and high frequencies caudally. In contrast to the one-dimensional frequency representation of the cochlea, cortical tonotopy spans two dimensions: a frequency gradient in one direction and isofrequency contours in the other. These contours group neurons with similar frequency preferences. Within each contour, tuning sharpness varies, with neurons at the center typically displaying narrower frequency bandwidths than those at the edges (Formisano et al., 2003; Moerel, De Martino, and Formisano, 2014).

Functionally, the belt and parabelt areas show specialization along the cortical axis, paralleling the division seen in visual processing (Goodale and Milner, 1992). Caudal and dorsal regions of the belt are involved in sound localization, whereas anterior-ventral regions are engaged in sound identification and spectrotemporal pattern analysis. These functional divisions support the dual-stream model of auditory processing, with dorsal "where" and ventral "what" pathways (Rauschecker and Tian, 2000).

In addition to spatial and spectral processing, the auditory cortex plays a crucial role in temporal coding. Along the ascending auditory pathway, the ability of neurons to phase-lock to rapid modulations diminishes: from several kilohertz in the auditory nerve, to 300 Hz in the thalamus, and down to 100 Hz in A1. Yet, humans can perceive modulation rates well above these cortical limits. This is explained by a transition from temporal coding to rate coding, wherein rapid modulations are represented not by precise spike timing but by average firing rates. Crucially, the change from temporal to rate coding in the auditory cortex reflects a transformation of temporal coding, from moment-by-moment to segment-by-segment basis, allowing for complex integration

over a larger window of time (Joris, Schreiner, and Rees, 2004; Lu and Brimijoin, 2022; Wang et al., 2008b).

As auditory information ascends the neural hierarchy, neuronal responses become increasingly selective and abstract. While auditory nerve fibers are tuned to single stimulus dimensions such as frequency or intensity, cortical neurons respond to specific combinations of features, including frequency, spectral bandwidth, intensity, modulation rate, and spatial location. These neurons occupy small receptive fields within a multidimensional feature space, enabling fine-grained auditory discrimination. In A1 and belt regions, neural responses often follow a two-phase pattern: a widespread onset response to any auditory input, followed by sustained firing in neurons that are selective for specific stimulus features (Wang, Gao, and Gao, 2005). Whereas core areas respond robustly to simple sounds like pure tones, belt regions are more responsive to spectrally and temporally complex stimuli, such as vocalizations and naturalistic sounds (Brodbeck, Presacco, and Simon, 2018; Norman-Haignere, Kanwisher, and McDermott, 2015; Norman-Haignere et al., 2022). This is of particular interest to the thesis, as simple tones elicit different responses compared to more complex naturalistic sounds. As will be shown next, a few recent studies have investigated the auditory cortex in response to natural sounds and speech specifically.

For natural sounds, Giordano et al. (2023) recently showed that the transformation from acoustic input to semantic representations of natural sounds begins in non-primary auditory regions. They compared acoustic models, semantic models, and the internal layers of a deep neural network trained for sound recognition against functional Magnetic Resonance Imaging (fMRI) data. Acoustic models best explained activity in Heschl's gyrus (primary auditory cortex), while intermediate Deep Neural Net (DNN) layers better predicted activity in the Super Temporal Gyrus (STG). Neither acoustic nor semantic models alone accounted well for STG responses, suggesting that this region supports intermediate representations bridging acoustics and semantics. The authors propose that these distributed, componential codes in the auditory cortex form the basis for downstream auditory object categorization in higher-order cortical areas, such as the ventrolateral prefrontal cortex.

In terms of speech processing, Hamilton, Edwards, and Chang (2018) showed that speech onsets and continuous speech are coded separately in non-primary regions of the auditory cortex, using Electrocorticography (ECoG) recordings. Specifically, posterior STG regions exhibited transient responses time-locked to sentence onsets, whereas more rostral STG regions showed sustained responses over the duration of speech. Importantly, this spatial parcellation was not specific to intelligible speech, as reversed and spectrally rotated stimuli elicited similar patterns, indicating that the responses

reflect general acoustic processing. These onset responses may provide crucial boundary markers that support the segmentation of continuous auditory input. Moreover, the encoding of phonetic features was not uniform but varied depending on whether they occurred at sentence onsets or within ongoing speech, suggesting that temporal context shapes phonetic representations. Hamilton et al. (2021) built on their earlier demonstration of onset and sustained response zones in STG by showing that speech processing in primary and non-primary auditory regions is more parallel and distributed. Specifically, the posterior STG areas responding to onsets showed similar temporal latencies of activity as the primary auditory cortex (Heschel's gyrus). For other non-primary regions, a temporal gradient from posterior medial areas to lateral frontal areas was found. Interestingly, they showed that stimulation of Heschel's Gyrus does not lead to deterioration of speech understanding, whereas lateral STG stimulation does. This suggests that parabelt areas are vital for speech perception.

The presented research provides a rough overview of how auditory information is processed by the brain. Most importantly, it provides a basis for what activity occurs and thus could be potentially measured in response to real-life soundscapes. This is critical, as the imaging modality of this thesis is non-invasive. Hence, the neural activity that can be measured is limited. How that is exactly the case will be explored in Chapter 4 in detail. Besides general processing of acoustic scenes, the adaptation to a dynamic environment is critical. How the auditory pathway adapts to the concurrent environment will be explored next.

2.2 NEURAL ADAPTATION

Adaptation in the brain is a fundamental process and refers to the capacity to dynamically alter ongoing neural activity to account for variability, or lack thereof, in the environment. The process of adaptation can occur on longer time scales, such as developmental adaptation induced by the lack of a diverse acoustic environment (Sanes and Bao, 2009), or on a day-to-day basis, such as in response to ongoing and unchanging sounds. This chapter exclusively deals with adaptation on a shorter timescale, specifically stimulus-driven adaptation. Auditory adaptation begins at the earliest stages of the auditory system, notably within the cochlea and the auditory nerve. This foundation of dynamic regulation is crucial for higher-order functions such as speech tracking, auditory scene analysis, and attentional modulation, which are implemented further along the auditory pathway. Throughout, it will become apparent that adaptation occurs at every stage of the auditory pathway to some degree inde-

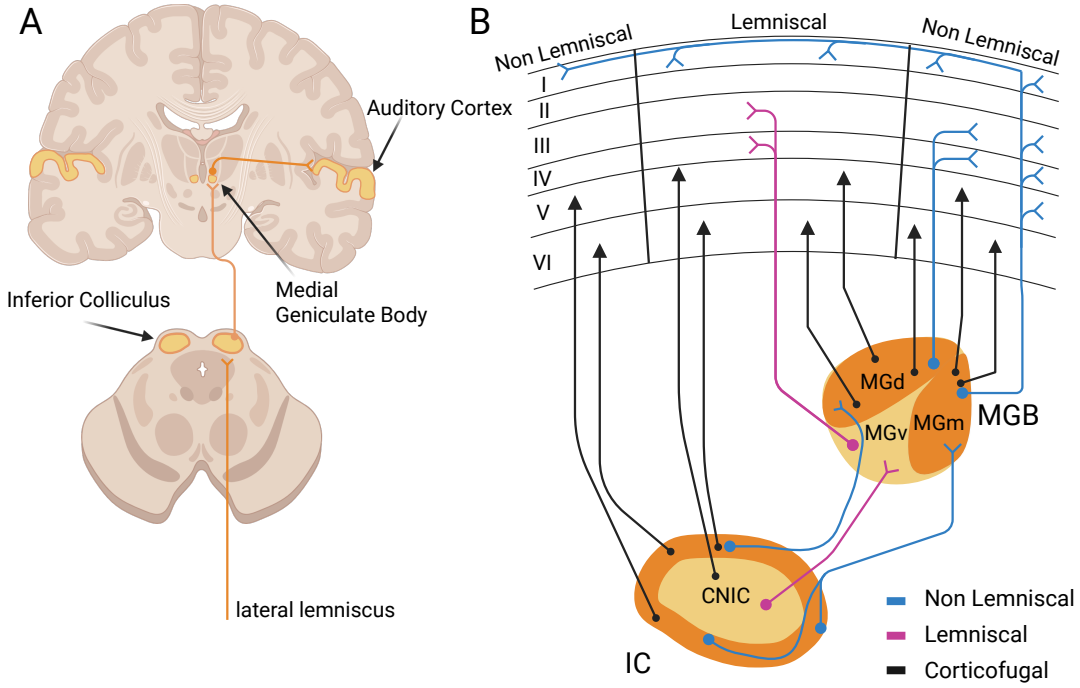

Figure 2: **A**: The lateral lemniscus projects to the IC in the midbrain. From there, the IC projects to the MGB of the thalamus, which in turn, projects to the auditory cortex. **B**: The auditory pathway comprises two main routes: the lemniscal (high fidelity) and the non-lemniscal (multimodal) pathway. The lemniscal pathway projects from the CNIC to the MGv of the thalamus. The MGv in turn projects to layers III and IV of A1. Both the CNIC and MGv receive descending cortical projections. The non-lemniscal pathway originates from the shell of the IC, which projects to the MGd and MGm. Neural projections from these two areas reach cortical layers I, III, IV, and V in lemniscal and non-lemniscal areas of the auditory cortex. Both the shell of the IC and MGm/MGd receive cortical back projections (adapted from Carbajal and Malmierca (2018) with modifications). Created with BioRender.com.

pendently of higher-order stages, but generally becomes more pronounced the higher the information traverses. This chapter follows the structure of the auditory pathway.

2.2.1 *Cochlear and Auditory Nerve*

The earliest adaption occurs in the cochlear and auditory nerve. Such early adaptation mechanisms serve to optimize the dynamic range of the auditory system. By adjusting sensitivity at the receptor and synaptic levels, the auditory system can amplify weak signals, suppress sustained redundant input, and remain responsive to rapid changes in the acoustic environment. For instance, in the cochlea, hair cells exhibit

mechanical adaptation, dynamically adjusting their sensitivity to maintain effective signal encoding over time (Fettiplace and Kim, 2014).

Outer hair cells, which are innervated by efferent projections from the medial SOC, modulate cochlear mechanics by decreasing sensitivity, a process thought to contribute to anti-masking, helping to extract relevant signals from background noise (Winslow and Sachs, 1987). Inner hair cells, in turn, receive feedback from the lateral SOC, which modulates the excitability of spiral ganglion neurons, influencing how signals are transmitted to the brainstem (Guinan, 2006).

Adaptation continues at the level of the auditory nerve, where the primary form of neural adjustment is firing-rate adaptation. Upon prolonged stimulation, the discharge rate of auditory nerve fibers progressively declines (Heil, 2004; Taberner and Liberman, 2005).

2.2.2 *Inferior Colliculus*

The IC integrates converging brainstem pathways and exhibits several forms of short-term adaptation to the recent acoustic context. First, IC neurons show dynamic range adaptation to sound-level statistics. In other words, neurons adapt their threshold to match the mean level intensity of the sound to avoid saturation. For more complex sounds (bi-modal distribution), neuronal populations divide their threshold adjustment to account for either of the probable sound pressure levels. Interestingly, the ability of the neurons in the IC to rapidly adapt to statistics depends on prior exposure. This has been suggested to reflect a learning effect, where a model of the environment has been generated, representing top-down cortical modulation (Dean, Harper, and McAlpine, 2005).

Besides mean-level sound statistics, the contrast, or variability of sounds, is also critical. For instance, evidence in mice IC neurons has shown that during a low contrast context (sound embedded in noise i. e., low Signal-to-Noise Ratio (SNR)), neurons respond more strongly to changes in sound level (higher gain). For a high contrast context (pure tone with now background noise, i. e., high SNR), the sensitivity of neuronal responses is reduced (lower gain). These gain mechanisms have been found in the CNIC, the lemniscal area of the IC (Willmore and King, 2023). Importantly, the adaptation to sound contrast is observed independently of the cortex; the response adaptation, however, appears to be prolonged when the same stimulus is re-encountered, suggesting top-down learning effects (Robinson, Harper, and McAlpine, 2016).

At last, Stimulus-Specific Adaptation (SSA), which refers to the reduced response to repeating tones, is predominantly found in the non-lemniscal area of the IC. Neurons are active when a sequence of tones is interrupted by a deviant tone, representing deviance detection/prediction error response (Carbajal and Malmierca, 2018). In terms of top-down control, cortical deactivation typically attenuates but does not abolish IC SSA. This indicates that corticofugal feedback modulates rather than generates these midbrain adaptations (Parras et al., 2017).

Together, mean-level adaptation, contrast gain control, and SSA illustrate how the IC re-centers, re-scales, and contextualizes the ascending code before it reaches the thalamus.

2.2.3 *Thalamus*

The thalamus represents the next stage of auditory processing, after the IC. Similar to the IC, the contrast gain control occurs in the lemniscal area (MGv). Here, neurons respond strongly (high gain) when the contrast is low. However, the time window of integration for the contrast control mechanism is longer than that in the IC, suggesting that the additional processing stabilizes gain representations. Another aspect that is shared with the IC is that this type of response adaptation occurs largely independent of the cortical activity, which takes a predominant modulatory role (Lohse et al., 2020).

Following the functional division of the IC, SSA is predominantly found in the non-lemniscal divisions (see Section 2.1.4). SSA persists when cortex is cooled/inactivated, yet its magnitude and persistence are shaped by descending projections, consistent with top-down gain control and predictive modulation (Carbajal and Malmierca, 2018). Functionally, the MGB represents another stage in auditory processing, which filters and re-emphasizes ascending signals of the IC. Specifically, the MGv maintains a high-fidelity, contrast-normalized channel to core cortex, whereas MGd/MGm convey context- and deviance-sensitive signals to belt/parabelt of the auditory cortex.

2.2.4 *Auditory Cortex*

Following the lemniscal/non-lemniscal distinctions established subcortically, the auditory cortex expresses complementary forms of short-term adaptation. A1 receives predominantly projections from the lemniscal areas of the thalamus, expressing contrast gain regulation-related activity (Lohse et al., 2020). Here, the longest time windows of integration are found compared to IC and MGB, suggesting additional cortical process-

ing. In line with the cross-brain region interaction, this type of adaptation appears to be largely independent of lower regions.

Notably, SSA has been observed in the A1 (Ulanovsky, Las, and Nelken, 2003), which receives limited projections from non-lemniscal thalamic and IC regions. This adaptation seems to result from local cortical processing rather than input from non-lemniscal subcortical areas, as network modeling suggests that intracortical dynamics account for the reduced response to repeated stimuli (Yarden and Nelken, 2017). Although deviance detection elicited a large response in the A1 (lemniscal), a greater response to deviance detection can be observed in the belt areas (non-lemniscal) (Carbajal and Malmierca, 2018). Additionally, in how far the top-down corticofugal network, described previously, modulates lower-level activity remains uncertain (King, Teki, and Willmore, 2018).

Besides external stimulation, Eliades and Wang (2008) have shown that during self-generated vocalizations (speech), auditory cortical neurons exhibit pre-emptive suppression, reducing responsiveness to expected input. This suppression likely originates in cortical circuits and appears earlier than similar effects in the IC or brainstem, suggesting a top-down predictive mechanism that distinguishes self-generated from external auditory stimuli.

Adaptation to more complex auditory environments has also been examined. For example, when individuals are exposed to continuous environmental background noise, the brain constructs an internal model of the noise's statistical properties. This model allows the auditory system to filter out irrelevant noise components and to detect meaningful sound events more efficiently (Hicks and McDermott, 2024). Supporting this mechanism, research has shown that the statistical features of background noise are actively suppressed to enhance the neural representation of speech in the left STG (Khalighinejad et al., 2019).

Most of the adaptation effects discussed so far have been observed using invasive techniques in mostly animal studies and some in humans (see Section 2.1.5 for investigations in humans using ECoG). The advantage of such methods is the ability to directly measure neural activity within the brain, offering higher spatial and temporal resolution than non-invasive alternatives. However, invasive recordings are not feasible for the majority of participants in human neuropsychological experiments. In the context of non-invasive measurements, the degree of stimulus-specific adaptation depends on the Inter-Stimulus-Interval (ISI) (López-Caballero, 2025). Where the neural response is reduced for shorter ISI. Furthermore, it has been shown that the direct context in which the stimulus was presented impacted the degree of modulation (Lanting et al., 2013). Another well-documented neural marker of neural adaptation involves

interrupting a train of similar tones by a deviant one. Here, a large negative potential is found, referred to as Mismatched Negativity (MMN), which peaks between 100-250 ms after stimulus onset (Garrido et al., 2009; Sams et al., 1985). Whether this is due to a memory trace or neuronal depletion remains debated (May and Tiitinen, 2010; Näätänen, 2001). Besides inter-stimulus distance, the frequency similarity between tones has been shown to lead to neural adaptation. Here, the degree of adaptation depends non-linearly on the spectral overlap of successive tones (Herrmann, Schlichting, and Obleser, 2014).

While these findings have been reliably replicated, they are largely based on simplified stimuli such as isolated, artificial tones. It remains unclear whether similar adaptation effects occur in response to more complex, continuous, naturalistic auditory environments. The second study in this thesis aims to address this gap by investigating whether these adaptation mechanisms also apply to natural soundscapes. Before, I will define the concept of soundscape and how it can be investigated.

3

SOUNDSCAPE

"Vi hör med örat, men lyssnar med hjärnan."

(Cognitive Hearing Science for
Communication Conference 2024)

Key Takeaways

- Natural soundscapes are inherently complex due to overlapping sounds and their temporal dynamics.
- An important distinction is between the physical properties of the sound source (proximal) and the subjective experience (perceptual).
- The proximal soundscape can be measured, and the perceptual needs to be inferred.

What you will learn:
What makes the soundscape complex, and how the proximal can be linked with the perceptual soundscape.

The concept of a soundscape may seem intuitive; most of us know what it means to "listen to our environment", yet defining it scientifically is challenging. As with the classic question of whether a falling tree produces a sound, the answer depends on whether sound is understood as a physical signal or a perceptual experience. Soundscapes share this duality: they can be described as measurable acoustic signals (proximal soundscape) or as subjective auditory experiences (perceptual soundscape). In this section, I will outline what makes soundscapes complex, trace how they have been defined in the literature, and highlight why the distinction between proximal and perceptual soundscapes is critical for their scientific investigation. To illustrate their diversity and richness, I invite the reader to explore examples of soundscapes from around the world: https://aporee.org/maps/.

3.1 SOUNDSCAPE COMPLEXITY

Compared to the simplified stimuli often used in laboratory experiments, real-world soundscapes are vastly more complex. Imagine standing on Times Square in New York and recording your environment. Chances are that this mental scenario was a rather busy and loud one. From all the cars driving by, honking, humans chattering, walking sounds, advertisements being played, music by street performers, dogs barking, to pigeons flapping their wings (Figure 3). All of these sounds overlap and converge into a dense mix, constituting the proximal soundscape. In the recording, there are an unknown number of concurrently active sound sources. Each source follows its own dynamics, some continuous, others marked by onsets and offsets, or by spatial movement. The superposition of these sound sources, their actions and events, as well as their dynamically changing nature, constitute the complexity of soundscapes. This complexity is central to the challenge of linking acoustic signals to perception and motivates the need for clear operationalization.

Figure 3: Capturing the proximal soundscape with microphones can yield a mixture of several sound sources. Exemplary, recording the soundscape at Times Square in New York may contain sounds of pedestrians, pigeons, and dogs barking (generated with ChatGPT 5).

3.2 SOUNDSCAPE DEFINITION

The scientific investigation of soundscapes began with Southworth (1969) definition of the soundscape as *"the quality and type of sounds and their arrangements in space and time"*. While this formulation captures the intuitive idea of soundscapes, it is too vague to serve as an operational framework. A significant step forward was made by Farina and Pieretti (2012), who defined the soundscape as *"the acoustic context produced and, in turn, perceived in different ways by both animals and humans"*. This definition introduced the perceiver as being a critical constituent of the soundscape, which represents a crucial step in guiding future soundscape research by explicitly acknowledging the subjective nature of perception. The role of the perceiver is also emphasized in the *ISO 12913-1:2014(en), Acoustics — Soundscape — Part 1: Definition and conceptual framework* (2014), which defines the soundscape as *"an acoustic environment, as perceived and/or understood by a person or group of people in context"*. These developments highlight a vital distinction between the physical properties of sound and the perceptual processes of the listener, a theme that recurs throughout this thesis. However, none of these definitions fully capture the physical aspects that constitute a soundscape.

A more recent approach addresses this gap by distinguishing three complementary categories: the distal, the proximal, and the perceptual soundscape (Grinfeder et al., 2022). The distal soundscape is defined as the totality of *"spatial and temporal distribution of sounds in a prespecified area in relation to sound propagation effects"*. While conceptually useful, it is more of a theoretical than a measurable concept. In contrast, the proximal soundscape is defined as *"the collection of propagated sound signals that occurs at a specific point in space"*. This category of soundscape is relevant for studying sound perception in everyday life, as it provides a point at which the soundscape can be obtained. In practice, however, it can only be approximated through recordings, which are limited by the acoustic range of the recording device. At last, the perceptual soundscape is defined as *"the individual subjective interpretation of a proximal soundscape"*.

Here, I only consider the proximal and perceptual soundscape. This dual perspective provides a framework for linking measurable acoustic signals with subjective auditory experience, which is the central goal of the thesis.

3.2.1 *Proximal Soundscape*

The first step is to investigating the auditory perception of complex soundscapes is to operationalize the proximal soundscape. This encompasses those aspects of the environment that can be quantified from the acoustic signal itself. This operationalization of acoustic models is often inspired by the perceptual organization, as will be shown below. Thus, the boundary between proximal and perceptual soundscape is blurry and can shift with additional novel modeling techniques (for a detailed discussion, refer to Section 11.2). The boundary in this thesis is the information and models that can be extracted from the proximal soundscape alone, without additional information on the perceptual state of the perceiver.

For instance, the segregation of complex soundscapes into separate sound sources has been shown to rely on acoustic information. Młynarski and McDermott (2019) showed that the co-occurrence of acoustic onsets, harmonicity, spectral similarity, and shared temporal modulation patterns are crucial acoustic cues for events to be grouped into the same or separate source/s. Furthermore, time-frequency characteristics such as harmonicity and temporal coherence bind features together into higher-order qualities into timbre, which are used to determine the source's identity (Ogg and Slevc, 2019).

Once the sound source has been determined, the process of chunking these continuous streams into perceptually meaningful events is referred to as event segmentation (Winkler, Denham, and Escera, 2013). Here, changes in acoustic attributes such as frequency and intensity can be determinants of events' on- and offsets. Speech processing is an example where event segmentation is relevant. Slow amplitude fluctuations correspond to syllable rate in speech, which are crucial constituents of communication (Ding et al., 2016; Oganian et al., 2023). Besides the chunking into separate events, the acoustic context can be depicted by the temporal proximity of sound events, and sudden energy changes over longer periods of time. For instance, Huang and Elhilali (2017) showed that sudden changes in loudness, pitch, and the spectral profile represent bottom-up acoustic cues of saliency.

These acoustic properties can be used to describe different aspects of the proximal soundscape. In practice, the acoustic properties of the recorded soundscape of Times Square (i. e., loudness, spectral properties, saliency) can be derived to estimate the sources of sounds. Thus, they provide the bridge to perceptual interpretations by linking physical signals to categories that listeners recognize. While, in principle, fairly straightforward, obtaining real-world recordings of a proximal soundscape with a sufficient level of description poses a significant challenge. Placing a microphone in

Times Square and recording continuously may capture the proximal soundscape, but extracting features to link to the perceptual soundscape is no easy task. For instance, separating the different sound sources, a task which is solved with millisecond precision by the auditory pathway (Section 2.1.5), is a methodological challenge from single-channel recordings (Ansari et al., 2023; Nourifard, 2025). The separation of neural sources, echoing the current issue, will be discussed in Section 4.3.1.

3.2.2 *Perceptual Soundscape*

The perceptual soundscape, in contrast, reflects how listeners organize, interpret, and experience incoming sounds (Grinfeder et al., 2022). Crucially, it cannot be directly inferred from the proximal soundscape, since perception depends on internal, cognitive, and contextual processes. For example, through acoustic feature analysis, I may have been able to extract the different sources of sounds that were present in Times Square, but whether each was perceptually relevant, we cannot know from the soundscape alone. This has far-reaching consequences as auditory perception is strongly shaped by top-down influences such as attentional focus (Ding and Simon, 2014; Obleser and Kayser, 2019).

Perception is further shaped by subjective dimensions such as pleasantness, eventfulness, and familiarity (Axelsson, Nilsson, and Berglund, 2010), as well as by affective states: acoustically identical sounds (e.g., footsteps from upstairs neighbors) may be experienced as neutral or highly irritating depending on the listener's affective state (Frescura et al., 2025). Finally, multimodal integration also plays a decisive role. Visual cues can override or reshape auditory processing, as in the ventriloquist effect (Bruns, 2019), the McGurk effect (Mcgurk and Macdonald, 1976), or audiovisual facilitation of speech in noise (Puschmann et al., 2019; Stekelenburg and Vroomen, 2007). These factors illustrate why the perceptual soundscape cannot be captured directly from the proximal soundscape: it is an active construction, shaped by top-down attention, affect, context, and multimodal cues.

This distinction highlights a central challenge for research: proximal soundscapes can be measured, but perceptual soundscapes must be inferred. The question, then, is how to obtain a measure of the perceptual soundscape. One approach would be through experimental manipulation of cognitive states (*"focus on the pigeon flapping on Times Square"*) or self-reports (*"does the pigeon flapping annoy you?"*). While valid, both approaches disrupt natural behavior (imagine being asked about your current level of annoyance every 5 minutes) and provide only momentary assessment in the latter

case. Besides these considerations, obtaining a sufficient level of description from real-world recordings already poses a challenge, due to the soundscape's complexity.

A common response has been to return to the laboratory, presenting isolated tones that vary along a single acoustic dimension (e.g., amplitude or frequency). Yet, such stimuli fall short of capturing the complexity of everyday listening (Hamilton and Huth, 2020). A more progressive strategy is to present increasingly complex stimuli under controlled conditions, allowing researchers to identify the driving factors of perceptual experience (Holleman et al., 2020; Shamay-Tsoory and Mendelsohn, 2019; Stangl, Maoz, and Suthana, 2023; Vallet and Van Wassenhove, 2023).

In this setting, a promising approach to obtaining continuous estimates of the perceptual soundscape is to approximate it by measuring the underlying neural processes outlined in Chapter 2. A suitable neuroimaging method is EEG, due to its temporal precision and high mobility for future beyond the lab recordings (Larson and Lee, 2013; Winkler, Denham, and Escera, 2013). Although it does not provide direct access to subjective qualia, it offers an indirect window into how proximal features are transformed into perceptual representations. By modeling brain responses to acoustic input and contrasting them with features tied to perceptual relevance, EEG enables us to approximate aspects of the perceptual soundscape in ways that go beyond passive description. This link forms the methodological and conceptual bridge between the soundscape framework and the empirical studies presented in the following chapters.

4

EEG

"It's a popular fact that 90 percent of the brain is not used and, like most popular facts, it is wrong. . . . It is used. One of its functions is to make the miraculous seem ordinary, to turn the unusual into the usual. Otherwise, human beings, faced with the daily wondrousness of everything, would go around wearing a stupid grin, saying "Wow," a lot. Part of the brain exists to stop this from happening."

Pratchett (2010)

Key Takeaways

- The neural signal that is captured by EEG is generated by the synchronous activity of predominantly cortical neuronal populations.
- EEG can be analyzed in the time and frequency domains and provides high temporal resolution.
- Mobile solutions enable the continuous recording beyond the lab to study auditory perception.

What you will learn:
How the signal that is measured by the EEG is generated and how it can be analyzed to investigate auditory perception.

EEG measures the brain's electrical activity using electrodes placed on the scalp. Importantly, EEG does not record the activity of single neurons but instead reflects the summed activity of large populations of neurons, integrated over cortical areas of approximately 10 cm^2 or more (Buzsáki, Anastassiou, and Koch, 2012). In the context of this thesis, EEG is of particular relevance because it allows us to investigate the neural processes underlying complex, naturalistic sound perception. Other neuroimaging

methods, such as fMRI, functional Near-Infrared Spectroscopy (fNIRS) or Magnetic Encephalography (MEG), either lack the portability needed for everyday life studies or do not provide the necessary temporal resolution. In the following sections, I will outline the fundamental principles that determine what EEG measures. This includes the biophysical basis of current generation at the cellular level, the neural origins of the EEG signal, the effects of volume conduction, the temporal characteristics of EEG, and finally, an overview of different measurement setups.

4.1 CURRENT GENERATION IN THE BRAIN

To understand the neural processes underlying sound perception, one needs to understand what neural signal the EEG measures. Thus, I explore how electrical current activity in the brain is generated in the first place. This starts at the single neuron level. Neurons maintain a resting potential of approximately –70 mV relative to their extracellular environment. This electrical gradient is established by the unequal distribution of sodium (Na^+), potassium (K^+), and chloride (Cl^-) ions across the cell membrane. When a presynaptic neuron releases neurotransmitters into the synaptic cleft, the permeability of the postsynaptic membrane changes, allowing charged ions to flow (Buzsáki, Anastassiou, and Koch, 2012).

Two major types of postsynaptic responses can occur, depending on the type of neurotransmitter. One type of neurotransmitter release leads to the influx of positive ions, and the postsynaptic membrane depolarizes, resulting in an Excitatory Postsynaptic Potential (EPSP). Importantly, once cell membrane depolarization passes a certain threshold, an action potential occurs. It represents an all-or-nothing strong depolarization of the cell membrane at the axon hillock and axons. Conversely, the other type of neurotransmission causes the outflow of positive ions, which leads to a hyperpolarization of the cell membrane, producing an Inhibitory Postsynaptic Potential (IPSP) (Amzica and Silva, 2017).

In the case of an EPSP, the entry of positive ions into the apical dendrites produces a local current sink, i.e., a region of negative extracellular potential at the site of ion influx. To preserve electroneutrality, this sink is balanced by a current source elsewhere on the neuron, typically near the soma, where positively charged ions accumulate extracellularly. The migration of ions between these regions, termed return current, generates extracellular potentials, also referred to as Local Field Potential (LFP) (Buzsáki, Anastassiou, and Koch, 2012) (Figure 4A). This spatial separation of sink and source establishes an electrical dipole along the somatodendritic axis. Importantly, the dis-

tance between the current source and sink significantly determines the size and shape of the LFP (Amzica and Silva, 2017). The voltage of single neuronal LFP, however, is much too small to be picked up by the EEG. Instead, it requires the synchronous activity of a large population of neurons.

4.2 SIGNAL GENERATION IN EEG

The voltage of single neuronal LFP, however, is much too small to be picked up by the EEG. Instead, it requires the synchronous activity of a large population of neurons. Importantly, all neuronal current sources contribute to the LFP, however, here I will only discuss the most dominant contribution, which is from Postsynaptic Potentials (PSP) (for an overview of other neuronal current sources, refer to Buzsáki, Anastassiou, and Koch (2012)).

The most prominent contributors to PSPs are pyramidal cells in the cortex. These neurons possess long apical dendrites, are arranged in parallel, and are oriented perpendicular to the cortical surface. This geometry is particularly favorable for generating strong extracellular fields. Their long dendrites create a large spatial separation of sinks and sources along the somatodendritic axis, which establishes strong dipoles. Their parallel alignment enables these dipoles of this neuronal population to superimpose constructively, which results in a signal that can be measured at the scalp (Figure 4B) (Buzsáki, Anastassiou, and Koch, 2012).

EEG is therefore most sensitive to neural activity at the cortical surface, where aligned pyramidal dipoles dominate. Here, the amplitude of the recorded potential depends on the distance between the current source and the electrode, decaying approximately with the square of the distance $1/r^2$ (Amzica and Silva, 2017; Buzsáki, Anastassiou, and Koch, 2012). Deeper sources generally lack the strength to be reliably detected, although recordings from deep brain structures are possible, such as measuring brain stem responses to sounds in infants (Hecox and Galambos, 1974; Kulasingham et al., 2024). This emphasizes why EEG is particularly well-suited to capture synchronous cortical surface activity, while being less sensitive to deeper or spatially dispersed sources.

4.3 VOLUME CONDUCTION

As outlined in the previous section, the ability of EEG to measure LFP depends on several neural factors as well as the distance between the recording electrode and the

neural current source. Importantly, electrical potentials generated in the brain do not reach the scalp directly, but rather spread through the intervening tissues. This process is known as volume conduction and refers to the passive propagation of electrical currents through the head. In the case of cortical activity, the potential generated at the source passes through multiple tissue layers, including the cortex, pia mater, cerebrospinal fluid, dura mater, skull, and skin (Figure 4B). Each of these layers differs in conductivity, with the skull in particular offering high resistance. As the potential spreads through these tissues, it becomes attenuated, distorted, and diffused. The resulting signal that arrives at the scalp electrode is therefore a spatially smoothed version of the underlying brain activity (Broek et al., 1998).

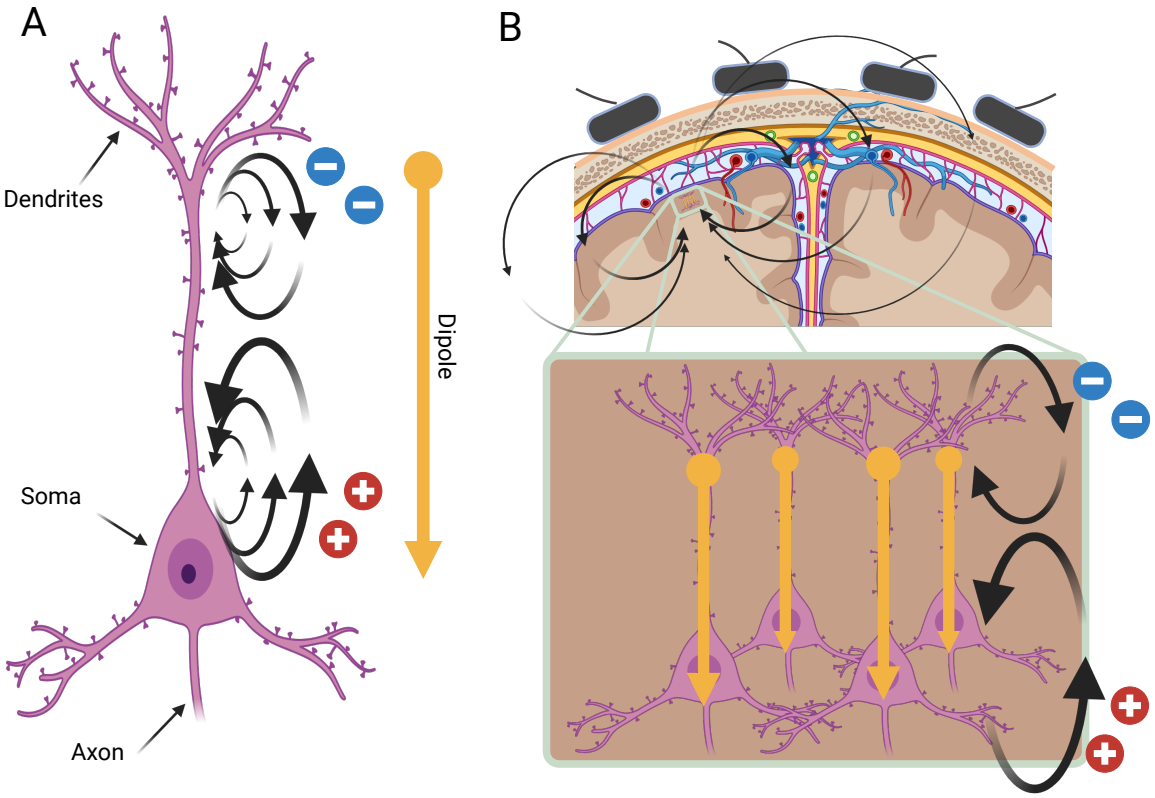

Figure 4: **A:** Neurons can be broadly categorized by consisting of dendrites, a soma, and an axon. An EPSP at the apical dendrites leads to a current sink (negative extracellular potential) at the site of ion influx, and is balanced by a current source near the soma. The resulting return currents establish an electrical dipole along the somatodendritic axis. **B:** Cortical arrangement of pyramidal neurons oriented orthogonally to the cortical surface. Synchronous activity of these neurons leads to LFP, which can be detected by electrodes placed on the scalp. The black arrows are set to visualize the effect of volume conduction, where their thickness represents the reduced source signal strength that is captured at distant compared to proximal electrodes. Created with BioRender.com.

4.3.1 *Sensor Source Relationship*

A highly related concept to volume conduction is the sensor (electrode) and source (LFP) relationship. EEG does not provide a direct measurement of isolated brain sources. Instead, each electrode records a superposition of activity from multiple neural generators. This includes the source of interest as well as unrelated or distant sources that also contribute to the scalp potential. This poses a significant challenge for EEG research, as activity measured at a single electrode cannot be attributed to a single neural population, which is known as the inverse problem. What makes the inverse problem particularly troublesome is that it has no unique solution, as the number of potential brain sources is unknown. This was proven as early as 1853 by Helmholtz (1853). Intuitively, the proof demonstrates that infinitely many source configurations can yield the same scalp distribution. This means that any solution is merely an approximation of the true underlying generators, making it difficult to draw definite conclusions from the results alone [1].

In practice, the sensor–source mapping is often assumed to be the result of linear mixing, whereby the scalp signal reflects a weighted sum of the underlying sources. This framing makes the problem mathematically tractable but still underdetermined. Several different approaches exist to offer solutions to the mixing problem. The most common in EEG research are Independent Component Analysis (ICA) (Hyvärinen and Oja, 2000; Makeig et al., 1995) and Principal Component Analysis (PCA) (Hotelling, 1936; Pearce and Hirsch, 2000), which belong to the blind source separation algorithms. Each has a different set of assumptions regarding the mixing process. For all these approaches, however, the number of sources that can be estimated is constrained by the number of electrodes used, that is, by the rank of the data matrix (Makeig et al., 1995). This is highly relevant, as the number of sources that can be estimated depends on the number of electrodes, and thus the type of EEG setup (discussed in Section 4.4). Besides separating neuronal sources, ICA is commonly applied to determine non-neural sources that introduce electrical noise in the measured signal.

4.3.2 *Noise*

Another crucial concept related to volume conduction is that of noise. Since electrodes record a superposition of activity from multiple sources, many of them unrelated to

1 This is the same problem encountered when determining the sound sources in a recorded complex soundscape, see Section 3.1

the source of interest, the measured signal can be conceptualized as the sum of the signal of interest and noise. This relationship is often quantified using the SNR. EEG generally has a low SNR, as it is not only influenced by distant neural sources due to volume conduction but also by a variety of non-neural signals that generate electrical activity.

Noise can arise from several sources. Physiological noise includes muscle activity, cardiac signals, respiration, and eye movements such as blinks and saccades. In addition, EEG is highly sensitive to environmental and instrumental noise, such as line noise, poor electrode contact, or cable movement (for a comprehensive list see Kaya (2021)). Because of volume conduction, noise from one source spreads widely across electrodes, further complicating the separation of signal and noise. While this list is not exhaustive, it illustrates how measuring the signal of interest is inherently prone to distortions from multiple sources.

To mitigate these effects, EEG data can be subjected to preprocessing. Common strategies include filtering (e.g., removing line noise), artifact rejection, and regression-based corrections. A powerful and widely used approach is ICA (Hyvärinen and Oja, 2000; Makeig et al., 1995), which assumes that sources mix linearly at the sensor level and can be statistically separated into maximally independent components. In practice, ICA decomposes the EEG into components that can be classified and, if appropriate, rejected as artifacts (Pion-Tonachini, Kreutz-Delgado, and Makeig, 2019). Importantly, ICA does not attempt to localize neural sources in anatomical space; instead, it separates the recorded signals into components that are maximally statistically independent, which may reflect distinct neural or non-neural processes.

Finally, it is important to note that the definition of noise is not absolute but depends on the research question. A signal that is considered noise in one context may be the feature of interest in another. For example, eye blinks are often removed because of their large voltage deflections, yet in certain perceptual or cognitive tasks, blink-related activity is itself highly relevant and therefore constitutes a meaningful signal (Holtze et al., 2023; Wascher et al., 2022).

4.4 EEG SETUP

EEG records weak electrical signals from the brain using electrodes placed on the scalp. Because voltage represents a potential difference, each EEG channel requires at least two electrodes: a recording electrode and a reference electrode. Together, they measure the difference in potential between two sites on the head. Since the raw

neural signals are extremely small, they must first be amplified, typically by a factor of 1,000 to 100,000 (Hari, 2017). The digitization on the recording device can range from a laboratory amplifier connected to a computer to more compact systems such as mobile phones.

After digitization, researchers often apply offline re-referencing to improve the SNR. The choice of reference electrode is crucial, as every channel reflects the potential difference relative to this site. Ideally, the reference is electrically neutral with respect to the processes under study, recording minimal and stable activity, and being placed sufficiently far from the electrode of interest to reduce contamination by volume conduction. Common approaches include linked mastoids and average referencing. In this thesis, a linked-mastoid reference is used, as this method is well validated in auditory EEG research (Mahajan, Peter, and Sharma, 2017) and is more suitable for the low-density electrode setup employed here, which is not optimal for average referencing (Lei and Liao, 2017).

EEG setups vary in montage electrode type and number of electrodes. Electrodes can be active or passive: active electrodes amplify the signal at the electrode site, whereas passive electrodes transmit the raw signal to the amplifier. Active electrodes reduce the effects of high impedance and mitigate external noise sources such as cable movement (Laszlo et al., 2014; Xu et al., 2017). However, in a real-world auditory task, Scanlon et al. (2021) reported no difference in data quality between active and passive electrodes during movement. In addition, electrodes may be wet, dry, or sponge-based. Wet electrodes, which employ conductive gel between the scalp and electrode surface, reduce electrical impedance and thereby increase SNR (Ehrhardt et al., 2024). The datasets in this thesis were acquired using passive, wet electrodes.

Another defining factor is the number of electrodes, which determines spatial coverage. This can range from ultra-high-density caps with up to 1024 electrodes (Schreiner et al., 2024), to high-density setups with 128–256 electrodes (Jaeger et al., 2020; Oostenveld and Praamstra, 2001), the standard 64-electrode montage (Mirkovic et al., 2015), sparse setups with 32 or 16 electrodes (Debener et al., 2015; Rosenkranz and Bleichner, 2022), and minimal systems with as few as 2 electrodes (Kosmyna et al., 2019). The choice of electrode density depends primarily on the research question. Additional considerations include participant comfort, setup time, and spatial coverage requirements. In the context of auditory perception EEG investigation, non-traditional approaches such as around-the-ear electrodes (cEEGrids; (Bleichner and Debener, 2017; Debener et al., 2015; Meiser, Knoll, and Bleichner, 2024) or in-ear electrodes (Geirnaert, Kappel, and Kidmose, 2025; Mikkelsen et al., 2021) have been proposed. These unob-

trusive cEEGrids are particularly promising for recordings outside the laboratory and will be explored in Chapter 9.

4.5 TEMPORAL CHARACTERISTICS OF EEG

This section provides a broad overview of how the EEG signal can be analyzed. Once the core concepts are outlined, specific examples for auditory perception will be provided. Importantly, EEG data can be analyzed in the time and frequency domain. To understand how the signal can be decomposed into different frequency bands and in how far these provide a physiological basis in the brain, I will first introduce the concept of oscillations. Next, I will introduce event-related potentials and how these transient, evoked responses can be derived and used to investigate the brain.

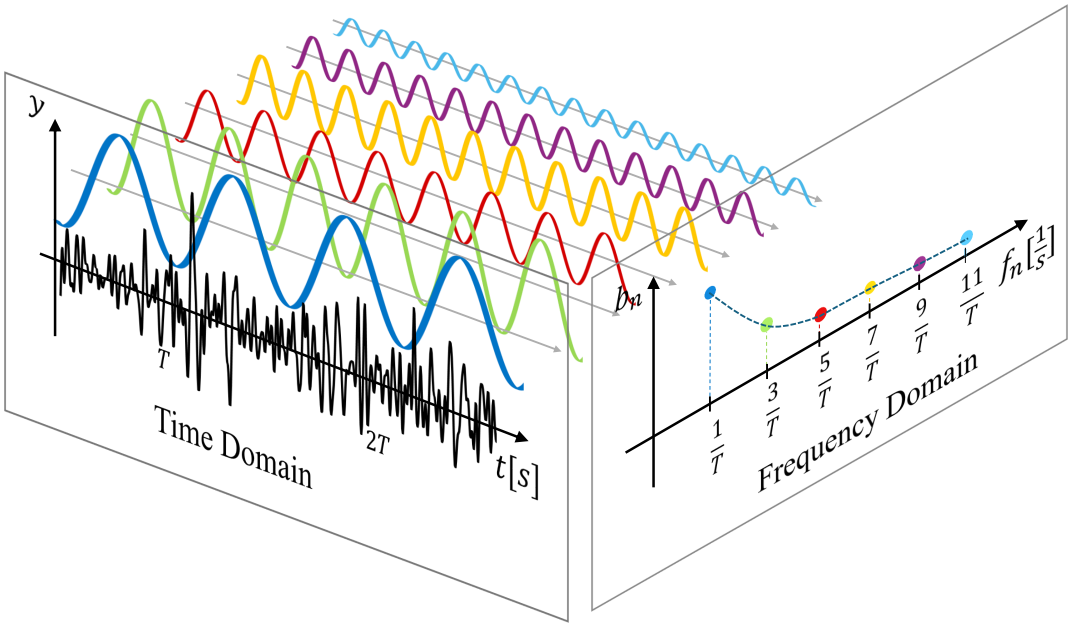

Figure 5: The neural signal can be described in the time or frequency domain. In the time domain, the signal is represented over time and the amplitude (for EEG in mV). Using the Fourier Transform, the signal can be displayed in the frequency domain. In the frequency domain, the signal is plotted over frequencies f and their respective magnitude b. Replicated from https://dibsmethodsmeetings.github.io/fourier-transforms/

4.5.1 *Oscillations in the Brain*

Fluctuations of PSP over time generate rhythmic patterns that appear in the EEG as oscillatory signals. By decomposing the time series into its frequency components using a Fourier transform (see Peters (1998) for a detailed explanation). The resulting Fourier coefficients are defined by their amplitude and phase for each respective frequency. Typically, the EEG spectrum follows a $1/f$ distribution, meaning that slower oscillations carry greater power than faster ones. This scaling reflects network properties: larger neuronal populations can be recruited over longer temporal windows, giving rise to stronger low-frequency activity (Buzsáki, Anastassiou, and Koch, 2012; Buzsáki and Vöröslakos, 2023). In conventional scalp EEG, the frequency range of interest spans approximately 0.1–100 Hz. Beyond this broad range, oscillations are commonly divided into canonical frequency bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (>30 Hz) (Figure 6A) (Burgess, 2012). Changes in power within these bands have been associated with a variety of cognitive and perceptual functions (Amzica and Silva, 2017). While it is tempting to map individual bands onto specific cognitive functions, these rhythms may reflect more general mechanisms of neural organization. Indeed, concerns have been raised that the precise boundaries between bands are somewhat arbitrary (Burgess, 2012). However, converging evidence supports the view that these divisions have a neurophysiological basis, as oscillations in different bands are generated by distinct neural circuits and are associated with different temporal scales of information processing (Buzsaki, 2006; Niedermeyer and Silva, 2005). In this thesis, particularly the low frequency components < 10Hz are of interest, as these have been linked to speech and sound perception (Ding and Simon, 2014).

In a seminal study, Luo and Poeppel (2007) demonstrated that spoken sentences could be classified based on the phase pattern of theta oscillations. This finding indicated that cortical activity in the phase of the theta band aligns, or entrains, to the dynamic fluctuations of the speech signal. Subsequent studies extended speech tracking relevant frequency components to the delta (Chalas et al., 2023; Gross et al., 2013; Howard and Poeppel, 2010) and gamma band (Zion Golumbic et al., 2013). The entrainment to the speech envelope has been a well-validated finding and has been extended to experimental designs.

For instance, neural entrainment is not purely stimulus-driven but also modulated by attention. In dichotic listening tasks, where two competing speech streams are presented simultaneously, entrainment to the attended stream is consistently stronger than to the unattended stream (Ding and Simon, 2012a; Jaeger et al., 2020; Zion

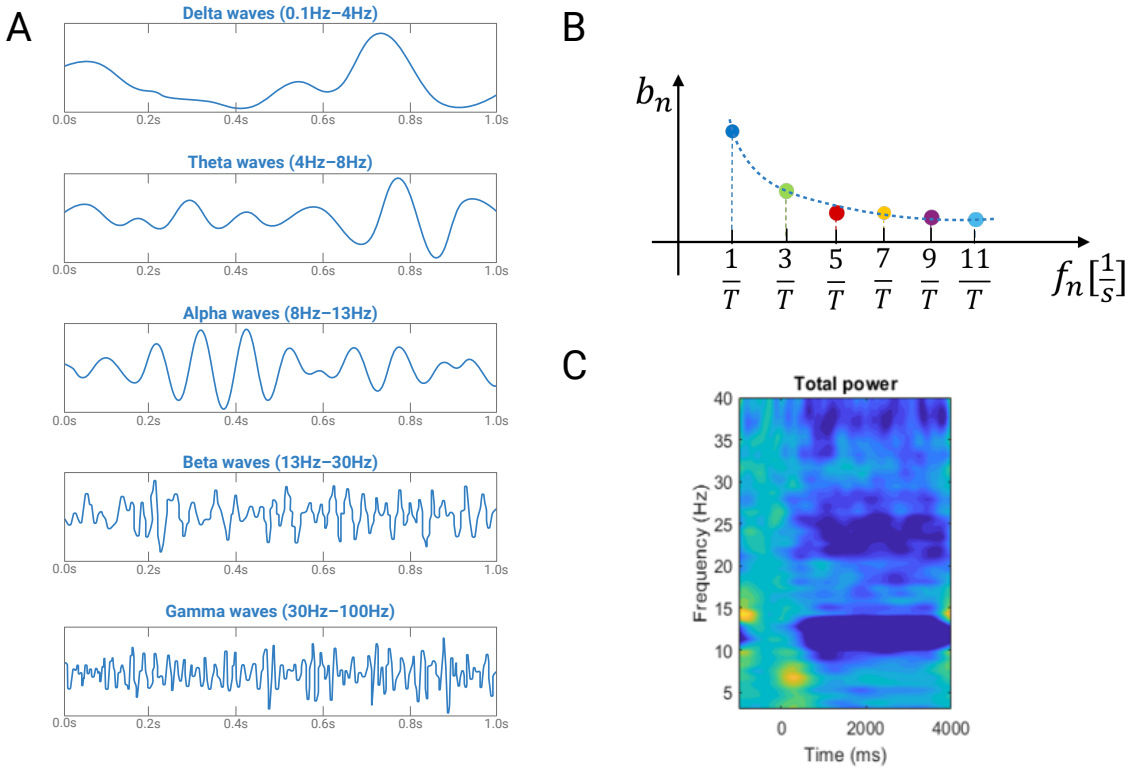

Figure 6: **A:** Several frequency bands can be distinguished in the brain. These are commonly grouped in the delta, theta, alpha, beta, and gamma bands. **B:** Their respective power can be inspected in the frequency domain. **C:** To determine fluctuations in the power of each of the frequencies, a time-frequency analysis can be plotted. Exemplary, the data of a participant performing motor imagery is shown. Here, a strong decrease in the alpha band power can be observed. Created with BioRender.com.

Golumbic et al., 2013). This suggests that oscillatory alignment to speech reflects an active process of selective attention, enabling listeners to focus on relevant auditory input in noisy environments. Furthermore, recent studies have investigated whether entrainment also reflects higher-order processes such as syntactic or semantic integration (Agmon et al., 2023; Brodbeck, Presacco, and Simon, 2018; Ding et al., 2016) or merely reflects acoustic processing (Daube, Ince, and Gross, 2019). Clarifying this distinction is crucial for understanding whether neural entrainment provides access primarily to low-level auditory encoding or also to the hierarchical processing stages underlying speech comprehension.

Despite the robustness of these findings, the underlying neural mechanisms remain debated. One account proposes that speech-driven fluctuations reflect phase resetting of ongoing, endogenous oscillations, aligning intrinsic neural rhythms to external acoustic events (Obleser and Kayser, 2019; Zoefel, Oever, and Sack, 2018). This

is believed to aid the parsing of speech into meaningful segments (phonemes, syllables) driven by acoustic landmarks (Giraud and Poeppel, 2012). An alternative view suggests that entrainment instead arises from the superposition of transient, evoked responses (see Section 4.5.2) that mimic oscillatory patterns when repeated over time (Breska and Deouell, 2017; Deoisres et al., 2023; Oganian et al., 2023). Current evidence does not fully resolve whether speech entrainment is best understood as endogenous oscillatory synchronization or as a byproduct of stimulus-locked responses. These considerations are crucial when extending the investigation of auditory perception to everyday life soundscapes that also contain non-speech objects (Section 3.2).

A natural question arises: if the focus lies only on those frequency bands that carry neural entrainment to continuous stimuli, are the other frequency bands simply noise that can be removed? The answer is yes, this is possible through digital filtering. In line with the general intuition of filters, their main function is to pass certain frequency components while attenuating others. In EEG, filters are typically designed to suppress activity outside the band of interest, thereby improving the SNR.

Filtering can be understood most intuitively in the frequency domain. Here, the Fourier spectrum of the EEG signal is multiplied by the frequency response of the filter. This multiplication acts like a window: frequency components within the passband are preserved, while others are reduced. The filtered signal can then be transformed back into the time domain. While this explanation is intuitive, it conceals an important complication: the effect of filtering on the phase of the signal.

Filters can be implemented with minimum-phase (causal) or zero-phase (acausal) properties. In the causal case, the filter introduces a temporal delay, which can shift the apparent timing of neural events. Zero-phase filtering, by contrast, eliminates net temporal delay by applying the filter forward and backward in time. However, this approach is acausal: future samples influence past points, which can smear responses backwards in time. Consequently, one might observe activity in the filtered signal before the actual event. This artifact is particularly problematic when interpreting event-related timing (Widmann and Schröger, 2012; Widmann, Schröger, and Maess, 2015).

For these reasons, careful design and documentation of filters is essential (for an excellent review, see Cheveigné and Nelken (2019)). It is also important to recognize that all filters, regardless of implementation, alter the signal in some way. Nevertheless, filtering is indispensable in EEG analysis—whether for suppressing slow drifts, removing line noise, or isolating frequency bands of theoretical interest. Beyond analysis, filtering also plays a role in stimulus design, for instance, when extracting the envelope of a speech signal, as will be introduced in later sections.

4.5.2 *Event Related Potentials*

Event-Related-Potential (ERP)s are a fundamental tool in EEG research for studying the brain's time-locked responses to discrete sensory (Adler and Adler, 1989; Lanting et al., 2013; Woodman, 2010), cognitive (Heidlmayr, Kihlstedt, and Isel, 2020; Proverbio, Santoni, and Adorni, 2020; Zuk, Teoh, and Lalor, 2020), or motor events (Jacobsen et al., 2022; Pfurtscheller, 1992). Given that the raw EEG signal is a superposition of several different sources, it is difficult to isolate the response to a single event within an individual trial. To improve the SNR of the event-specific neural response, researchers present participants with the same stimulus repeatedly and then average the corresponding EEG segments. This averaging process is based on the assumption that event-unrelated activity, being random across trials, will average out, while the consistent, event-related neural response will remain. As a result, averaging enhances the SNR and reveals voltage deflections that correspond to various stages of neural processing (Figure 7A) (Luck and Kappenman, 2011). Although potentially not directly apparent, these voltage deflections largely originate from PSP of synchronously active neuronal populations in response to the stimulus, as established in Section 4.1. Importantly, there remains an ongoing debate whether ERPs reflect an event-locked phase resetting of continuous oscillations or if they arise from the additive summation of stimulus-evoked neural activity (Obleser and Kayser, 2019; Oganian et al., 2023).

The trajectory of anERP waveform is characterized by a sequence of positive and negative voltage deflections, commonly referred to as peaks, that occur at distinct time points following stimulus onset. These peaks are typically labeled according to either their polarity and sequence (e.g., P1 for the first positive peak, N1 for the first negative peak) or their approximate latency in milliseconds (e.g., N170 for a negative peak occurring around 170 ms post-stimulus) (Luck and Kappenman, 2011, p. 72). In addition, some components are named after the cognitive processes they are thought to reflect, such as the MMN (May and Tiitinen, 2010). ERP components ideally reflect information processing related to the stimulus at different points in time. Carefully controlled experimental designs allow researchers to link specific ERP components to distinct cognitive or perceptual operations. For example, the brief presentation of faces reliably elicits a larger N170 compared to non-face stimuli, suggesting that the perceptual differentiation of faces from other objects occurs approximately 170 ms after stimulus onset (Rossion and Jacques, 2012). Although voltage values of peaks at each electrode contribute to a scalp topography that provides spatial distribution, the actual neural sources of these signals are difficult to pinpoint due to volume conduction (see Section Section 4.3). Even after averaging across trials to minimize

unrelated activity, this challenge persists. Although no definitive solution currently exists, combining EEG with other neuroimaging modalities, such as fMRI or MEG, provides physiological constraints under which the sources can be estimated reasonably (Michel and Brunet, 2019; Nunez and Srinivasan, 2006). Furthermore, confidence in the anatomical interpretation of ERP components depends heavily on the replication of findings across independent datasets (Luck and Kappenman, 2011, pp. 48, 55).

Auditory Evoked Potential (AEP) have been extensively studied, with the earliest systematic observation attributed to Davis (1939), who reported auditory responses in single-trial EEG data. AEPs are typically evoked by abrupt changes in auditory input, such as sound onsets, offsets, or rapid changes in continuous sounds. Such changes tend to recruit large neuronal populations in a synchronized manner, resulting in prominent scalp-recorded components (Winkler, Denham, and Escera, 2013). The high temporal resolution of EEG allows for the separation of auditory processing across different stages of the auditory pathway, ranging from brainstem responses (e.g., auditory brainstem response, ABR), to mid-latency responses from the medial geniculate body, and finally to long-latency cortical responses. Detecting responses from deeper brain structures generally requires a high number of repetitions to overcome attenuation due to distance from the source to the electrode and volume conduction. Cortical responses, on the other hand, are more readily accessible (Winkler, Denham, and Escera, 2013). This section focuses on the long-latency cortical AEPs, particularly the P1–N1–P2 complex (Figure 7B).

The P1-N1-P2 complex consists of two positive and one negative peak. The P1 is a positive deflection occurring approximately 50 ms after stimulus onset and is associated with activity in the primary auditory cortex (Godey et al., 2001). It has been linked to the early acoustic processing of auditory features (Haumann et al., 2021) and the segregation of auditory streams (Jaeger et al., 2020; Snyder, Alain, and Picton, 2006; Snyder et al., 2012). The N1 is a prominent negative deflection at around 100 ms, typically maximal at central scalp sites (Winkler, Denham, and Escera, 2013). It is a complex component arising from multiple generators, both in primary (lemniscal) and secondary (non-lemniscal) auditory cortices (Kohl, Parviainen, and Jones, 2022; Pantev et al., 1995) and is sensitive to stimulus features (Beauducel et al., 2000; Drennan and Lalor, 2019; López-Caballero et al., 2023), presentation rate (Lanting et al., 2013; López-Caballero et al., 2023), and attentional manipulations (Debener et al., 2002; Hillyard et al., 1973). Its amplitude and latency are influenced by the regularity of auditory input, although its exact mechanism remains debated (May and Tiitinen, 2010; Näätänen, 2001). The N1 shows both tonotopic and amplitude-specific organization, with the location of its cortical generator varying based on the frequency and

intensity of the auditory stimulus (Pantev et al., 1988, 1989). Following the N1, the P2 emerges between 175 and 200 ms post-stimulus. It exhibits a positive polarity over the vertex and has been shown to consist of two peaks, one in the auditory cortex anterior to those of the N1 (Bosnyak, Eaton, and Roberts, 2004; Steinmetzger and Rupp, 2023) and a later component in distributed areas of planum polare and planum temporale (Steinmetzger and Rupp, 2024). Although less well understood, the P2 has been shown to respond to the spectral complexity of sounds (Shahin et al., 2005), the tracking of rapid acoustic changes (Steinmetzger and Rupp, 2023), learning (MacLean et al., 2024), and increased attention (Snyder et al., 2012). Rather than providing an exhaustive overview of all reported functions and localization results, the present section is intended to highlight the temporal sequence of auditory processing and to illustrate how EEG captures the progression of information along the auditory pathway.

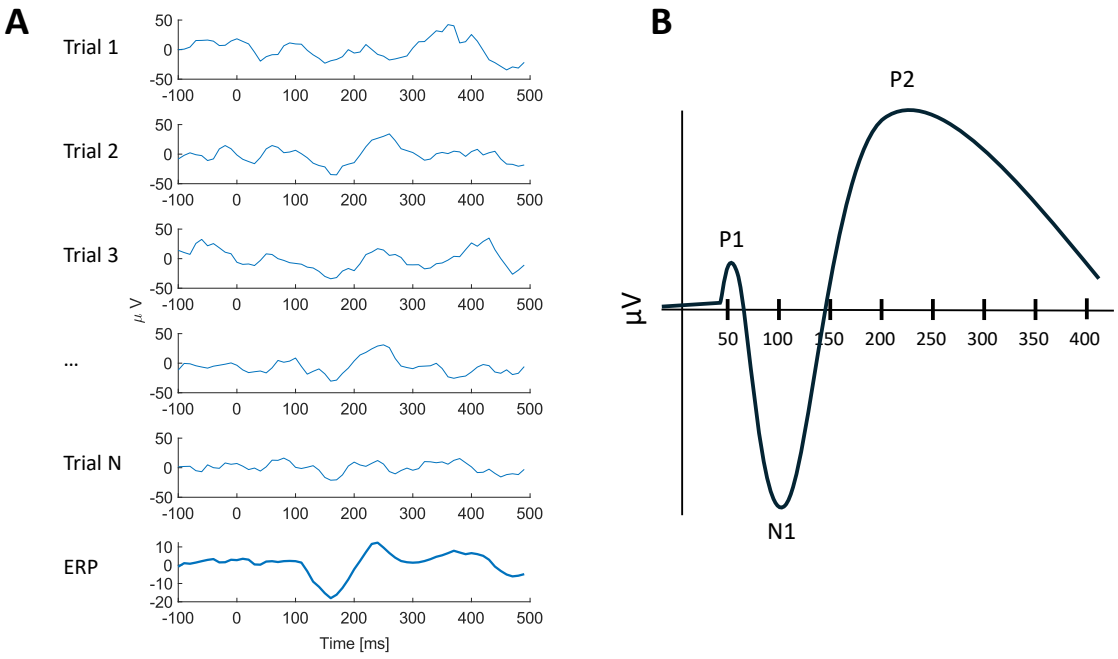

Figure 7: **A:** The EEG data is segmented, locked according to the external event that was being presented. Repeating this procedure results in several Trials (N). Averaging over these trials is done to increase the SNR and results in an ERP. **B:** Presentation of a tone elicits a neural response. Averaging over trials of repeated presentation of the same tone yields a AEP. This AEP, here simulated, is characterized by the typical P1-N1-P2 peak complex.

4.6 MOBILE EEG

When monitoring neural activity underlying auditory perception during natural behavior, minimizing disruption from recording hardware is crucial. Portable setups, termed mobile EEG, have been developed to address this need to measure perception BTL (Bleichner and Debener, 2017; Bleichner and Emkes, 2020; Debener et al., 2012). Mobile EEG systems can be evaluated along three dimensions: system specification, participant mobility, and device mobility. For example, maximal device mobility requires a fully head-mounted system, while full participant mobility allows for unconstrained running or other vigorous activity. High system specification, in turn, would demand gel-based, active, shielded electrodes, combined with high sampling rates, high bit resolution, long battery life, and reliable wireless transmission. In practice, no existing setup achieves the highest score across all three dimensions simultaneously. Importantly, not all research questions require maximum specification, but this framework highlights the trade-offs in mobile EEG development (Bateson et al., 2017).

In the context of everyday auditory perception, Hölle and Bleichner (2023) investigated sound processing in an office setting using an unobtrusive setup with cEEGrids, a neckspeaker, and a phone to record the signals. The system included an integrated amplifier and microphones to capture ambient sounds. AEPs were successfully obtained for tones presented via the neckspeaker, but not for automatically derived sound onsets in the ambient environment. This outcome highlights both the potential and the limitations of mobile EEG for real-world soundscape research. While neural activity can be recorded unobtrusively, the description of the acoustic environment is critical. As discussed in Section 3.1, soundscapes are highly complex, and simplistic representations such as indiscriminate sound onsets may be insufficient. Furthermore, while AEPs are widely used to study auditory perception and associated cognitive processes, they are limited in their applicability to naturalistic, continuous stimulation. Averaging across trials requires repeated presentations of the same stimulus, which rarely occurs in real-world listening situations. For this reason, classical ERP approaches are not suitable to capture how the brain processes continuous input, such as ongoing speech. Therefore, I will introduce methods for analyzing neural responses to continuous sound streams in Chapter 5.

5

TEMPORAL RESPONSE FUNCTIONS

"Probably the last sound heard before the Universe folded up like a paper hat would be someone saying, 'What happens if I do this?' "

Pratchett (2008)

Key Takeaways

- Temporal Response Functions enable the mapping between the input and the output of linear time-invariant systems.
- Forward models map stimulus (input) to neural data (output)
- Decoding models reconstruct the stimulus (output) from the neural data (input).
- Selection of features, which represent characteristics of the stimulus, influences the extent to which neural variability can be explained.

What you will learn:
The conceptual and mathematical formulation of Encoding and Decoding models and how they can be used to investigate continuous auditory perception.

5.1 THE CONTINUOUS WORLD

We experience the world in a continuous manner, with sensory impressions integrated into the perception of one seamless flow of experience. To better understand how the brain processes information in realistic settings, it is essential to analyze continuous time series data. Such analyses acknowledge the ongoing and dynamic nature of perception and align more closely with how sensory input is encountered in everyday life.

One promising approach to address this challenge is the use of Temporal Response Functions (TRF). TRFs provide a mathematical framework that describes the relationship between specific features of a stimulus and the neural responses using mapping functions. The are two directions of modelling: encoding and decoding. Both of these approaches fall under the broader framework of system identification, which refers to estimating a mathematical model that characterizes how an input signal is transformed into an output signal. In the case of an encoding approach, a mapping function from stimulus property to neural activity is derived. In the case of decoding the input and output switches Figure 8 (Marmarelis, 2004). The TRFs are rooted in the framework of Linear Time-Invariant (LTI) systems. LTI systems assume that the brain's response is proportional to the stimulus (linear) and stable over time (time-invariant). These assumptions make them particularly useful in this context, as they provide a simple yet powerful way to capture how continuous stimulus features consistently shape neural responses (Crosse et al., 2021; Holdgraf et al., 2017). This enables the estimation of interpretable models such as TRFs. While it is well recognized that the brain is neither stationary nor always scales linearly with the sensory input (Bünau, 2012; Drennan and Lalor, 2019), non-linear models, which could capture non-linear neural dynamics, are difficult to interpret (Crosse et al., 2021). Thus, LTI models offer a practical and interpretable approximation of how stimulus features are processed.

Returning to the auditory pathway illustrates this point: information is progressively transformed as it travels from the ear to the cortex, such that the raw acoustic signal no longer directly matches cortical activity. The question then becomes, how should the stimulus be represented so that it aligns with this transformed neural information? This line of reasoning reflects a central aim of neuropsychological research: to determine where and how properties of the external world are represented in the brain, and how these representations are used to guide perception and behavior.[1]

5.1.1 *Encoding Models*

The encoding models used in this thesis are a class of linear models that map stimulus features onto neural data. Conceptually, they function as transfer functions, testing whether a given feature is represented in the neural activity recorded with EEG. These

1 There is an interesting point to be made, in how far the concepts of psychological research describing behavior, perception, or cognition i.e. emotions, attention, working memory, are actually meaningful concepts for the brain. That is, rooting representations of stimuli in psychological concepts to explain neural activity. Conceptually, this refers to the outside-in approach. Although vital to hypothesis generation and scientific investigation, it would diverge too much from the point of this section and the interested reader is directed to the discussion between Poeppel and Adolfi (2020) and Buzsáki (2020).

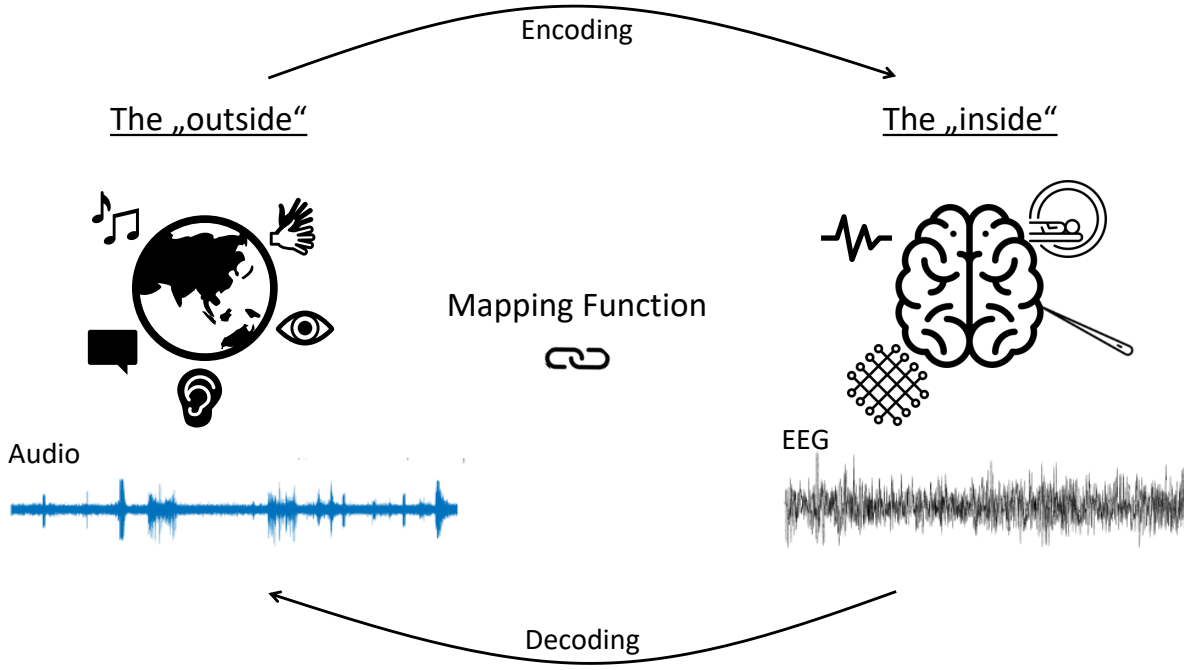

Figure 8: The TRF describes how an input is mapped to an output over time. In neuroscience, this framework links external stimuli to the recorded neural activity. The versatility of this approach allows linking a variety of external stimuli (i.e., auditory, visual) with neural recordings (i.e., EEG, fMRI, ECoG). In this example, the forward approach maps audio recordings onto the EEG data using the TRF. In the decoding approach, the audio is reconstructed from the EEG recording.

concepts (i.e., mapping functions, LTI systems) can be challenging to grasp, particularly for those coming from non-technical backgrounds. A useful way to think about these encoding models is that they provide a direct link between the activity of a brain area of interest and the sensory information driving it. One can imagine the EEG data as a fogged window, through which we see only a blurred reflection of the true underlying neuronal activity. Different aspects of the sensory signal can then be thought of as different pairs of glasses placed in front of this window. Each pair emphasizes certain features while obscuring others, thereby shaping what becomes visible. In this analogy, the glasses represent the mapping functions: they determine how raw stimulus features are transformed into the patterns of brain activity we observe. In practice, one might ask whether the cortex tracks changes in the amplitude of a sound signal. Here, tracking refers to systematic changes in the activity of neuronal populations in response to fluctuations in loudness.

To approach this, we first derive the envelope $s(t)$ of the sound signal over a recording period of time $t = 1...T$. A naïve starting point would be to ask: if the envelope increases at time point $s(t+1)$, does the neural response $r(t+1, c)$ (measured using

channel $c = 1...C$) at the same time point also increase? Repeating this across all time points and computing the inner product (assuming the signals are mean-centered) between the two signals yields a measure of covariance. The outcome, normalized by the variance, provides a measure of correlation.

However, this simple approach overlooks an important fact: changes in the stimulus are not instantaneously reflected in cortical activity. Due to biological processing delays, changes in the envelope at time $s(t+1)$ may only be observable in the neural signal at some later time $r(t+k, c)$. To capture this, we systematically introduce time lags by shifting the stimulus relative to the neural response and computing correlations across these delays. The result is the cross-correlation function, which reveals the temporal relationship between the stimulus feature and the neural response at each lag.

Finally, to fully understand how temporal response functions operate as encoding models, it is essential to account for the autocorrelation of the stimulus time series itself. In forward modeling, this step ensures that correlations between stimulus and response are not artificially inflated by the inherent structure of the stimulus feature, such as the temporal smoothness of an acoustic envelope (Figure 9).

In mathematical terms, we assume that the neural data $r(t, c)$ can be estimated through the convolution of the stimulus $s(t)$ with a channel-specific set of weights (TRF) $w(k, c)$.

$$r(t, c) = \sum_{k} w(k, n)\, s(t - k) + \varepsilon(t, c), \tag{1}$$

$\varepsilon(t, c)$ presents the error term of residual variance not explained by the model. Importantly, k defines a range of lags and has to be defined a-priori. The selection of lags $k = 1...K$ is typically based on the time range one would expect to observe cortical response components, which are based on laboratory studies. Interestingly, the $w(k, c)$ at lag $k = 100ms$ describes how one unit change in i.e. the envelope, affects the predicted neural response 100ms later. The way to derive the TRF weights $w(k, c)$ is through minimizing the Mean Squared Error (MSE) between the predicted $\hat{r}(t, c)$ and actual neural response $r(t, c)$.

$$\min \varepsilon(t, c) = \sum_{t} [r(t, c) - \hat{r}(t, c)]^2 \tag{2}$$

The minimization of the mean squared error can be solved using ordinary least squares (OLS). This approach provides a closed-form solution, in which the weight vector is estimated as

$$w = \left(S^T S\right)^{-1} S^T r, \tag{3}$$

Importantly, S is the lagged time series of the stimulus of dimension T x K, the envelope in this case. The matrix is zero-padded at non-zero lags to ensure causality (Mesgarani et al., 2009). Here, the similarity between the neural response and the envelope at each lag is computed and normalized by the autocovariance $\left(S^T S\right)^{-1}$ of the stimulus. This is an important step and distinguishes TRF estimation from cross-correlation, as autocorrelated stimuli, such as speech, would result in smearing. The result is a set of weights $w(k, c)$ at each time lag k.

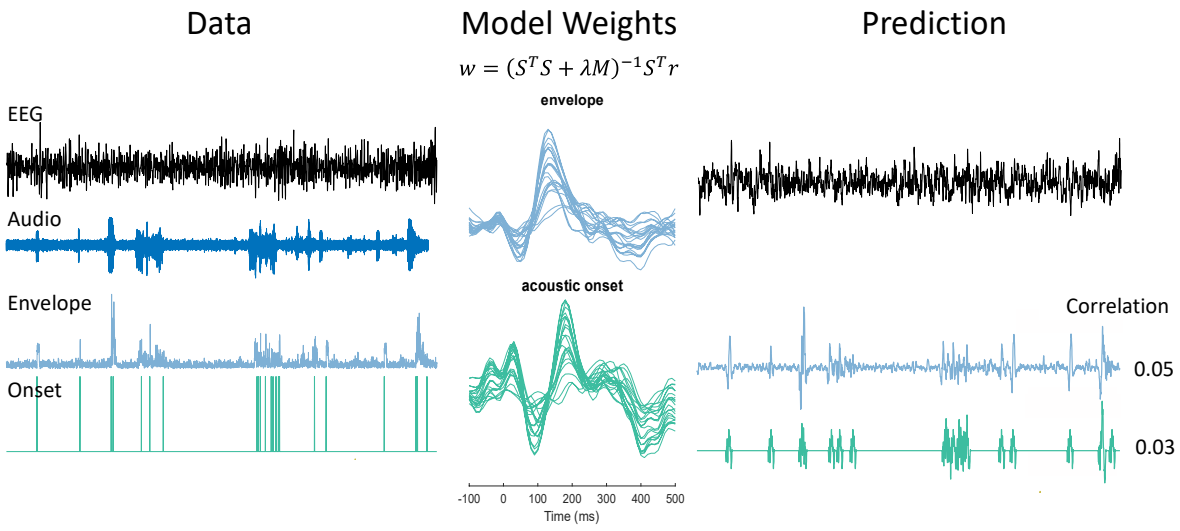

Figure 9: To investigate auditory perception, features of the audio can be mapped onto the EEG using TRFs to determine how they are processed. The derived model weights are used to predict the EEG data of a held-out test set, by convolving the model weights with the respective feature. The predicted time series is then correlated with the actual EEG data. Here, exemplarily, the envelope and sound onsets are used as features, and their respective model weights are shown. Furthermore, their respective predictions and correlations are displayed on the right side. Replicated from Haupt, Rosenkranz, and Bleichner (2025a), with permission).

5.1.2 *Decoding Models*

Decoding operates in the inverse direction compared to encoding. Instead of mapping stimulus features onto neural data, the goal is to reconstruct the stimulus from the recorded neural activity. While encoding models are primarily used to investigate how the brain processes information, decoding models reveal which stimulus properties are represented in the neural signals at a given point in time.

Returning to the earlier metaphor of the fogged window and glasses, encoding describes how different glasses shape the way the sensory input is projected onto the neural signals we observe. Decoding, in turn, asks whether we can look back through the blurred window, using the patterns in the neural activity, to infer which glasses were being worn and thus reconstruct what the original stimulus must have looked like.

The decoder uses the distributed information across EEG channels to predict how the stimulus evolved at different time lags. This contrasts with the encoding model, where a separate set of weights is derived for each channel, effectively resulting in a mass-univariate analysis. By combining information across all channels, decoding exploits not only shared variance related to the stimulus but also channel correlations, including those introduced by noise. This multivariate pooling typically leads to higher reconstruction accuracies compared to forward prediction in encoding models (Hebart and Baker, 2018; Kriegeskorte and Douglas, 2019). A consequence of this, however, is that the weights of the backward model should not be interpreted neurophysiologically, as large weights do not necessarily represent being critical for the stimulus (Haufe et al., 2014).

In practice, the decoder weights $g(k, c)$ are derived analogously to the encoding, merely the order of response and stimulus changes, as we try to minimize the MSE of the actual stimulus s and the predicted $\hat{s}$.

$$g = \left(R^T R\right)^{-1} R^T s, \tag{4}$$

where R refers to the time-lagged neural response. Since backward modeling is acausal, that is, the mapping of the neural data to the stimulus is backward in time, the lags should consider this.

5.2 REGULARIZATION

A central challenge in both encoding and decoding models is overfitting. Ordinary least squares provides an optimal solution for a given dataset, but this solution may not generalize well to new data. For models fitted and tested on the same participant, as is the case in this thesis, this is due to the noisy[2] nature of neurological data (Crosse et al., 2016; Wong et al., 2018). The regularization can be tuned to make the resulting weights less specific to the dataset it was trained on and thus perform better on unseen data.

Another issue is that the solution to the inverse of the autocovariance matrix can be ill-posed, resulting in numerical instability. This can impact the resulting model weights to be inflated and thus suboptimal. Regularization can reduce the variance by adding a bias term, improving the estimated inverse covariance matrix. Lastly, the space over which features can be fitted can be unconstrained (Diedrichsen and Kriegeskorte, 2017; Kriegeskorte and Douglas, 2019). That means that any number of features could yield the same feature weights. In order to constrain this space, regularization can be applied.

In the context of encoding and decoding, regularization involves adding a bias term to the autocovariance matrix. A commonly used framework is ridge regression, for which the mathematical expression looks as follows:

$$w = \left(X^T X + \lambda I\right)^{-1} X^T y, \tag{5}$$

here the magnitude of λ controls the amount of bias and has to be tuned. In practice, this involves testing several different lambda values and choosing the most optimal solution. In this thesis, the most optimal solution is the set of model weights that yields the highest reconstruction/ prediction accuracy.

Several types of regularization exist, including ridge regression (Wong et al., 2018), Tikhonov (Tikhonov and Glasko, 1965) regularization, and lasso (Tibshirani, 1996) methods. In a comparative study, Wong et al. (2018) found that regularization had the strongest impact on decoding performance, while its effect on forward encoding was limited. The authors attributed this difference to the fact that the inverse problem for neural data is particularly prone to being ill-posed. Nevertheless, regularization remains crucial for ensuring stable and interpretable solutions in both modeling approaches.

2 refer to Section 4.3.2 for an explanation of noise in neural data.

5.3 FEATURE SELECTION

An important consideration when building encoding or decoding models is the selection and transformation of stimulus features. The brain does not respond in a purely linear manner to sensory input, and this non-linearity must be taken into account when preparing features (Buzsáki and Mizuseki, 2014; Rahman et al., 2020). For example, neural responses to aspects of sounds such as loudness are known to be non-linear (Drennan and Lalor, 2019). To address this, features can be linearized by applying appropriate non-linear transformations prior to modeling (Crosse et al., 2021). This step ensures that the features fed into linear models more accurately reflect how the brain processes sensory information.

Another key aspect is that most natural stimuli are not adequately described by a single univariate feature. Instead, they are inherently multidimensional. Speech, for example, can be represented at multiple levels of description, ranging from acoustic properties such as the envelope, onsets, and spectrogram, to higher-level representations such as phonetics and semantics (Brodbeck, Presacco, and Simon, 2018). Incorporating multiple features $f = 1...F$ into a model allows these complex and interrelated aspects of the stimulus to be represented jointly. In practice, this is achieved by concatenating the features, resulting in a matrix of size TxFk.

At last, echoing the proximal and perceptual soundscape, features regarding relevancy cannot be obtained from the soundscape alone. That is, if a certain sound is a target is not directly apparent from the soundscape alone.

By analyzing multiple features, one can capture not only the unique contribution of each feature but also their interactions. This multivariate approach offers a richer and more realistic account of how the brain processes naturalistic stimuli, moving beyond single-feature explanations to provide a comprehensive understanding of neural encoding. I aim to answer the open question, which features of the soundscape are most relevant for capturing neural responses. To address this, it is necessary to first examine sound–brain relationships under controlled conditions, providing a reference framework that can then be extended to more naturalistic settings. This issue is the focus of Chapter 7 in this thesis.

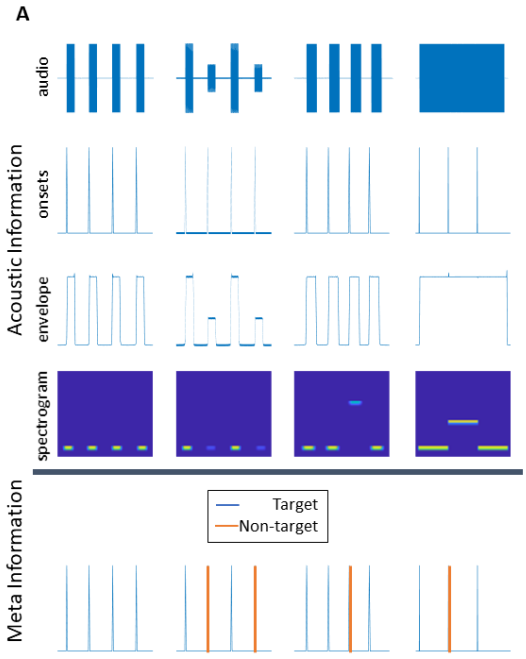

Figure 10: Several features can be derived from the soundscape, each differing in their descriptive properties. Where onsets mark the beginning of a sound, they do not contain continuous amplitude fluctuation information. Conversely, envelopes do not depict changes in the frequency contents. The type of acoustic information that can be characterized through specific features depends on the research question and the quality of the audio recording. Meta information, such as whether a tone is perceptually relevant by being declared a target, is not apparent from acoustics alone. Replicated from Haupt, Rosenkranz, and Bleichner (2025a), with permission.

6

OBJECTIVES

"The purpose of a storyteller is not to tell you how to think, but to give you questions to think upon."

Sanderson (2010)

This thesis aims to advance the understanding of real-world soundscape perception using mobile EEG neuroimaging. While studies have already attempted to measure soundscape perception in everyday life and succeeded in obtaining ERPs to artificial stimuli, they lacked a foundational understanding of which aspects of the soundscape are driving the neural response. To address this, this thesis adopts a stepwise, nuanced approach by first investigating complex soundscapes in controlled settings and then moving to more uncontrolled settings beyond the lab. This balances ecological validity with experimental control.

The first study addresses the foundational question of which features of a naturalistic soundscape elicit measurable EEG responses in a controlled setting. Participants listened to a naturalistic soundscape, for which complete annotation was present, while EEG was recorded using mobile-compatible hardware. Using encoding (TRF) models, we predict the brain's response to both bottom-up acoustic features (e.g., onset, envelope, spectral modulation) derived from the proximal soundscape and higher-level annotations (e.g., task relevance) representing top-down perceptual interpretations. This approach allows for a direct comparison of how well each type of feature explains neural variance. The findings provide a conceptual and methodological basis for interpreting EEG responses in later, less controlled environments and clarify the limits of what can be inferred from proximal acoustic data alone.

Building on the first study, the second experiment investigates whether contextual information, specifically, the temporal proximity of sound onsets, modulates EEG responses. This was investigated under the constraint that only proximal soundscape information is available. Here, we analyzed the EEG data of the first study and used

inter-onset intervals as a proxy for contextual acoustic structure. The results reveal that EEG responses decrease with increasing onset proximity, suggesting adaptation effects that generalize previous findings from isolated tones to continuous, complex soundscapes. This demonstrates that contextual information can be inferred from the acoustic signal (proximal soundscape) alone and shapes neural response dynamics in naturalistic, dynamically changing environments.

The third study tests the generalizability of EEG-based auditory attention decoding by extending established lab paradigms into real-world contexts using a minimal and mobile setup (cEEGrids, a portable hearing lab, smartphone). Participants listened to narrated stories (either one or two) under progressively more naturalistic conditions: starting in the lab with free-field playback, followed by added cafeteria noise, then moving outside the lab to a seated public hall, and finally walking along a street. This stepwise immersion allowed us to evaluate whether findings from controlled environments hold in mobile, everyday settings and to assess how real-world movement and ambient noise affect EEG signal quality. Importantly, we observed that while attention decoding is feasible using a minimal, mobile setup in stationary settings, movement introduces artifacts that can obscure neural responses. These were only detectable via forward modeling, which underscores the need for careful investigation of the underlying signal. This study thus complements the previous two by testing whether the theoretical and methodological insights derived from controlled experiments can generalize to real-world auditory perception, thereby supporting the broader aim of developing robust neuroimaging approaches for investigating soundscape perception in daily life.

Together, these three studies establish an empirically grounded approach to studying naturalistic sound perception. In the general discussion, I will put the results in the context of my personal motivation for investigating auditory perception in everyday life situations. Furthermore, I will discuss and critically reflect on what types of information EEG can meaningfully capture in naturalistic settings and the pitfalls of model interpretation. At last, a future outlook on how these results can be used for future studies will be presented.

Part II

STUDIES

This part showcases the empirical work that has been published or submitted to peer-reviewed and open access journals.

7

RELEVANT SOUNDSCAPE FEATURES

"He hadn't changed in one giant leap,
but across a million little steps."

Sanderson (2017)

Exploring Relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes

Thorge Haupt[1], Marc RosenKranz[1], Martin G. Bleichner[1,2]

[1]Neurophysiology of Everyday Life Group, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
[2]Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

This chapter is identical in content to the version published in:
Thorge Haupt, Marc Rosenkranz, and Martin G. Bleichner (Jan. 2025a). "Exploring Relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes." en. In: *eNeuro* 12.1. Publisher: Society for Neuroscience Section: Research Article: New Research. DOI: 10.1523/ENEURO.0287-24.2024

ABSTRACT

A comprehensive analysis of everyday sound perception can be achieved using Electroencephalography (EEG) with the concurrent acquisition of information about the environment. While extensive research has been dedicated to speech perception, the complexities of auditory perception within everyday environments, specifically the types of information and the key features to extract, remain less explored. Our study aims to systematically investigate the relevance of different feature categories: discrete sound-identity markers, general cognitive state information, and acoustic representations, including discrete sound onset, the envelope, and mel-spectrogram. Using continuous data analysis, we contrast different features in terms of their predictive power for unseen data and thus their distinct contributions to explaining neural data. For this, we analyse data from a complex audio-visual motor task using a naturalistic soundscape. The results demonstrated that the feature sets that explain the most neural variability were a combination of highly detailed acoustic features with a comprehensive description of specific sound onsets. Furthermore, it showed that established features can be applied to complex soundscapes. Crucially, the outcome hinged on excluding periods devoid of sound onsets in the analysis in the case of the discrete features. Our study highlights the importance to comprehensively describe the soundscape, using acoustic and non-acoustic aspects, to fully understand the dynamics of sound perception in complex situations. This approach can serve as a foundation for future studies aiming to investigate sound perception in natural settings.

Significance Statement

This study is an important step in our broader research endeavor, which aims to understand sound perception in everyday life. Although conducted in a stationary setting, this research provides foundational insights into necessary environmental information to obtain to understand concurrent neural responses. We delved into the analysis of various acoustic features, sound-identity labeling, and cognitive information, with the goal of refining neural models related to sound perception. Our findings particularly highlight the need for a thorough analysis and description of complex soundscapes. Our study provides key considerations for future research in sound perception across various contexts, from laboratory settings to mobile EEG technologies, and paves the way for investigations into more naturalistic environments, advancing the field of auditory neuroscience.

7.1 INTRODUCTION

Mobile electroencephalography (EEG) has opened new avenues for studying neural activity Beyond the Lab (BTL), offering insights to expand our current understanding of brain function in real-world settings (Gramann et al., 2011; Hölle and Bleichner, 2023; Studnicki, Downey, and Ferris, 2022). One critical aspect of interpreting EEG data recorded BTL is understanding the environmental context driving the neural response (Holdgraf et al., 2017; Robbins et al., 2021). Thus, sufficient information about the environment needs to be captured to accurately determine how it influences neural responses. For BTL recordings, however, there is typically a tradeoff between the aspects that can be captured and the overall mobility implicating unobtrusive recordings (Bateson et al., 2017). Therefore, a selective approach is required, focusing on environmental aspects that are within the scope of the setup and are most pertinent to the research objectives (Bleichner and Debener, 2017).

In the study of auditory perception, particularly in understanding how we perceive naturally occurring sounds in everyday life, it has yet to be shown, which auditory features are most relevant to understand the concurrent neural response.

One well-studied feature category depicts the acoustic properties of the soundscape on different levels of abstraction (Hamilton et al., 2021; Heer et al., 2017). Starting from the most abstract, discrete sound onsets, to continuous broadband amplitude changes captured by the envelope, to the mel-spectrogram depicting power fluctuations across different frequency bands over time (Figure 11A). However, the choice of abstraction is crucial; with greater abstraction relevant fine structure information about the audio is lost.

Beyond the acoustic depiction, additional layers of information (meta-information), as illustrated in Figure 11B, should be considered for a comprehensive understanding of auditory processing (Robbins et al., 2021). For instance, Zuk, Teoh, and Lalor (2020) have demonstrated that the EEG response to music and speech stimuli is stronger compared to other natural sounds. Here, knowing the sound identity, i.e., the category of a sound, such as speech, rustle, etc., is a valuable insight for neural model building (Figure 11B). Beyond the soundscape description, the cognitive state, or 'cognitive priors', of the participant, such as how attentional resources are allocated, profoundly affect how auditory information is processed and, consequently, the resulting EEG responses (Holtze et al., 2021; Obleser and Kayser, 2019; Rosenkranz et al., 2023).

From a pragmatic perspective, the effort required to acquire various auditory features can vary significantly. Where Acoustic features can be directly and readily obtained from the auditory waveform, deriving meta-information presents a greater

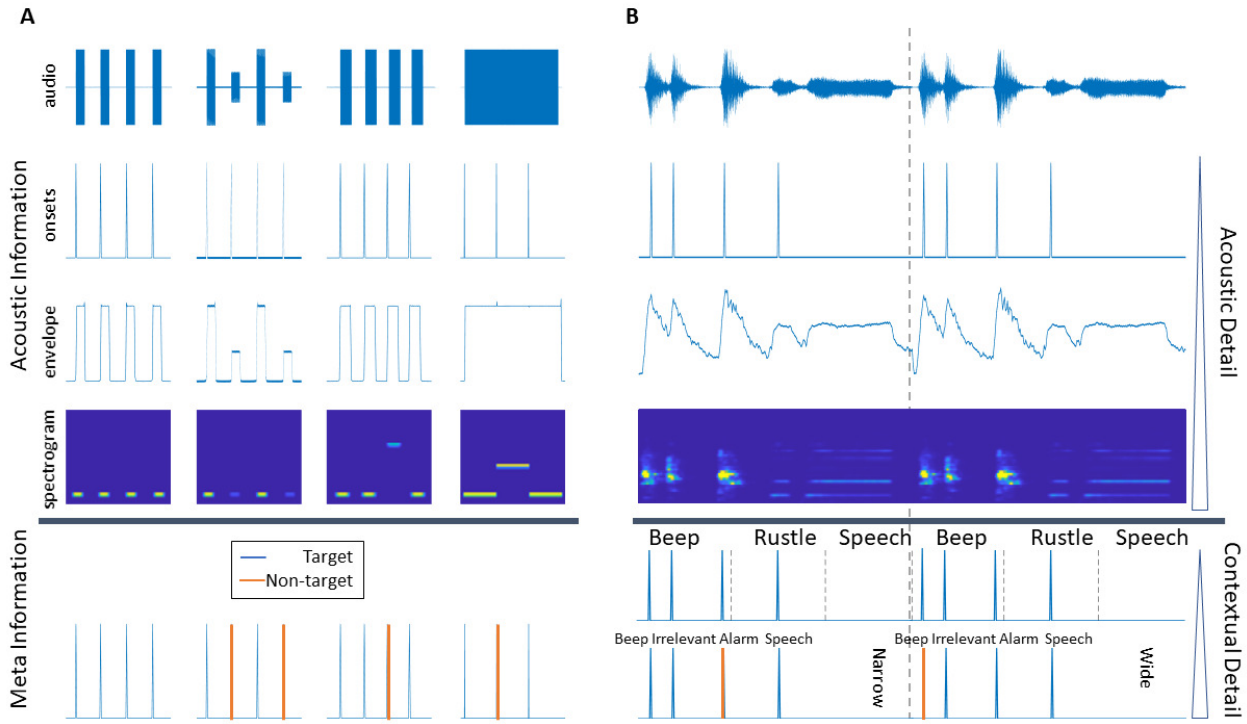

Figure 11: **A:** This panel illustrates the levels of abstraction with which acoustic features depict different tone sequences. Furthermore, it highlights the impact of the acoustic detail the acoustic onset, envelope, and mel-spectrogram capture. In the first tone sequence, onsets reveal the timing of sound occurrence, while the envelope extends this information by showing both the timing and duration of the sound; the mel-spectrogram further adds details about the frequency content. For the second tone sequence, the limitation of onsets becomes apparent, as they do not convey any information about amplitude changes. In the third sequence, the frequency changes are distinctly and exclusively captured by the mel-spectrogram, with the onset and envelope methods remaining indifferent to these frequency variations. The fourth sequence underscores a similar limitation of the envelope, which fails to distinguish frequency variations in a continuous tone. The lower section of this panel demonstrates the crucial role of meta-information. For instance, understanding which sound in an experiment requires a behavioral response cannot be discerned from the acoustic features alone. Meta-information thus complements the acoustic analysis by providing essential contextual and functional information. **B:** This panel showcases acoustic feature representations—onsets, envelope, and mel spectrum – for a complex naturalistic soundscape. Additionally, it highlights the diversity of meta-information that can describe these soundscapes. The upper row features continuous classification from Yamnet, providing labels for individual sound segments. The lower row presents manual annotations, categorizing sounds into different conditions (narrow and wide) and distinct types (beep, irrelevant, alarm, and speech sounds). These labels, informed by task-specific knowledge, offer deeper contextual and descriptive insights into the soundscape, illustrating the multi-layered nature of acoustic analysis in naturalistic settings.

challenge. These cannot be directly extracted from the audio signal but require additional processes (Robbins et al., 2021).

The question, then, is which features are relevant to the construction of neural models, particularly in the context of naturalistic soundscapes. The current study aims to explore the optimal feature selection and the impact of including sound identity and cognitive priors in neural models, bearing in mind the practical limitations of data collection in natural settings.

Our approach employs continuous forward modeling to investigate the link between the derived features and their representation in the neural signal (Crosse et al., 2016; Ding and Simon, 2012b; Hamilton and Huth, 2020; Holdgraf et al., 2017). This methodology allows for a nuanced understanding of the neural underpinnings of auditory perception and examine our two primary study goals: First, if sound identity and cognitive prior information enhance the estimation of neural response models. This involves examining whether meta-information provides a significant advantage in modeling the neural correlates of auditory perception.

Second, we seek to understand the extent to which the level of acoustic detail influences the effectiveness of neural response model estimation for natural soundscapes. This aspect of the study focuses on assessing how varying degrees of detail in acoustic feature representation impact the accuracy and robustness of our neural models.

By addressing these two objectives, our research contributes to a deeper understanding of how different types of auditory information are encoded in neural responses, particularly in rich, real-world acoustic environments.

7.2 METHOD

7.2.1 *Data Set*

7.2.1.1 *Experimental Design*

The analysis in this study is based on an existing dataset by Rosenkranz et al. (2023). The study aimed to investigate how altering attentional focus affects the perception of auditory information. This was done in two separate conditions, where participants had to respond to two different tones while they engaged in a complex audio-visual motor task. These tones are integrated within a soundscape designed to mimic a surgical auditory environment.

Specifically, the complex task involved an adapted 3-dimensional version of a Tetris game, where participants received occasional auditory instructions to place a specific

Tetris piece at a distinct location. The auditory background soundscape consisted of various sounds, including alarms, monitor beeps, and speech as they are common in an operation room. Additionally, a non-relevant coherent conversation split into smaller segments was played to simulate background conversation.

In each of the conditions, the participants listened to the same audio track. Only the instruction regarding the specific tone participants had to respond to changed. In the first condition, participants were instructed to press the space bar when hearing an alarm sound, which introduced a narrow focus. In the second condition, participants had to respond to beep tones which were played from multiple directions together with other sounds, necessitating a more comprehensive (wide) attentional focus. For both conditions, the task-relevant instructions had to be monitored concurrently. In this study we refer to conditional differences as cognitive priors, as these have been shown to induce two different brain states. For a detailed procedural account, please refer to the open source material of the previous study: Rosenkranz et al. (2023).

7.2.1.2 *Auditory Stimuli*

The soundscape included three added tones, each of which was played 48 times. A hospital alarm and a task-irrelevant hospital monitor sound, which each lasted approximately 200ms. The last sound that was included, a beep tone, was generated using Matlab, with a frequency of 800Hz, and a duration of 60ms.

Participants heard 12 times one of four instructions they had to comply with ("Place the next stone in the [upper left | lower left | | upper right | lower right] corner"). The same instruction was never played consecutively. The irrelevant speech segments were taken from podcast conversations unrelated to the task and medical setting. There were in total 48 task-irrelevant speech segments that were presented in a pre-defined order and only once. Each snippet lasted approximately $3.5(\pm 1.5)$ seconds and was extracted using Audacity®.

All the extracted sounds were processed in Matlab, such that their Root-Mean-Squared (RMS) value matched that of the average RMS of all sounds. Adjusting for differences in loudness was achieved by applying specific gain parameters to each auditory sound. Additionally, the tones were spatially separated using the head-related impulse function (Kayser et al., 2009).

7.2.1.3 *Code Accessibility*

The code described in the paper is freely available online at `https://github.com/ThorgeHaupt/RelevantFeaturesSoundPerception`. The analyzed dataset can be found

under https://zenodo.org/records/7147701. A Dell Precision 3650 Tower running Microsoft Windows 10 Education was used.

7.2.1.4 *Data Acquisition*

For measuring the EEG, participants were fitted with 24 Ag/AgCI passive electrodes according to the 10-20 international system (EasyCap GmbH, Hersching Germany). The collected data was amplified using a wireless SMARTING system (mBrainTrain, Belgrade, Serbia) and referenced to Fz, and grounded to AFz. The data was sampled at 500Hz and the impedance of electrodes was kept below 20Ω before the recording. The audio presented during the experiment was sampled at 44.1 kHz. All data streams recorded were synchronized using the Lab Recorder software, which is based on the Lab Streaming Layer. Before measuring EEG data, participants were informed about the procedure and had to sign the informed consent.

7.2.1.5 *Preprocessing of EEG Data*

The EEG data was preprocessed in MATLAB (version 2021a, MathWorks, Natick, MA) using the EEGlab plugin and custom scripts. For detecting artifacts in the data we ran an Independent Component Analysis (ICA). To obtain optimal ICA weights, Winkler et al. (2015) proposed a pre-processing pipeline that runs separately from the actual pre-processing of the data that will be analysed later on. In other words, the pre-processing of the data for the ICA analysis does not impact the data used for analysis later on, as merely the ICA weights are extracted and added back to the raw data. In particular, we merged the two experimental conditions for each participant. Subsequently, the data has been resampled to 250Hz, after which a high- and then a low pass filter were applied (*pop_firws*(EEG, 'fcutoff',1,'ftype','highpass','wtype','hann','forder', 568), *pop_firws*(...,'fcutoff',42,'ftype','lowpass',...,'forder',128)). The cutoff frequencies were chosen to eliminate drifts and line-noise to obtain optimal ICA weight estimates (Winkler et al., 2015). Furthermore, channels with poor signal quality were identified and removed using the *clean_channels* function. Lastly, the data was segmented into 1-second epochs and converted to double digits. Artifactual epochs were removed using the *pop_jointprob* function using a threshold of 3 standard deviations.

The ICA was computed using the *pop_runica* function utilizing the extended version. The weights were added back to the raw and unfiltered data of each condition. Next, the ICA components were automatically flagged as either, muscle, eye, heart,

line noise, or channel noise using the *pop_icaflag* function within a corresponding, specified threshold range of probabilities ([0.7 1;0.7 1;0.6 1;0.7 1;0.7 1]).

Following artifact removal, the raw data were filtered using the same filter setup as previously described, with slight modifications to the filter order and cutoff frequencies. First, the lowpass filter was applied: fcut 20Hz, forder: 100. The data was subsequently resampled to 100Hz and subjected to high-pass filtering: fcutoff 0.3Hz, forder 518. The choice of lower filter order for the low pass filter was motivated by recommendations of Crosse et al. (2021), as to minimize artifacts caused by sharp roll-over introduced by higher-order filters. Reducing the passband to [0.3 20] Hz was done according to literature from speech tracking, demonstrating that most auditory processing activity is found in the lower frequency ranges (Crosse et al., 2016; Di Liberto, O'Sullivan, and Lalor, 2015). Finally, the data was referenced to the mastoids (TP9/TP10).

7.2.2 *Sound Features*

A full description of the soundscape is available, including acoustic information, the label of specific sounds (sound identity), behavioral responses, and conditional information. Here, we derived three types of features: Acoustic Features (AC), Sound Identity (SI) markers, and Cognitive Priors (CP).

The acoustic features that we derived in our study were acoustic onsets (i.e. transient sound events), the envelope, and the mel spectrum. These features vary in their level of acoustic detail, with acoustic onsets being the most sparse and the mel spectrogram providing the most detailed representation. Importantly, acoustic onsets are discrete in contrast to the continuous envelope and mel-spectrogram. This has crucial consequences for predicting unseen data, as for sufficiently spaced onsets, not every sample can be predicted (Figure 12). These features have been validated in many different studies using speech stimuli (Brodbeck, Hong, and Simon, 2018; Daube, Ince, and Gross, 2019; Desai et al., 2021; Heer et al., 2017; Mesik and Wojtczak, 2023). Here we want to test whether these features can be extended to soundscapes with non-speech soundscapes.

Sound identity and cognitive priors were available through the meta-information. This information is not necessarily readily available from everyday recordings (Hölle and Bleichner, 2023). For the current study, we selected a subset of tones for which sound identity markers and cognitive prior information were present; alarm, beep,

and irrelevant tone. The experimental relevance of the first two tones changed between conditions, whereas the last tone remained irrelevant throughout the experiment.

In addition to directly comparing different features, we also explored multi-feature models. For instance, we combined several acoustic features' information to assess whether their properties were processed differently in the brain. A combination of acoustic onsets and the envelope yielded a model we referred to as OnsEnv. Alternatively, we investigated the effect of providing SI and CP information in addition to acoustic features. Here, combining acoustic onsets with alarm tones yielded a model we referred to as AlarmOns. This approach proved particularly valuable for assessing overall model improvement when incorporating additional information and examining the overlap and unique contributions of each feature. Throughout the analyses, we employed the combination of different features multiple times.

It is important to note that when combining features, it was advised to normalize them to a common scale. This was crucial as the magnitude of the feature impacted the derivation of the model weights (Crosse et al., 2021). Thus, when different features were combined, we applied a min-max normalization, such that the values were bound to be between [0 1]. We opted for this particular normalization as it transformed each non-binary feature to share the same scale as that of the acoustic onsets and sound identity marker.

7.2.2.1 *Acoustic Onsets*

Acoustic onsets were derived for the entire soundscape, indicating the time point when the intensity of the soundscape exceeded a predefined threshold. The acoustic onsets here were represented as binary vector, where 1 indicates sound onset and 0 no sound onset. As we are interested in the impact acoustic detail has on explaining neural variability, we opted for the most parsimonious representation of the soundscape: the onset of a sound. The extraction of onsets occurred unsupervised, thus for any sound, indiscriminate of experimental relevance or presence of meta-information.

To derive the onsets, we used an energy novelty function (Müller, 2021). Initially, the audio waveform was squared and smoothed using a Hann window (length= 2048 and overlap=128 samples). The resulting smoothed signal was then logarithmically compressed with a gamma scaling factor of 10 to approximate the perception of sound intensities in humans.

Subsequently, the difference function was applied to obtain the rate of change in sound intensity, which was further smoothed. The resulting approximation of the first-order derivative was half-rectified to produce the final output of the energy novelty

function. This smoothed rate of change in sound intensity served as a representation of sound gain.

To identify acoustic onsets, we employed a threshold peak detection function. In this function, a peak was detected when the sound gain exceeded the threshold value. The resulting feature vector was binary, where a value of one indicated the presence of a sound onset.

7.2.2.2 *Envelope*

A plethora of studies have shown that EEG-measured neural activity tracks the acoustic envelope of auditory stimuli (Drennan and Lalor, 2019; Giraud and Poeppel, 2012; Luo and Poeppel, 2007), even with transparent setups (Holtze et al., 2022; Mirkovic et al., 2016). There are different ways to derive the acoustic envelope (Oganian and Chang, 2019; Petersen et al., 2017). To highlight that not only the feature but also the way of deriving the envelope impact model estimates, we have used two different ways. For all envelope computations, we first converted the stereo audio recording to mono and computed the envelope over the entire signal.

The first method, as described by Petersen et al. (2017), involves taking the absolute value of the Hilbert transform of the audio. To ensure a smooth envelope, a third-order Butterworth low-pass filter with a cut-off frequency of 30 Hz was applied to the resulting signal. Finally, the envelope was downsampled to match the sampling rate of the EEG signal.

Second, the mTRFenvelope function from the mTRF toolbox was used. This method involved computing the signal's power, followed by resampling through a moving average filter. To compress the signal, the square root of the power was applied, using a compression parameter of $\log_{10}(2)$.

The resulting envelopes are highly correlated rho = 0.94 indicating that they contain almost the same information content. As can be seen from Figure 16, the spectrum differences occur mostly in the below 1Hz spectrum and do not follow any systematicity. The stark deviation around 0Hz is caused by an offset. Given the high correlation and almost non-existent frequency differences, we inspected the estimated model weights. Here, edge artefacts for the envelope are visible at extreme latency values. Reducing the time lags manually and thus removing the edge artefacts did not impact the model performance.

In conclusion, albeit being very similar, the mTRF contains more power in the lower frequency ranges, which is highly relevant for tracking of soundscapes (Deoisres et al., 2023; Howard and Poeppel, 2010). We believe that this may be the reason why the

mTRF envelope outperformed the envelope. Henceforth, only the results of the mTRF envelope are considered.

7.2.2.3 *Mel-Spectrogram*

The mel-spectrogram is a time-frequency decomposition of a continuous signal. The 40 frequency bins are spaced according to the mel-scale, which has been shown to mimic the auditory perception of the human ear (Stevens, Volkmann, and Newman, 1937). We have used the MIR-toolbox (Lartillot, Toiviainen, and Eerola, 2008) function with the following settings: *mirenvlope* ('audio.wav','Spectro','Mel','Sampling','100'). The output was adjusted to the length of the EEG signal.

The mel-spectrogram offers the most acoustic detail out of the three acoustic features selected. While the broadband envelope depicts amplitude changes over time, the mel-spectrogram represents power changes in the different frequency bins over time. Like the envelope, the mel-spectrogram describes the soundscape at every sample.

7.2.2.4 *Sound Identity Markers*

Part of the original experiment was the presentation of three sounds that were embedded in the soundscape. Through available meta-information we extracted their marked onset and experimental condition information. We embedded this information in a feature vector, for each tone and condition respectively, in the form of ones and zeros. Here, ones mark the onset of the tone, whereas every other sample is 0. This is in contrast to the acoustic features, which solely comprise acoustic data and lack information pertaining to the sound identity and conditional information.

Similar to the acoustic onsets, however, sound identity markers are discrete. Thus, using these features we cannot make predictions regarding unseen data at every sample point. Due to the sound specificity of the features we can predict even fewer samples compared to the acoustic onsets. This yielded a tradeoff between building highly sound specific models and a proportion of explainable data (Figure 12).

7.2.2.5 *Cognitive Priors*

Besides acoustic characteristics, the cognitive state of the perceiver critically impacts the neural response to any type of auditory stimulation. For example, a cognitive prior could be the instruction given to a participant to pay attention to a specific sound. When a participant is instructed to focus on a particular sound, the neural

response is generally larger compared to when the sound is unattended (Rosenkranz et al., 2023). This information cannot be derived from acoustic properties alone but is crucial for understanding the neural response.

In the current data set, there were two conditions where participants receive separate instruction: attend to the alarm sound (narrow), or attend to the beep tone (wide). The attentional manipulation should impact the perception of the soundscape and thus the underlying neural activity. Including this information in the model estimation should lead to more neural variability being explained. Specifically, we have integrated this information by building models separately for the different conditions. Exemplary for the alarm tone, we used two feature vectors, one indicating the sound onset in the narrow condition and another vector depicting the alarm tone onsets in the wide condition.

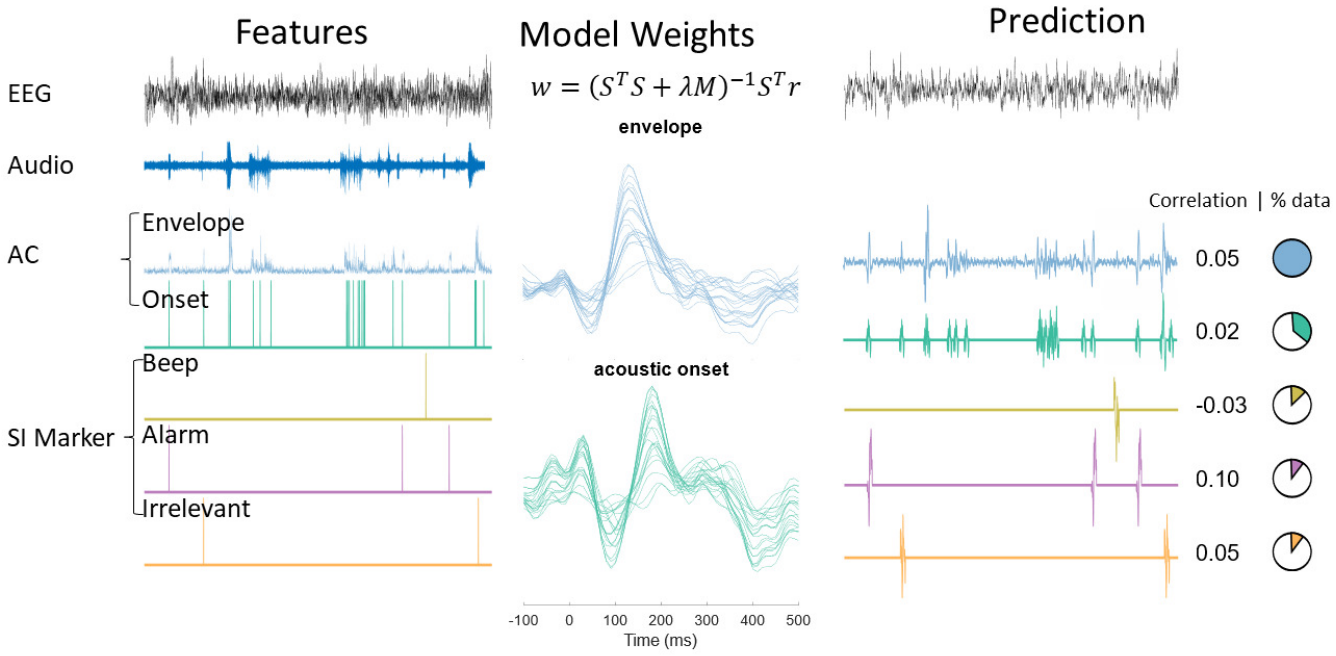

Figure 12: This figure depicts the process of forward modeling. From the audio several features are extracted. Using the EEG data and the features the temporal response function is derived for each EEG channel. This temporal response function is then used to predict unseen neural data based on the corresponding feature information. The resulting prediction is correlated with the actual signal. If discrete features are used, not every sample can be predicted. This is visualized in the pie charts, showing the proportion of samples that can be predicted. AC = acoustic features, SI = sound identity.

7.2.3 *mTRF*

To analyze the neural time series, we employed the mTRF toolbox developed by Crosse et al. (2016) in MATLAB. This toolbox enables one to derive a set of weights, which establish the relationship between the output of a system to a given set of input vectors through convolution. In our study the output of the system is the instantaneous neural time series denoted as $r(t, c)$, where $c = 1, \ldots, C$ represents the channel index and $t = 1, \ldots, T$ the time points. The measured time series can be modeled as the convolution of a set of channel c-specific weights w at a distinct timelag τ with the stimulus features shifted by τ. This approach aims to replicate the brain's processing, wherein the response to a stimulus is not immediate but rather delayed by an unknown duration. The activity not accounted for by the response weights is captured by the residual term $\varepsilon(t, c)$.

$$r(t, c) = \sum_{\tau} \omega(\tau, c) s(t - \tau) + \varepsilon(t, c)$$

In this study, each of the features' set of channel weights was used to interpret the morphology and topography, investigate model performance, compute multivariate models, and utilized for cross-prediction.

The determination of the TRF involves solving an optimization problem aimed at minimizing the MSE between the actual and predicted neural time series:

$$\min_{\epsilon(t,n)} = \sum_{t} [r(t, c) - \hat{r}(t, c)]^2$$

The solution is obtained by computing the weight vector $\mathbf{w}$:

$$\mathbf{w} = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \mathbf{r}$$

Where S is the design matrix containing the stimulus time series at the different time lags τ. The dimensionality of the resulting design matrix is sample points by numbers of features and lags ($T * N^{ft}\tau$). In this study, this would yield weight dimensionality of 40x61x22 for the mel-spectrogram and 1x61x22 for all other AC and SI features. Furthermore, the stimulus matrix is padded with zeros at non-zeros lags to ensure causality (Mesgarani et al., 2009). The matrix operation where the transposed matrix $\mathbf{S}$ is multiplied by the neural time series $\mathbf{r}$ yields an inner product between the stimulus and neural time series at each time lag τ, indicating the similarity between stimulus and neural data. The autocorrelation of the stimulus is accounted for by the inverse

of the autocovariance (note: the inverse operation can be read as a division for square matrices). The resulting weight matrix **w** is of dimensionality N^{ft} x τ x C and describes the set of weights that optimally predict the neural time-series at channel C.

Generally, a time lag of [-100 500]ms is used, unless stated otherwise. Furthermore, if cross-validation is applied, the lambda parameter search is conducted on a linearly spaced set of values ranging from 10e-4 to 10e4 in steps of 10.

Now that the mathematical operations underlying the analysis of continuous data have been defined, the data set described, and the derivation of features explained, we will detail the analyses that compare the different features.

7.2.4 *Analyses*

The data analysis was performed using MATLAB (version R2021a, MathWorks, Natick, MA) with the aid of custom scripts. Part of the analyses were inspired by the work of Desai et al. (2021). For all of the following analyses, we split the data into 6 segments (M = 3.14 minutes, SD = 0.18), where each consists of the neural data and the concurrent soundscape. At least one segment always served as a held-out test set for which the performance metric was computed. The remaining segments were used for deriving model weights and cross-validation parameter estimation. The neural data was split accordingly and z-scored to ensure comparability between segments.

The model testing is done by predicting the neural response based on the held-out test set. The performance metric used to assess each feature was the correlation between the predicted and actual neural time series, providing a measure of the model's ability to capture the underlying neural activity. We opted to contrast the different correlational distributions using the Wilcoxon sign rank tests at α = 0.05, since the assumption of normality was violated for several prediction distributions. We have refrained from computing the corresponding effect size, as the interpretation of such does not provide meaningful information regarding the magnitude of differences between prediction performance. Multiple testing correction was applied according to Benjamini and Yekutieli (2001) False Discovery Rate (FDR) method. For all analyses, the chance level of the correlation for using acoustic onsets was estimated by creating a random acoustic onset vector. For this, we randomly shuffled the temporal position of the acoustic onsets, while preserving the inter-onset interval of the original data. This was done for each condition and participant 100 times and the resulting 95% confidence interval was determined as noise floor.

7.2.4.1 *Nested Model Design*

The first crucial step in our analyses was to establish whether the selected meta-information provides meaningful information for model estimation beyond the existing acoustic depictions. We opted for a nested design, where different factors of meta-information were iteratively added to model estimation and resulting prediction values compared. In particular, we tested three levels. The acoustic features represented the base level onto which we added sound identity, and at last, integrated cognitive prior information.

The first level involved testing the different acoustic features and contrasting their model prediction accuracy. For this, we concatenated the separate EEG data sets and corresponding feature vectors. These were partitioned into 10 segments through which iterated such that each segment became a test set once. In each iteration, the regularization parameter was determined on the remaining 9 segments, and model weights estimated. In the last step, we predicted the neural data on the held-out test set and averaged the resulting correlation values over channels. As this was done for each iteration, we finally averaged over iterations.

To address the impact of adding sound identity information to model estimation we added the sound identity marker of the three tones to each of the acoustic features separately, i.e. combining the acoustic onsets once with the alarm, the irrelevant, or the beep tone. The same training and testing procedure as described earlier was repeated.

The last step involved taking cognitive priors into consideration. To assess their impact, we split each of the tones from the prior analysis (i.e., alarm, irrelevant, and beep) into two separate feature vectors. One feature vector indicated the occurrence of the tone in the narrow condition and the other feature vector indicated the occurrence of the tone in the wide condition. These vectors were zero-padded such that the length remained consistent with that of the neural data.

The statistical analysis involved contrasting the different levels of information content for each of the acoustic features and paired sound identity markers. The corresponding correlational distributions were compared using the Wilcoxon signed rank test and multiple comparison corrections was applied as outlined above. As an additional analysis, we extracted the model weights corresponding to either condition for each tone and compared these using cluster-based permutation testing as detailed in (Mirkovic et al., 2019) using the fieldtrip toolbox (Oostenveld et al., 2011). This was done to test whether the different cognitive priors from the study would impact model weight estimation.

7.2.4.2 *Variance-partitioning*

To address in how far the different models explained similar neural activity, we used variance partitioning (Crosse et al., 2021; Desai et al., 2021; Heer et al., 2017). This analysis allowed us to investigate the unique variance explained of each feature and to pinpoint the degree to which combining different features led to model improvement.

The underlying concept of variance partitioning is that if models A and B share some degree of similarity, the explained variance of the combined model should be less than the sum of the individual models ($A + B > A \cup B$). In other words, the features are not independent. In our study, the correlational values are transformed into the variance explained R^2.

When determining the variance explained from each single and combined model, the unique and mutual explained variance can be computed as $A \cap B = A + B - A \cup B$. This can be extended to three variables as well, where:

$$A \cap B \cap C = A \cup B \cup C + A + B + C - A \cup B - B \cup C - A \cup C$$

In the case of the current study, an example would be to derive a set of model weights based on the envelope, onset, and mel-spectrogram information. For this particular example, we could test how overall model performance improves compared to single feature models i. e., $A_{ons} \cup B_{env} \cup C_{mel}$ vs. $[A_{ons}, B_{env}, C_{mel}]$, but also which feature contributed the most to the improvement and which features are redundant.

$$A_{ons}/(B_{env} \cup C_{mel}) = A_{ons} - A_{ons} \cup B_{env} - A_{ons} \cup C_{mel}$$

We segmented the data into six segments, five for training and cross-validation, and one held out test set for which we calculated the R^2. This was done for all features and their combinations. For the subsequent variance partitioning analysis, we considered only models with a maximum of three different features.

In our study, we calculated the R^2 value for each feature directly from the average channel correlational values. However, in some cases where we sought to determine the unique contribution of a feature in a multi-feature model, the resulting R^2 value was negative. The negative R^2 values, theoretically impossible, according to set theory, are likely caused by overfitting of noise and by too large predictor sets. To address this issue, we introduced a bias estimator *post hoc* based on Heer et al. (2017) work. The optimal bias estimator was derived based on the underlying constraint that variance partitioning should yield values that are at least zero. Here we expect that model similarity is expressed by the degree of overlap.

7.2.4.3 *Cross Prediction*

The similarity of information content between each model was assessed by the variance partitioning analysis. Our next goal was to evaluate the generalizability of a model to predict the neural response to different features. Concerning the results of the variance partitioning analysis, we expected the shared variance observed to reflect the strength of the correlation in this analysis.

In detail, we derived model weights on the training set containing information of feature A and feature B. Then we used the derived weights (A_w,B_w) to predict neural time series on a test set based on feature B information. An example of such a procedure was to select acoustic onsets and mTRF envelope for model training, and applying both of the derived weights on the mTRF envelope of the test data set to predict the neural time series. We pre-defined pairs of features to be tested, which were derived for each participant and condition.

To ensure robust conclusions regarding cross-prediction scores, each of the segments (N=6) served as a testing set once. Within each fold, we derived two sets of model weights, corresponding to either of the feature pair. These weights were used to estimate the neural time series using both the original feature information and the information from the other feature. This approach yielded four prediction scores (averaged over channels) for each segment: two within-prediction scores and two cross-prediction scores. The resulting cross-validation correlational scores were averaged over folds. The mel-spectrogram was excluded from this analysis, as feature dimensions had to be consistent.

We adopted a comparative approach by analyzing the relationships between within and cross-prediction scores for the different feature pairs. The selected approach eliminated the need for normalizing explained variance, as some features inherently explain more variance than others, thus rendering a comparison based on absolute prediction scores unsuitable.

A high correlation between scores suggests that if the within-feature model weights accurately predict the neural time-series, so do the model weights from another feature. This indicates the generalizability of the model weights to other feature information.

For some feature pairs, weights looked highly similar in their trajectory, but peaks were shifted in time. This led to low cross-prediction scores, despite model similarity. Correcting for that shift improved scores tremendously. Further testing revealed that the weights' difference did not imply separate neural processes, but was caused by the feature's properties. This further warrants caution when interpreting model weights

(Haufe et al., 2014; Kriegeskorte and Douglas, 2019; Popov, Ostarek, and Tenison, 2018).

7.3 RESULTS

The main purpose of this study was to investigate feasibility of different features to explain neural data in the context of a naturalistic soundscape. Specifically, we compared acoustic features with respect to their level of acoustic detail. Furthermore, the additional benefit of including meta-information, specifically sound identity markers and cognitive priors, in the neural model estimation was assessed systematically.

7.3.1 *Nested Model Analysis*

In the first analysis, we investigated the impact of including meta-information in the model weight estimation. Specifically, we wanted to test whether adding information of sound identity markers and cognitive priors would enhance model estimation. Here we combined each of the three different acoustic features with sound identity markers separately. This was done to determine the impact of sound identity information on model estimation.

When combining the acoustic features with any of the three sound identity markers, model performance improved compared to the acoustic features alone. Combining the mTRF envelope with the alarm tone (W = 0.0, Z=-3.92, p= 0.001), with the irrelevant (W= 0.0, Z=-3.92, p= 0.001), and with the beep tone (W= 1, Z=-3.88, p= 0.001) significantly improved model performance. Similarly, for the acoustic onsets, specifying experimental tones in the form of the alarm tone (W = 0.0 Z=-3.92, p= 0.001) (Figure 13), with the irrelevant (W = 0.0 Z=-3.92, p= 0.001) and with the beep tone (W= 1, Z=-3.88, p= 0.001) improved model prediction compared to using only the acoustic onsets. The mel-spectrogram also improved, but by less compared to the other two acoustic features. Here specifying the alarm tone (W=12, Z=-3.47, p= 0.0036), the irrelevant (W = 0, Z=-3.92, p= 0.001) and the beep tone (W = 32, Z=-2.73, p= 0.034) improved model prediction.

Next, we tested whether additional consideration of cognitive priors would yield further model improvement. Contrary to the expectations, including cognitive prior information did not improve any feature combination, but led to decreased performance for the irrelevant and beep tone in combination with mTRF envelope (Irr: W=

198, Z=3.47, p= 0.0036; Beep: W= 184, Z=2.94, p= 0.018), and the acoustic onsets with the irrelevant tone (W= 195, Z=3.36, p= 0.005).

The models that included cognitive prior information outperformed the respective acoustic models in predicting unseen data. The only case where this effect was not observed, was the combination of the mel spectrogram and cognitive prior information of the beep tone (W = 40, Z = 2.43, p = 0.073).

In order to validate the finding that CP did not yield any beneficial information for model estimation, we applied cluster-based permutation testing on the derived model weights of the sound identity markers. The results revealed no significant cluster that compared to the windows selected in Rosenkranz et al. (2023), which found significant condition differences for both the alarm [336-432]ms. and beep [308-404]ms. tone. Since we applied a more conservative measure of permutation testing compared to the original linear mixed-model analysis for the pre-specified windows, we re-ran our analysis using the same statistical testing reported in Rosenkranz et al. (2023). Again, the results did not yield any significant differences between conditions (alarm: ß = 0.1564, SE = 0.1444, t(19) = 1.083,p = 0.29; beep: ß = 0.1759, SE = 0.1839, t(19) = 0.975,p = 0.35). A visualization in (Extended Data Figure 14-1) revealed that the TRFs of the narrow and wide condition did not deviate strongly during the pre-defined period, for neither the alarm nor the beep tone marker, explaining the non-significant findings. The TRF trajectories deviate from those of Rosenkranz et al. (2023), which can be attributed to different pre-processing decisions. Albeit no significant differences were found in the pre-defined windows, significant deviations were detected for other latencies (alarm:[180 310], [360 800]ms, beep: [400 800]ms.). However, when testing whether these differences in TRF weights would impact the prediction accuracy, no significant differences were found, thereby supporting our previous finding of the nested model analysis. Based on this result we do not consider condition differences to be relevant in the interpretation of the results of the subsequent analyses.

7.3.2 *Variance Partitioning*

One shortcoming of the previous analysis was that it did not reveal whether any of the derived models explained similar aspects of the underlying neural processes. Thus, the current analysis aimed to determine the degree of similarity of the different models. First, we computed the model prediction accuracies of the different features and used these values for the subsequent computation of the variance partitioning. The statistical assessment was performed on the model predictions only.

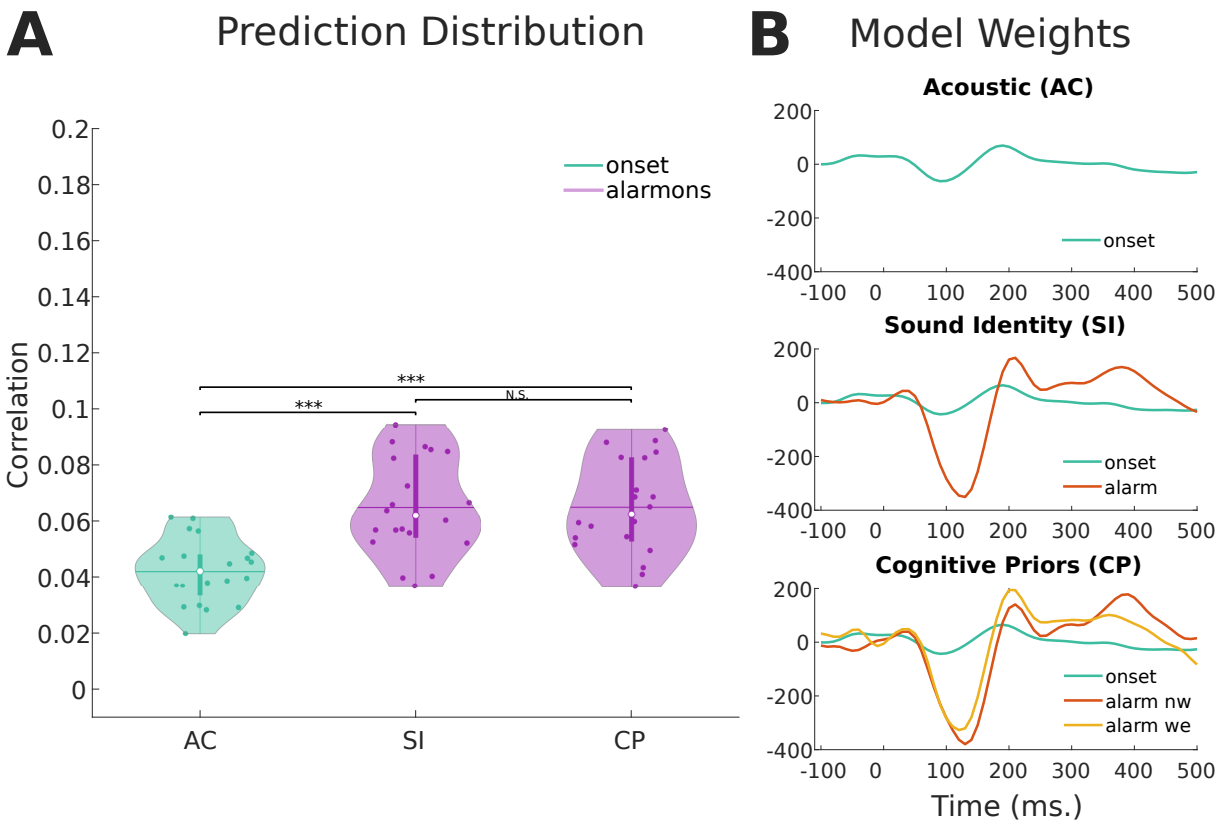

Figure 13: Exemplary model building of combining acoustic onset information with alarm tones and the respective cognitive priors. **A:** Statistical comparison of prediction distributions based on AC = Acoustic, combined with either SI= sound identity, CP = cognitive prior information. * p< 0.05, **p< 0.01 ***p<0.000, N.S.=non-significant **B:** Example of model weights for acoustic onsets alone, extended by SI, or CP information. A comparison of the condition effect to the results of Rosenkranz et al. (2023) can be found in the Extended Data Figure 14-1. The x-axis displays the time lags for the different weights. In the last plot, the nw= narrow and we= wide condition weights are displayed for the alarm tone.

The results of contrasting the acoustic features showed that model prediction differed between them. A direct comparison between models revealed that the mel-spectrogram yielded model weights that were most successful in predicting unseen neural data compared to all other base models (mel | ons: W 210, Z = 3.92, p= 0.002; mel | menv: W = 207, Z = 3.36, p=0.010). The model based on the acoustic onsets performed significantly worse than the other models. Furthermore, the comparison between the two different envelope models showed that the mTRF-implemented function to derive the envelope yielded a better performance (W = 210, Z = 3.92, p=0.002). Each model's performance presented in Figure 15A was above chance level.

Comparing the results of combining the acoustic features pairwise revealed that performance did not differ when acoustic onsets were added to either the mTRF envelope

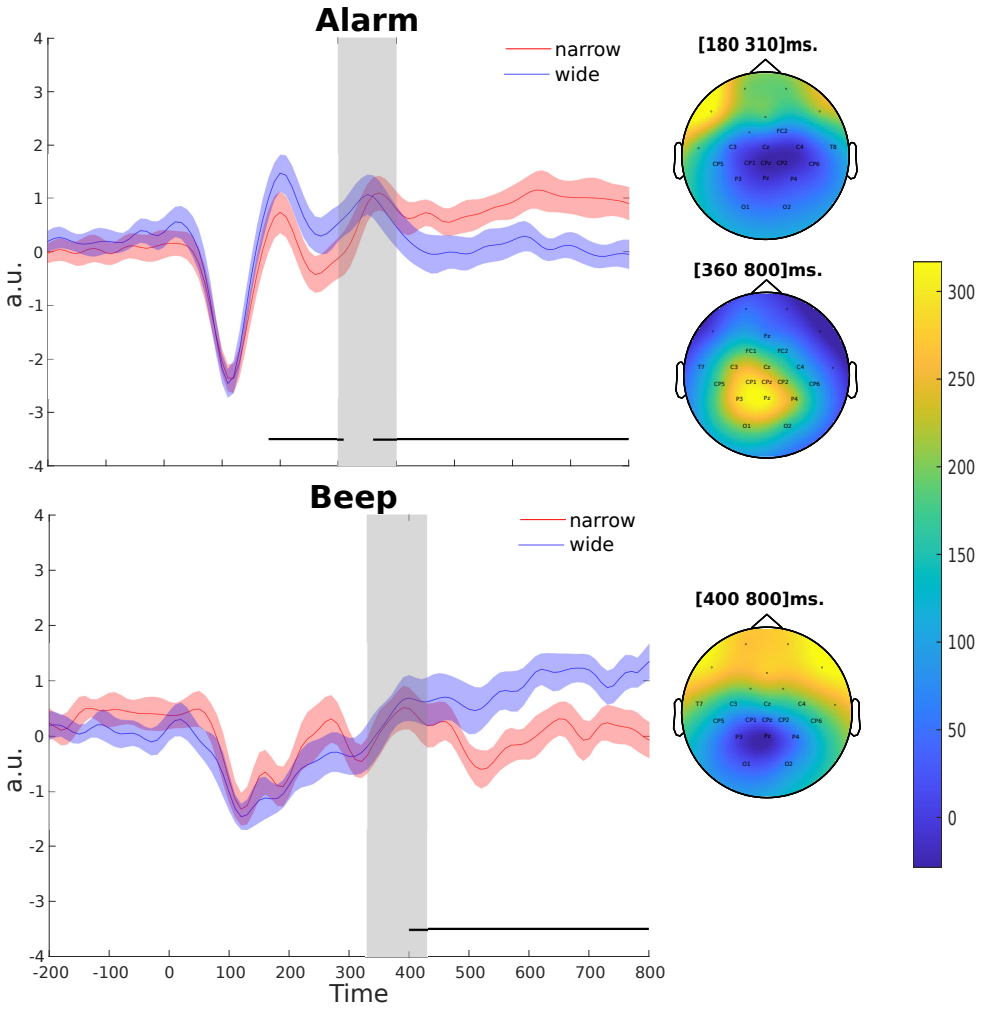

Figure 14: **1**. Upper panel shows the TRF weights of the alarm tone for the two conditions. The grey shaded area marks the window of interest as reported by Rosenkranz et al. (2023). The black lines are the clusters detected by the permutation testing with the corresponding topographies. The lower panel shows the same but for the beep tone.

(W = 85 Z = -0.75, p = 1) or the mel-spectrogram (W = 108, Z = 0.11, p=1) compared to respective, better performing, base model. Combining the mTRF envelope with the mel-spectrogram, however, significantly outperformed the respective base (menv: W = 3, Z = 3.81, p = 0.003; mel: W= 0, Z = 3.92, p = 0.002) and any model combination that included the acoustic onsets. Any acoustic base model that was outperformed by the other acoustic base model was also outperformed by their combined model.

These results were also reflected in the variance partitioning. Nearly all of the variance that was explained by the acoustic onsets was also captured by the mel-spectrogram (r^2_{ons}=0.0023, r^2_{ovlp}= 0.0023) (Figure 15B). Similarly, most of the variance explained by the mTRF envelope was contained in the mel-spectrogram (r^2_{ovlp}= 0.0039).

The mTRF envelope did, however, explain unique aspects of the neural processing (r^2_{unq}= 0.0012).

Similar to the acoustic features each sound identity marker model yielded model weights that predicted neural data above chance level (Figure 15A). Contrasting the different sound identity markers with each other, the alarm tone yielded the highest and the beep tone the lowest performance. While the alarm tone's performance did not differ significantly from that of the irrelevant tone (W = 171, Z = 2.46, p=0.136), both differed significantly from the beep tone's prediction accuracy (alarm: W = 200, Z = 3.55, p = 0.005; irrelevant: W = 191, Z = 3.21, p = 0.015).

Inspecting the results of the pairwise combination of sound identity markers revealed that only the combination of the beep and irrelevant tone did not outperform both base models (Beep: W = 207, Z = 3.81, p = 0.003; Irr: W = 173, Z = 2.539, p = 0.112). When all three sound identity markers were used to derive a combined model all of the pairwise combined models were outperformed significantly (IrrAlarm: W = 185, Z = 2.99, p = 0.031; BeIrr: W =210, Z = 3.92, p = 0.002; BeAlarm: W = 206, Z = 3.77, p = 0.003).

The results of the variance partitioning of the sound identity marker showed that most models explain relatively more unique variance than shared variance. This finding is unsurprising, given that the sound identity markers describe different sections of the soundscape. There were, however, differences in terms of the proportional amount of unique variance explained between the different base models in their respective pairings. In particular, the beep tone model shared the most overlap with any other base model proportional to its total variance explained (r^2 = 0.0009, r^2_{unq} = 0.0007). In contrast, if the irrelevant tone was combined with the alarm tone, most of the explained variance is unique (Alarm: r^2_{unq} = 0.0041; Irr: r^2_{unq} = 0.0027; r^2_{ovlp}= 0.0001)(Figure 15B).

7.3.3 *Proportion Explained*

One aspect that has not been considered in the previous analysis, but is vital if one aims to compare different sets of features, is the consideration of the features' structure. For instance, the acoustic envelope depicts amplitude changes over the entire time course and hence allows to relate the ongoing EEG to the ongoing soundscape at each time point. Whereas discrete features, e.g. sound onsets, only allow to model the neural response around specific time points and do not provide a prediction of the neural signal where no sound onsets occur. Considering the typical lag used in

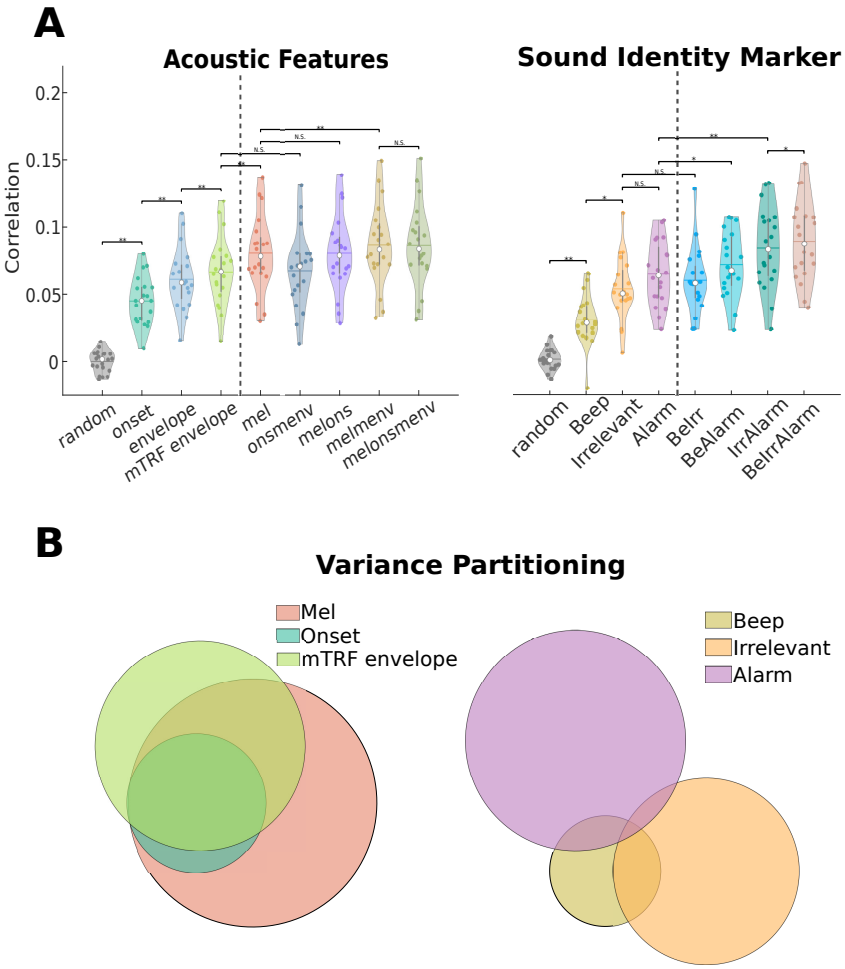

Figure 15: **A:** A comparison of the prediction distributions of the acoustic features and their combinations. * p< 0.05, **p< 0.01 ***p<0.000, N.S.=non-significant. A detailed comparison between the envelope and mTRF envelope can be found in Extended Figure Figure 16-1. **B:** The result of the variance partitioning showing the unique contribution and shared explained data. The letters correspond to Table 1, which displays the values of the variance partitioning analysis.

our model computation, which covers 500 milliseconds post-stimulus onset, viable predictions are only feasible for periods immediately following an auditory event. In contrast, during 'event-free' periods where no auditory onset event is present, the model's prediction of EEG activity is effectively zero, rendering these predictions non-informative. To determine the impact this distinction has on model prediction we re-evaluate the accuracy of discrete features for sections where data can be predicted using the analysis pipeline of the variance partitioning.

Here, we observed a non-linear negative trend in terms of prediction accuracy with an increasing proportion of explainable data (Figure 17). While the sound identity markers contributed only a fraction of the total explainable data (alarm: 3%, irrelevant:

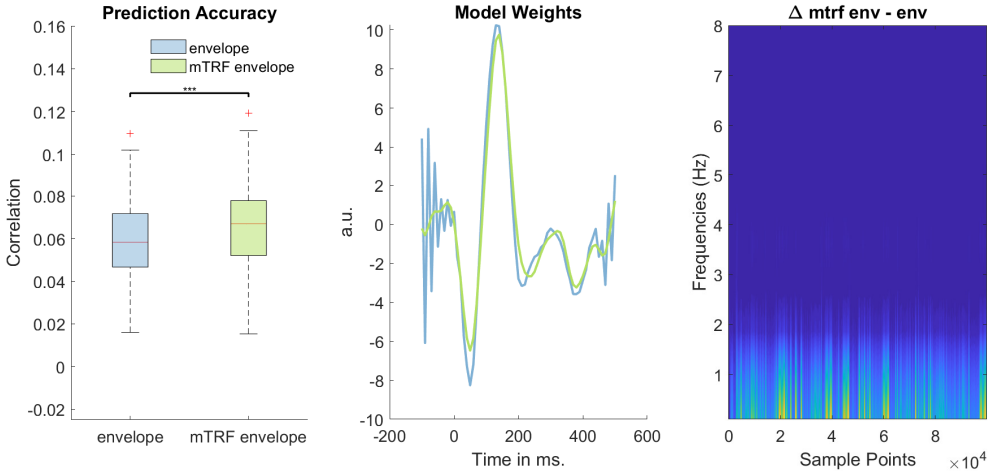

Figure 16: The left panel shows the distribution of the prediction accuracy for the envelope and mTRF envelope model. Significance is tested at * p< 0.05, **p< 0.01, ***p< 0.000. The panel in the middle shows the model weights averaged over participants, conditions, and channels. The panel on the right highlights the frequency decomposition of the difference curve of the mTRF envelope and the envelope.

3%, beep: 4%), they showed high prediction accuracies for these small sections. We again observed that the alarm tone yielded the best model predictions (M = 0.34, SD = 0.11), followed by the irrelevant (M = 0.28, SD = 0.11), and then the beep tone (M = 0.15, SD = 0.06). Interestingly, combining these three markers into one model yielded overall more data that can be explained (9%), but with reduced accuracy compared to the separate models (M = 0.24, SD = 0.08). The prediction distribution of the combined model appeared to be a linear combination of the results of the three separate models.

Using the acoustic onsets we could make predictions for roughly 26% of the data using the described TRF settings. For the mTRF envelope and the mel-spectrogram on the other hand, we were able to predict the neural data at every sample. Contrasting the prediction distributions by accounting for the difference in samples where predictions were made, we found that the difference between the acoustic onsets and the mTRF envelope was no longer significant (W = 144, Z = 1.46, p = 0.668). Furthermore, the direct comparison between the acoustic onsets and the mel-spectrogram revealed they were also no longer statistically different (W = 155, Z = 1.867, p=0.290). The contrast between the mel-spectogram and mTRF envelope remained significant (W 0 195, Z = 3.36, p=0.004).

The prediction accuracy as well as the proportion of explainable data were improved if the sound identity markers were added to the acoustic onsets (29%, M = 0.11, SD = 0.04) (Figure 17). The increased proportion of data that could be explained indicates that our detection of onsets can be further improved, as seemingly not all

		r^2	r^2_{unq}	A∩B	A∩C	B∩C	A∩B∩C
SI	Beep (A)	0.0009	0.0007	0.0003	0.0002	0.0001	0.0003
	Irr (B)	0.0028	0.0027				
	Alarm (C)	0.0041	0.0041				
AC	Mel (A)	0.0072	0.0031	0.0023	0.0039	0.0020	0.0020
	Ons (B)	0.0023	0				
	mEnv (C)	0.0052	0.0012				

Table 1: Shows the variance explained for the acoustic (AC) and sound identity (SI) features. Specifically, the first two columns display the total and unique variance explained by each feature. The last four columns highlight the shared variance (∩), where A refers to the first, B to the second, and C to the third feature, for SI and AC respectively (Figure 15).

SI markers were detected. We believe that the improved prediction accuracy for the acoustic onset and SI markers was not only due to the increased data quantity that could be explained. Rather, we found that the additional information of the SI markers also improved the data prediction qualitatively, as highlighted by finding increased performance of combining SI markers with the mTRF envelope (M = 0.1, SD = 0.03) or mel-spectrogram (M = 0.9, SD = 0.03). Since the mTRF envelope and mel-spectrogram made predictions at every sample, the improved prediction accuracy can be attributed to the additional information of the SI markers. Specifically, the combination of SI markers with the mTRF envelope (W = 210, Z = 3.92, p = 0.001) or mel-spectrogram (W = 207, Z = 3.808, p = 0.001) outperformed either base model.

The correlational values found for the sound identity markers for the explainable data periods exceed those commonly found in other speech tracking studies (Drennan and Lalor, 2019; Gillis et al., 2021; Mesik and Wojtczak, 2023). Here, generally, values ranging around 0.02-0.05 are detected. To investigate if the values we observed are plausible, we created 10 minutes of data where we randomly placed 396 triphasic ERP responses (P1-N1-P2). On top of this signal, we added pink noise, where the noise distribution followed the $1/f$ distribution. The noise was added at various amplitudes, achieving SNRs that varied between –30 dB to 10 dB. We subjected this simulated data to the same analysis pipeline as for the original data, and computed correlation coefficients for the explainable segments as well as for the whole segment (Figure 18).

Another crucial aspect was determining the SNR level of our sound identity markers and acoustic onsets model at a single trial level. For this, we computed the standard deviation of a baseline period in each trial [-0.1 -0.01]s. and took the mean absolute

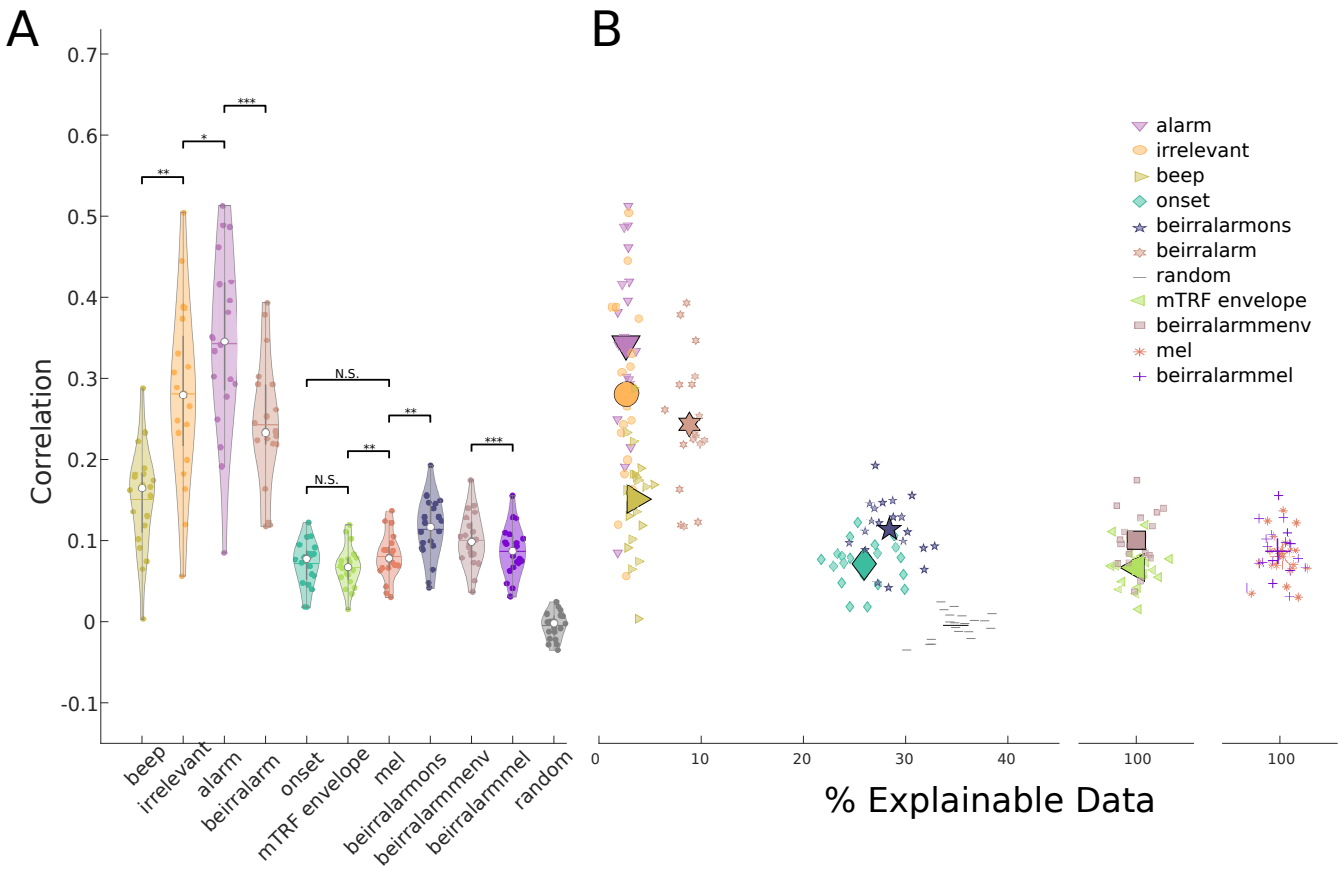

Figure 17: **A:** Displayed are the distributions of prediction accuracies for a subset of different features and feature combinations. Each dot represents the average condition and channel correlation score of one participant. These distributions are the collapsed results over proportion of explainable data as presented in **B**. The results of the Wilcoxon sign rank test are presented exemplary at * p<0.05, **p<0.01 ***p<0.000, N.S.=non-significant. **B:** prediction distributions share the same y-axis as that of part A. Additionally, we computed the proportion of data that is explainable using the different features. For the continuous features such as the envelope or mel-spectrogram, the variance was added for visual purposes.

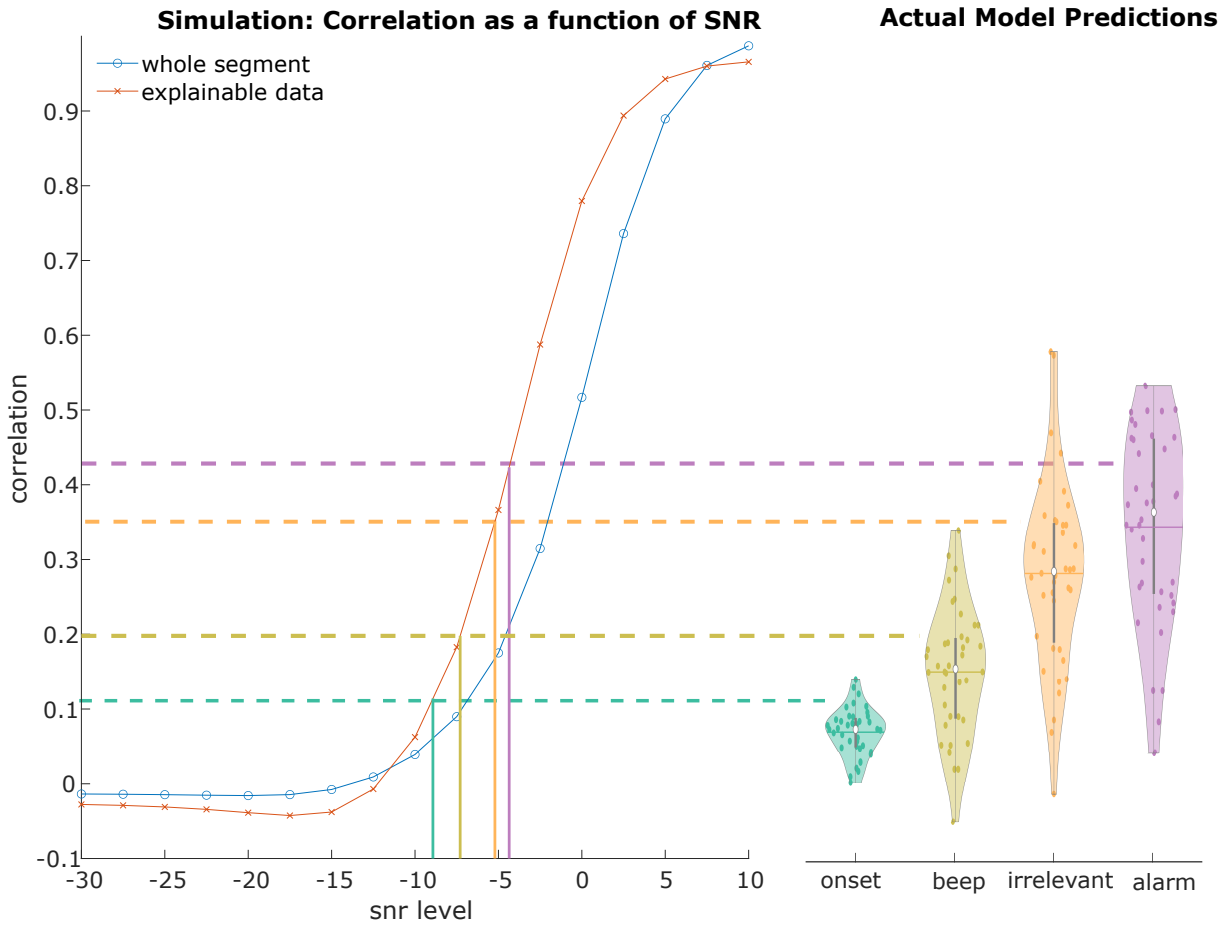

Figure 18: The left panel shows the results of the simulation study. The x-axis depicts the different simulated SNR levels in dB. The y-axis shows the correlational values. The two graphs show the mean correlational values as a function of SNR levels for either the whole segment (blue) or only where data was predicted (red). The vertical lines show the estimated SNR level for the onset, beep, irrelevant, and alarm tone respectively. The panel on the right shows the prediction scores of the actual data.

value of the corresponding ERP from [0 0.5]s. The ratio of these two values was converted to dB and used as a single trial estimate of SNR. Vertical lines represent the SNR values of the different features, and the plot on the right reflects the actual data predictions.

As can be seen, the expected correlational values of the simulation are similar to the ones obtained for the actual data. The results for the simulation are consistently higher than the mean correlation values of the real data. The higher values observed in the simulation study may be due to the assumption of identical neural responses for each trial, which is not realistic. Despite this limitation, we believe that our simulation still demonstrates that these high correlation values are possible.

7.3.4 *Cross Prediction*

The results of the cross-prediction indicated in how far a model trained on one feature would be able to predict data based on other feature information. A positive correlation of prediction scores indicates that segments that are predicted well by the matching feature model, are also predicted well by the other feature model.

Inspecting the corresponding cross-prediction for the mTRF envelope and onset, we observed a highly linear relationship. Using mTRF envelope information, we show a high correlation between prediction values using the mTRF envelope model and acoustic onset model. ($\rho = 0.76$, $p < 0.001$). Conversely, using acoustic onset information, we show a high correlation using the acoustic onset model and mTRF envelope. ($\rho = 0.79$, $p < 0.001$) (Table 2).

All model cross-predictions showed a significant linear relationship. Here, the beep tone performed better at predicting the alarm ($\rho = 0.63$, $p < 0.001$) or irrelevant tone ($\rho = 0.72$, $p < 0.001$), than the other way around (alarm on beep ($\rho = 0.51$, $p < 0.001$) irrelevant on beep ($\rho = 0.63$, $p < 0.001$). The highest correlation was found when using either the alarm or the irrelevant model to predict the neural data based on either alarm or irrelevant feature information. Despite the positive correlations, segments were always predicted better when the neural model fit the features used, compared to the cross-prediction (Table 2 and Extended Data Table Figure 19-1, Figure 20-2).

We also tested how models based on general acoustic properties generalize to models based on sound identity markers and vice versa. We observed the same positive relationship of feature models generalizing to other features. Importantly, at face values these correlational scores overall are lower compared to those of AC predicting other AC or SI other SI features (Table 2).

		base				
		AC		SI		
		onset	mTRF envelope	alarm	irrelevant	beep
Cross	onset	1	0.76	0.72	0.81	0.49
	mTRF envelope	0.79	1	0.70	0.69	0.38
	alarm	0.69	0.28	1	0.85	0.51
	irrelevant	0.72	0.44	0.81	1	0.63
	beep	0.48	0.18	0.63	0.72	1

Table 2: Table shows the correlational values for the different cross-prediction pairs. The table should be read from left to top. The feature labels on the left are the models that were trained on these features. The labels at the top are feature information used to predict segments using the model on the left. A visualization of the results can be found in Extended Data Table Figure 19-1 and Figure 20-2. All correlations are significant at $p<0.01$.

7.4 DISCUSSION

Understanding the neural response to natural soundscapes necessitates comprehensive information about the auditory environment. The abstraction of this information into features represents a fundamental parameter that influences the estimation of neural response models. Beyond depicting acoustic aspects the soundscape can also be described in terms of how the perceiver interacts with it. For this, we derived sound identity markers and cognitive priors. We set out to test in how far the features' comprehensiveness in depicting the soundscape impacts the amount of neural variability that can be explained.

Our findings show that as we provide both, sound identity information and a more acoustically detailed description of the auditory signal, the accuracy of model prediction improves. Simultaneously, deriving the most parsimonious abstraction of the complex soundscape, that is the acoustic onset of sounds, can be used to explain significant portions of neural variability. These findings provide crucial insights to understand and determine what aspects of the naturalistic auditory soundscape are important to capture to investigate the underlying neural activity.

7.4.1 *Features Comprehensiveness's Impact on Explaining Neural Variability*

For the acoustic aspects, we abstracted the complex soundscape in varying degrees of acoustic detail to determine the impact on explaining neural variability. Features

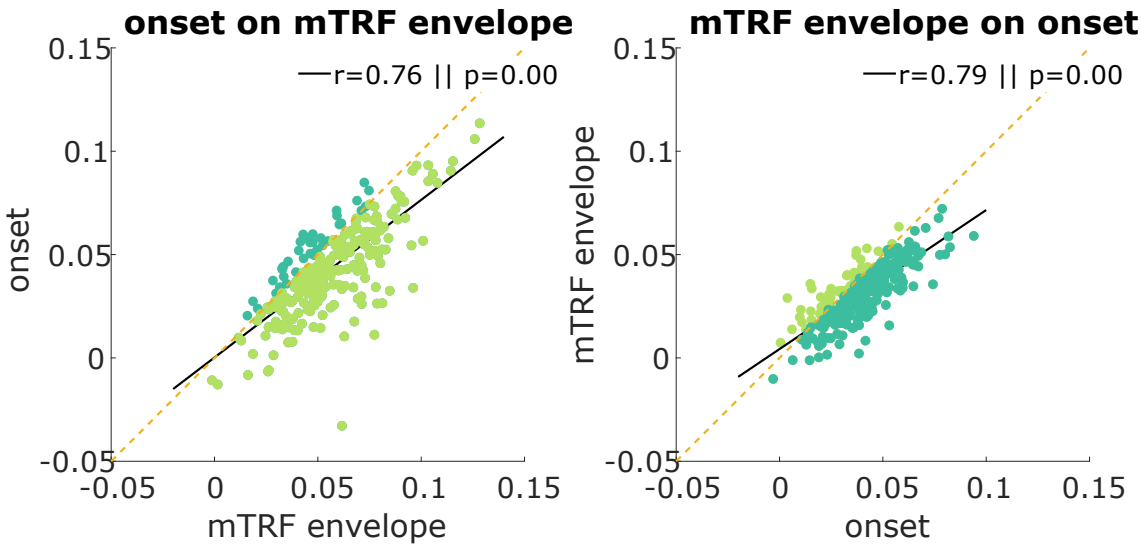

Figure 19: This figure shows the results of the cross-prediction analysis. On the x-axis are the correlational score of the testing data segment with the prediction based on feature information that the model was initially trained on. On the y-axis are the correlational scores for the same segment and feature information as on the x-axis, but using model weights derived from the depicted feature.

ranged from the most fundamental aspects, the onset of sounds, to the envelope, and the highly detailed mel-spectrogram. Replicating previous results, we find that acoustic features explain significant portions of neural data, where the explained neural variability increases with the acoustic detail of the features (Desai et al., 2021; Drennan and Lalor, 2019). Here the mel-spectrogram explains most neural variability, replicating the results of previous studies (Daube, Ince, and Gross, 2019; Di Liberto, O'Sullivan, and Lalor, 2015). This can be explained by the feature's characteristic resembling human auditory processing, the decomposition of the signal into non-linear spaced frequency bands, which captures the non-linear neural response more accurately (Brodbeck et al., 2023; Rahman et al., 2020). Interestingly, the most parsimonious representation, the acoustic onsets, explains a significant portion of the data, that depending on the analysis, is on par with the continuous envelope.

Close examination of combining acoustic features—onsets, with either the envelope or mel-spectrum—reveals no enhancement of prediction accuracy, suggesting they explain the similar aspects of the neural signal. This finding aligns with our variance partitioning results, indicating shared explained neural variance among these features. The information overlap is attributed to a common basis: onsets are derived from the envelope, which is in some sense a broadband representation of the mel-spectrogram.

Furthermore, the generalization between acoustic onsets and the mTRF envelope underscores their high comparability. From this, we contend that the predominant information content in the mTRF envelope relevant to deriving a neural response model is largely associated with acoustic transients. This assertion aligns with established findings (Deoisres et al., 2023; Oganian et al., 2023; Zuk, Teoh, and Lalor, 2020), supporting the notion that acoustic transients are a key determinant in driving the neural response. Notably, our findings gain further significance when considering that, upon excluding samples with no predictions, the mTRF envelope and acoustic onsets are not significantly different from each other. One could argue against this line of reasoning by pointing out that peak latency differences between the acoustic onset and mTRF envelope model had to be corrected for the cross-prediction analysis. Subsequently, these peak differences could suggest separate neural processing. Here we would like to refer to the work of (Lalor et al., 2009) who contrasted the response functions to discrete unit impulses and continuous stimulus characteristics. Not only did they observe a similar shift of peak latencies for discrete and continuous stimuli, but also their source localization revealed the continuous model to be a generalized version of the discrete response. The shift of latencies intuitively makes sense given that the envelope reaches maximum energy later compared to the discrete onsets. Here, the neural response is closer to the later occurring peak of the envelope, resulting in earlier peak latencies for the continuous model (Brodbeck, Hong, and Simon, 2018). The finding that acoustic onsets capture crucial aspects of a complex soundscape is particularly relevant as acoustic onsets represent only a fraction of the soundscape, making them a suitable feature for BTL recordings (Hölle and Bleichner, 2023).

Besides the abstraction into acoustic features, we also depicted the soundscape in terms of how the user interacts with it, using sound identity markers and cognitive priors. Regarding sound identity markers, there are considerable differences in their predictive power. Specifically, the irrelevant and alarm sounds show higher prediction accuracy compared to the beep tone. This is likely due to their heightened salience (Huang and Elhilali, 2017), evoking stronger neural responses. Despite the sound identity models' uniquely explained data, the cross-prediction suggests model weights to be similar. We argue that weights fitted on specific sounds share a common neural basis and that the observed variations are partly due to differences in the acoustic profiles of the sounds.

It is important to note that there is not a single "best" feature; several distinct feature sets could all yield similar prediction accuracy. Therefore, it is crucial to keep the research question in mind and recognize that a feature set explaining a large proportion of the variance might provide a plausible explanation, but it is only one of many

possible explanations for the observed data (Diedrichsen, 2020). This complexity underscores the need to consider multiple feature combinations and remain cautious about over-interpreting the significance of any single feature set. The results in this study indicate a general pattern, where the set of features that explain the most neural variability is a combination of detailed acoustic features and sound identity markers. The overall improvement of combining acoustic features with sound identity markers can be explained by sound identity markers accounting for sound-specific variance and acoustic models representing an average neural response to acoustic aspects. For instance, we argue that the acoustic onsets represent a suboptimal one-fits-all solution, as they encompass a broad range of different sounds, thus yielding weights representing a smeared average. Conversely, sound identity models, derived from recurring sounds, offer a more accurate representation of the underlying processing. The similarity between these models shown by the cross-predictions supports the notion that general acoustic processing properties are represented in both types of models. Somewhat related findings come from speech analysis, where Di Liberto, O'Sullivan, and Lalor (2015) found that the inclusion of sound identity markers, in this case, phonemes, improved prediction accuracies compared to a model where only the spectrogram was used. Whether the improved prediction is solely due to a more refined estimation of the acoustic processing to the specific sounds (Daube, Ince, and Gross, 2019), or shows higher-order processing cannot be determined from these results alone and requires further investigation.

Generally, our results suggest that general acoustic models are suitable to explain neural variability, but can be improved by accounting for sound-specific variance. However, these results should be interpreted cautiously. The sounds in our study that have identity markers are simple beep-like and presented identically each time they occur. This contrasts with the rest of the signal, where the acoustics are more varied. Despite this variability, we can reasonably hypothesize that providing more detailed information about sound identity can significantly improve prediction accuracy. For studies, where this information is not readily available (Hölle et al., 2022), this proves to be a vital finding, as efforts should be directed toward obtaining better descriptors of the acoustic environment when recording beyond the lab. One way to obtain a better description of the acoustic environments where sound identity information is not readily available, novel online, deep neural nets such as the Yamnet could be applied. The continuous classification of sound categories could aid in comprehensively describing the soundscape and thus improve neural model estimation.

7.4.2 *Comparing Discrete to Continuous Features*

Beyond exploring acoustic properties, sound identity, and cognitive priors, our study extends to understanding how the temporal distribution of sounds and their representation as features affect model weight estimation in EEG analysis.

An important aspect to consider when comparing continuous and discrete features is that the latter only explains specific sections of the data. To ensure an accurate assessment of model performance, we chose to concentrate on EEG data segments following auditory events and exclude 'event-free' periods from the correlation computation. Here the derived feature does not contain a depiction of the soundscape, thus no mapping onto the neural data can occur. The prediction is essentially zero. Including periods where no event was detected will impact the computation of the correlation. For the sound identity markers, this is most vivid as they significantly outperform acoustic features in terms of prediction accuracy. It has to be noted, however, that they apply to only a limited portion of the overall data and are repeated identically in the soundscape. Whether the observed effect holds for soundscapes without repeating identical tones has to be shown. Yet, the relative decrease of prediction accuracy for discrete features is seldom recognized in studies that contrast discrete with continuous features (Mesik and Wojtczak, 2023). Here researchers should decide whether a comparison is suitable over samples that can be predicted, or whether the inability of discrete features to predict every sample should be factored into the comparison to continuous features.

It has to be noted that our approach to derive acoustic onsets uses a somewhat arbitrary threshold of what is considered an onset and what is ignored. This approach originated from our previous study (Hölle et al., 2022). This has several implications for the analyses. Specifically, the selection of the threshold alters the type of onsets detected and thus inevitably the derived model and hence the prediction accuracy. Although the choice of the threshold is arbitrary to some degree a too-low threshold will result in many onsets being detected, ranging from clearly audible to minor, barely perceptible sound gains. These might not be meaningful when setting the brain in relation to the acoustic description. Concurrently, a too-high threshold will limit the analysis to only very few distinct sounds, failing to describe large portions of the soundscape. To strike the balance of this selection is difficult and will depend ultimately on the choice of the researcher. Despite the arbitrary nature of threshold selection, we have shown that the most fundamental depiction of a soundscape: the onset of sounds is sufficient to explain neural variability.

This study contributes to the field by extending findings from speech studies to complex soundscapes, such as the operating room environment we used here. We demonstrate that approaches commonly used for studying speech tracking—such as breaking down the acoustic signal into different feature sets (e.g., phonemes, word surprisal)—can also be applied to natural soundscapes. Similar to speech tracking, different feature sets can capture different aspects of the soundscape.

It is important to note that the nature of the soundscape (e.g., man-made vs. natural sounds) and our relationship to these sounds can influence which feature sets are most informative. For instance, a few studies have looked at processing differences of speech and music using EEG (Shan, Cappelloni, and Maddox, 2024; Zuk, Teoh, and Lalor, 2020). Future studies could therefore explore the interaction between speech and non-speech aspects of the soundscape to further understand these dynamics.

7.4.3 *Conclusion*

In this study, we have shown that estimating the neural response to a naturalistic soundscape is possible using a combination of acoustic features, sound-identity information, and cognitive priors. While parsimonious acoustic onsets suffice for robust neural modeling, a detailed and specific description of the soundscape generally enhances model estimation. However, this specific information is not consistently available when relating the neural signal to an everyday soundscape. For instance, neither the cognitive state of the perceiver nor the exact sound labels may be known when monitoring everyday life sound perception.

This variability in feature availability underscores the nuanced nature of EEG data analysis in natural soundscapes. Our findings do not pinpoint a single 'optimal' feature set but rather highlight several advantages and limitations to consider. By addressing these challenges, we aim to provide a more holistic understanding and set the groundwork for future research in EEG analysis of complex auditory environments.

Funding Information

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Emmy-Noether program - BL 1591/1-1 - Project ID 411333557.

Acknowledgements

We would like to thank Manuela Jäger and Silvia Korte for the fruitful discussions throughout the development of the study. We also thank Sebastian Puschmann and Jörn Anemüller for their insightful comments.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Potential Declarations

During the preparation of this work, the author(s) used ChatGPT 4o and the free version of ChatGPT (mid 2024) in order to improve language and readability of selected sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author Contribution

TH, MR, and MB Designed Research, TH Performed Research, TH Analyzed Data, TH and MB Wrote the Paper

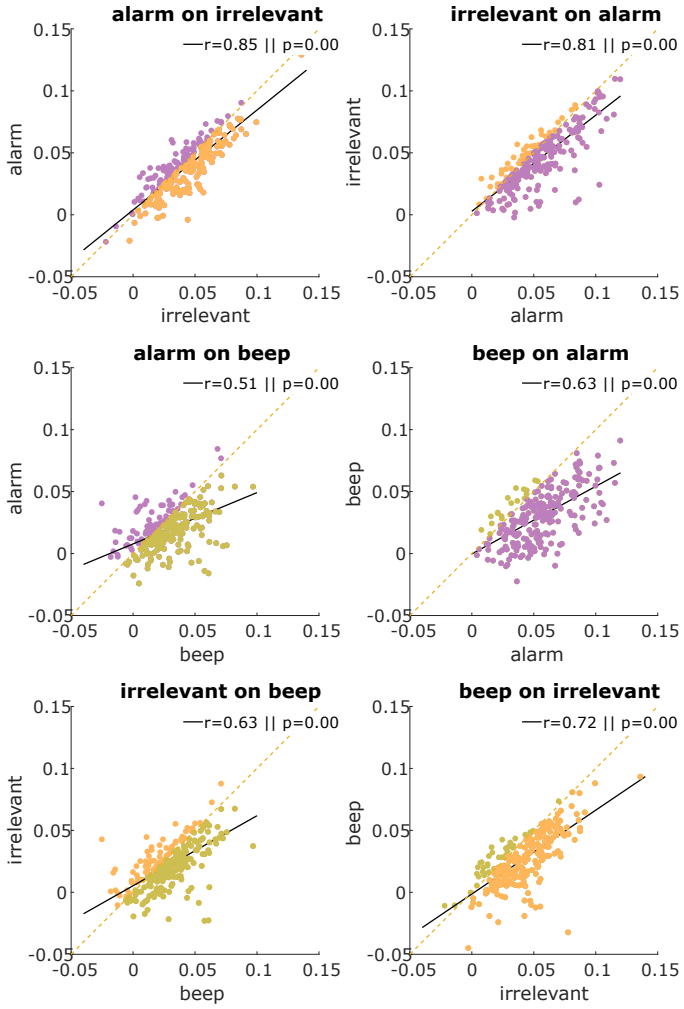

Figure 20: This figure shows the results of the cross-prediction analysis for the sound identity marker. On the x-axis are the correlational scores of the testing data segment with the prediction based on feature information that the model was initially trained on. On the y-axis are the correlational scores for the same segment and feature information as on the x-axis, but using model weights derived from the depicted feature.

8

NEURAL RESPONSE ATTENUATION

"Science may be described as the art of systematic oversimplification."

Popper and Poper (1991)

Neural response attenuates with decreasing inter-onset intervals between sounds in a natural soundscape.

Thorge Haupt[1], Marc RosenKranz[1], Martin G. Bleichner[1,2]

[1]Neurophysiology of Everyday Life Group, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[2]Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

This chapter is identical in content to the version published in:
Thorge Haupt, Marc Rosenkranz, and Martin G. Bleichner (Sept. 2025b). "Neural response attenuates with decreasing inter-onset intervals between sounds in a natural soundscape." en. In: *eNeuro*. Publisher: Society for Neuroscience Section: Research Article: New Research. DOI: 10.1523/ENEURO.0210-25.2025

Abstract

Sensory attenuation of auditory evoked potentials (AEPs), particularly N1 and P2 components, has been widely demonstrated in response to simple, repetitive stimulus sequences of isolated synthetic sounds. It remains unclear, however, whether these effects generalize to complex soundscapes where temporal and acoustic features vary more broadly and dynamically. In this study, we investigated whether the IOI, the time between successive sound events, modulates AEP amplitudes in a complex auditory scene. We derived acoustic onsets from a naturalistic soundscape and applied temporal response function (TRF) analysis to EEG data recorded from normal hearing human listeners (N = 22, 16 females, 6 males). Our results showed that shorter IOIs are associated with attenuated N1 and P2 amplitudes, replicating classical adaptation effects in a naturalistic soundscape. These effects remained stable when controlling for other acoustic features such as intensity and envelope sharpness and across different TRF model specifications. Integrating IOI information into predictive modelling revealed that neural dynamics were captured more effectively than simpler onset models when training data were matched. These findings highlight the brain's sensitivity to temporal structure even in highly variable auditory environments, and show that classical lab findings generalize to naturalistic soundscapes. Our results underscore the need to include temporal features alongside acoustic ones in models of real-world auditory processing.

Significance Statement

Employing automatic onset detection in a complex, ecologically valid soundscape, we enable fine-grained analysis of temporal auditory processing. Specifically, we find that neural responses (i.e. the N1 and P2 components) to sound events are attenuated when inter-onset intervals are short, replicating classic attenuation effects within a naturalistic soundscape. These findings demonstrate that temporal sensitivity in auditory processing persists even in the presence of substantial acoustic variability, which is characteristic of real-world settings.

8.1 INTRODUCTION

Non-invasive neuroimaging tools, such as electroencephalography (EEG), have been invaluable in unraveling the neural underpinnings of auditory perception (Alain and Winkler, 2012; Gutschalk and Dykstra, 2014; Lee et al., 2014; Rahman et al., 2020). Many of the mechanisms uncovered using EEG have relied on highly controlled, low-complexity stimuli (Crosse et al., 2016; Schutz and Gillard, 2020). Often, unnatural and repetitive, click-like tones have been used, reducing experimental investigation to changes along a single stimulus dimension, such as intensity (López-Caballero et al., 2023), frequency (Herrmann, Schlichting, and Obleser, 2014; Herrmann et al., 2013), or inter-stimulus interval (ISI) (Zacharias, König, and Heil, 2012). An emerging question is whether established neural mechanisms, such as the attenuation of auditory evoked potentials (AEP), can be applied to understand human perception in response to complex and naturalistic soundscapes, where many of the investigated factors occur simultaneously.

A well-documented finding is that the amplitude and latency of an AEP in response to a sound are dependent on the characteristics of the preceding sound, as well as the context. Studies have shown that repeated presentations of tones modulate AEP components associated with acoustic processing, particularly the N1. It has been shown that the N1 amplitude is reduced by a preceding tone for up to 10 seconds (López-Caballero et al., 2023; Wang et al., 2008a). Furthermore, it has been shown that the N1 amplitude scales non-linearly with ISI (Zacharias, König, and Heil, 2012). While the peak modulation has been observed reliably, the exact neural mechanisms driving this attenuation remain debated (May and Tiitinen, 2010; Näätänen and Picton, 1987). Taken together, these findings demonstrate that the N1 component is sensitive to specific stimulus properties, such as temporal spacing. Importantly, changes in these acoustic properties lead to predictable patterns of modulation in amplitude and latency of the neural response.

Recent advances in data analysis and experimental designs have made it feasible to study auditory processing in response to more naturalistic sound environments (Brodbeck et al., 2023; Crosse et al., 2021; Holdgraf et al., 2017; Lalor et al., 2009) and situations (Ladouce, Mustile, and Dehais, 2021; Rosenkranz et al., 2023, 2024). The trend towards using more naturalistic stimuli is particularly notable in language research, where studies have gradually moved from presenting isolated words and phonemes (Lutzenberger, Pulvermüller, and Birbaumer, 1994; Näätänen, 2001) to sentences (Desai et al., 2021), continuous speech (Ding and Simon, 2014; Howard and Poeppel, 2010), and, ultimately, naturally recorded speech (Agmon et al., 2023). This

shift towards naturalistic stimuli provides new insights into auditory processing and raises the question of how far results based on experiments using isolated tones generalize to real-world soundscapes (Hamilton et al., 2021; Schutz and Gillard, 2020; Vallet and Van Wassenhove, 2023).

Temporal response functions allow the study of neural responses to continuous acoustic stimuli (Crosse et al., 2021; Holdgraf et al., 2017; Kriegeskorte and Douglas, 2019), allowing us to investigate whether auditory mechanisms derived from isolated tones extend to real-life sounds. The benefit of these models is that they are straightforward to interpret and allow for the comparison of multiple models (Crosse et al., 2016). However, many standard TRF implementations assume, either explicitly or implicitly, that neural responses to repeated instances of a given feature type, such as peaks in the speech envelope, are uniform. This assumption oversimplifies neural dynamics, as accumulating evidence suggests that brain responses to acoustic features are often nonlinear and context-dependent (Buzsáki and Mizuseki, 2014; Herrmann et al., 2016; Stam, 2005; Wang et al., 2008a). This raises a critical question: How can prediction-based models account for the non-linear and temporally dynamic nature of auditory processing?

Drennan and Lalor (2019) approached the modeling of nonlinear neural dynamics by partitioning the acoustic speech envelope into discrete amplitude-based bins. This enabled a more precise characterization of intensity-dependent neural responses to continuous auditory stimuli.

We aim to utilize the approach of Drennan and Lalor (2019) and investigate whether the influence of inter-onset interval (IOI) (i.e., the temporal distance between two onsets) on neural response amplitude, which was previously observed using simple, isolated stimuli, can be extended to complex, naturalistic soundscapes. The investigation of the effect of IOI on neural response amplitude is non-trivial, since sound events rarely occur at a steady rhythm and differ widely in their acoustic properties. Generalizing this relationship to real-world auditory input would provide a framework for understanding brain dynamics in a more ecologically valid setting. Specifically, we test whether the amplitude of the neural response to a sound onset depends on the duration of the interval preceding it, even in the presence of continuous, naturalistic auditory input.

8.2 METHOD

8.2.1 *Data Set*

The current study uses an existing data set by Rosenkranz et al. (2023), where they investigated the effect of attentional modulation on auditory perception during a complex audio-visual motor task. Specifically, the soundscape was created to simulate sounds encountered in an operating room to determine the neural response to different types of relevant and irrelevant sounds depending on the attentional instructions. In this dataset, 22 healthy, right-handed adults (age range: 20–30; 6 males, 16 females) were recruited through an online announcement. All participants provided informed consent and received monetary compensation. The sample size was determined based on previous studies investigating similar neural markers in natural settings (Hölle and Bleichner, 2023; Scanlon et al., 2017). Eligibility criteria included normal or corrected-to-normal vision, self-reported normal hearing, absence of psychological or neurological conditions, right-handedness, and compliance with COVID-19 hygiene regulations in place at the time of data collection. Two participants were excluded from analysis: one due to poor EEG data quality and another for not following task instructions. Therefore, the final analyzed sample comprised 20 participants (14 females, 6 males).

8.2.1.1 *Code Accessibility*

The code described in the paper is freely available online at `https://github.com/ThorgeHaupt/Attenuation.git`. The analyzed dataset can be found under `https://zenodo.org/records/7147701`. A Dell Precision 3650 Tower running Microsoft Windows 10 Education was used.

8.2.2 *Task*

The goal of the original study was to investigate attentional effects in a surgical workplace scenario. For this, participants had to perform a complex visual-motor task that comprised playing 3-dimensional Tetris. In addition to the standard Tetris rules, vocal instructions occasionally told the participants specific locations where to place the blocks. Besides the vocal instructions, participants had to respond to tones. There were two conditions in which the participants had to respond to different tones. Specifically, participants were instructed to respond either to a distinct alarm tone (narrow attentional scope) or a less distinct beep tone (wide attentional scope). Both tones occurred

within each condition, with only the target instructions differing between conditions. They are, however, not relevant for the current analysis.

8.2.3 *Soundscape*

The soundscape was designed to mimic an operating room, specifically geared towards a surgeon's perspective. The acoustic environment consisted of speech sounds and environmental sounds (e.g., clattering of tools, footsteps, conversations). The sounds were either vocal instructions on where to place the next block or conversation snippets from a podcast. In total, each participant had to comply with the vocal instructions 12 times, where they were told to "Place the next stone in the [upper\lower left \right] corner". Importantly, the instructions were played randomly and never consecutively repeated. The conversation snippets were taken from a podcast and also placed randomly, but in semantically coherent order. The content of the conversation snippets was irrelevant to the experiment. In total, 48 snippets were played and lasted roughly $3.5(\pm 1.5)$s. The total soundscape was played for roughly 16 minutes per condition, totaling 32 minutes of recorded data on average.

Besides speech segments, the soundscape also contained hospital sounds of people moving around and air conditioning. Furthermore, there were three different tones inserted: alarm, beep, and irrelevant. The alarm and irrelevant sound were 200ms, and the beep tone was 60ms long. Each tone was played 48 times and was also randomly placed into the soundscape. Importantly, the experimentally relevant alarm tone was always played from the same direction, whereas the beep was played from multiple directions. Both tones (alarm and beep) were always presented in both experimental conditions, differing only in which tone participants were instructed to respond to. The timing of tone presentations within the soundscape was randomized individually for each participant. When randomization resulted in the beep tone overlapping with other tones (e.g., beep and alarm, or beep and irrelevant sounds), the overlapping sounds were returned to the stimulus pool and presented again at a new randomized time. This was done to obtain 48 non-overlapping trials to unbias the following EEG analysis of the tones. Consequently, slight variations in total condition duration occurred across participants. However, each participant consistently received the full stimulus set: 48 vocal instructions (2–3 s each), 48 conversation snippets (mean duration 3.5±1.5 s each), and 144 total tone presentations (48 each: alarm [200 ms], beep [60 ms], irrelevant [200 ms]), embedded within continuous environmental background sounds (Figure 21). Including brief silent intervals between stimuli and background

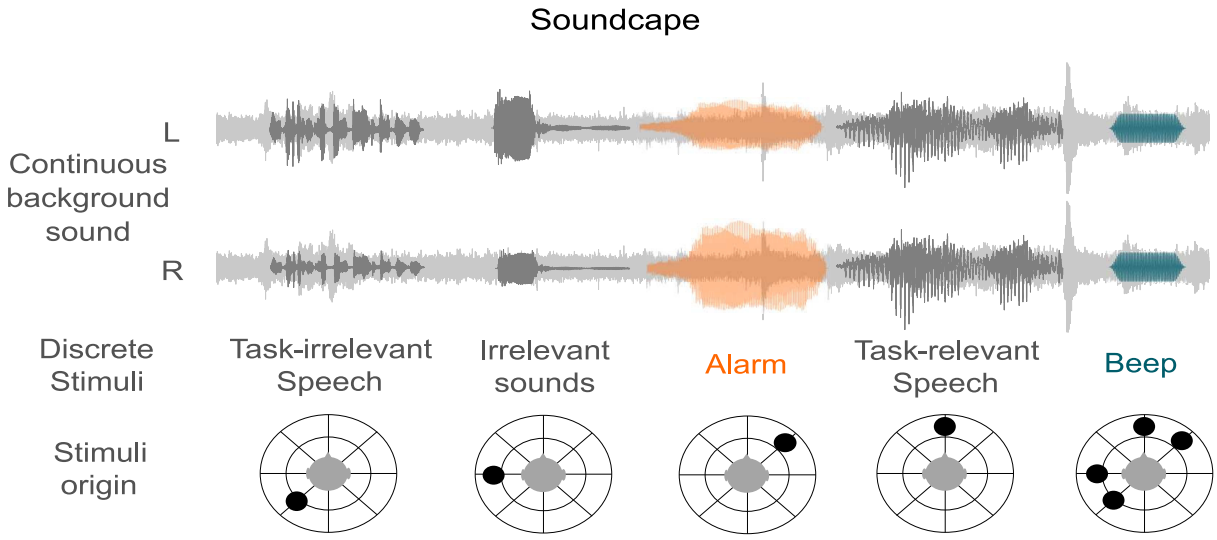

Figure 21: Illustration of the experimental soundscape presented binaurally via headphones (left and right channel shown separately). Light grey indicates continuous surgical background noise. Dark grey marks task-irrelevant sound events, including vocal instructions and irrelevant speech snippets. Orange indicates the alarm tone relevant in the narrow-attention condition, while dark green marks the beep tone relevant in the wide-attention condition. The circular schematics below each discrete stimulus illustrate their spatial positions, manipulated using head-related transfer functions. Adapted from Investigating the attentional focus to workplace-related soundscapes in a complex audio visual motor task using EEG by M. Rosenkranz, T. Cetin, V. N. Uslar, & M. G. Bleichner, 2023, Frontiers in Neuroergonomics, 3, Article 1062227 (https://doi.org/10.3389/fnrgo.2022.1062227). Licensed under CCBY.

environmental sounds, this procedure yielded an average soundscape duration of approximately 18 minutes per condition, totaling around 36 minutes across both conditions. Although durations varied slightly due to randomization, these differences were marginal and not expected to introduce systematic effects on time-on-task analyses. For additional clarity, a schematic timeline illustrating stimulus sequencing and timing has been added (adapted from Rosenkranz et al. (2023)).

All sounds included in the soundscape were processed in MATLAB, such that the RMS was consistent across them. Accounting for differences in loudness was done by adjusting the loudness through sound-specific gain parameters. Lastly, using the head-related impulse function, tones were spatially separated. The experimental audio stimuli were sampled at 44.1 kHz. All recorded data streams were synchronized via the Lab Recorder software, utilizing the Lab Streaming Layer for integration. Participants provided informed consent after being briefed on the procedure. For more details, see the original paper Rosenkranz et al. (2023).

8.2.4 *EEG Measurement*

Participants were fitted with 24 Ag/AgCl passive electrodes positioned according to the 10-20 international system (EasyCap GmbH, Hersching, Germany) for EEG recording. Data collection was performed using a wireless SMARTING system (mBrainTrain, Belgrade, Serbia), with signals referenced to Fz and grounded to AFz. Sampling occurred at a frequency of 500 Hz, and electrode impedance was kept below 20 Ω prior to recording.

8.2.5 *Preprocessing of EEG Data*

EEG preprocessing was conducted using MATLAB (version 2021a, MathWorks, Natick, MA) with the EEGLab plugin and supplementary custom scripts. Artifact detection was performed using ICA. To optimize ICA weight estimation, separate preprocessing steps were employed, as recommended by Winkler et al. (2015). This preprocessing pipeline was solely designed for ICA computation and thus did not influence the data ultimately used for analysis. After deriving the ICA weights, they were applied to the unprocessed raw data.

Initially, data from both experimental conditions were combined for each participant. The combined data was resampled to 250 Hz and subjected to a series of filters, starting with a high-pass filter (cutoff: 1 Hz, order: 568) followed by a low-pass filter (cutoff: 42 Hz, order: 128). These cutoff frequencies were chosen to mitigate drifts and line noise, facilitating optimal ICA weight estimation (Winkler et al., 2015). Channels exhibiting poor signal quality were removed using the clean_channels function. The data was segmented into 1-second epochs, converted to double-precision format, and artifactual trials were removed using the pop_jointprob function with a threshold of three standard deviations.

ICA was executed using the pop_runica function with the extended ICA algorithm. The resulting ICA weights were then reapplied to the raw, unfiltered data from each experimental condition. Automatic classification of ICA components as muscle, eye, heart, line noise, or channel noise artifacts was performed using the pop_icaflag function, with a predefined probability threshold ([0.7, 1; 0.7, 1; 0.6, 1; 0.7, 1; 0.7, 1]). Here the conservative rejection thresholds were chosen to account for the button presses throught the conditions.

Following artifact removal, the raw data underwent a second round of filtering, this time with modified parameters. A low-pass filter was applied first (cutoff: 20

Hz, order: 100), followed by resampling to 100 Hz and high-pass filtering (cutoff: 0.3 Hz, order: 518). The reduced low-pass filter order minimized artifacts associated with steep roll-offs, as recommended by Crosse et al. (2021). The frequency band was restricted to [0.3, 20] Hz, aligning with findings from speech-tracking studies highlighting the dominance of auditory processing in lower frequency ranges (Crosse et al., 2016; Di Liberto, O'Sullivan, and Lalor, 2015). Finally, the data was rereferenced to the mastoids (TP9/TP10).

8.2.6 *Temporal Response Function*

The neural time series were analyzed using the mTRF toolbox (Crosse et al., 2016) in MATLAB. This toolbox estimates weights that relate neural responses to stimulus features through convolution. The neural response, $r(t, c)$, is modeled as the convolution of channel-specific weights (temporal response function), $\omega(\tau, c)$, with the stimulus features shifted by a time lag, τ, plus a residual term, $\varepsilon(t, c)$:

$$r(t, c) = \sum_{\tau} \omega(\tau, c) s(t - \tau) + \varepsilon(t, c). \tag{6}$$

Here, t and c denote time points and channel indices, respectively. This approach captures the delayed nature of neural responses to stimuli. The resulting weights were analyzed for morphology, topography, model performance, multivariate modeling, and cross-prediction.

The TRF is determined by minimizing the Mean Squared Error (MSE) between observed and predicted neural responses:

$$\min_{\hat{r}} \sum_{t} [r(t, c) - \hat{r}(t, c)]^2. \tag{7}$$

The optimal weights, w, are computed using the formula:

$$w = (S^\top S)^{-1} S^\top r. \tag{8}$$

Here, S is the design matrix containing the stimulus features across time lags. Its dimensionality is determined by the number of features and lags. Zero-padding was applied at non-zero lags to maintain causality (Mesgarani et al., 2009). The opera-

tion $S^\top r$ represents the inner product between stimulus and neural time series, while $(S^\top S)^{-1}$ accounts for stimulus autocorrelation.

The dimensions of the resulting model weights are determined by the number of features, the time window of integration, and the number of channels. For instance, dividing the soundscape into 8 different bins of IOI intervals yields a model with the dimensionality of $8 \times 61 \times 22$.

8.2.7 *Model Training*

To train the model, the data was partitioned into 6 segments, where 5 served for training and one segment served as the held-out segment for testing. Within the 5 training segments, cross-validation was applied to derive the optimal lambda for regularization. The resulting model was used to predict the data of the test segment and correlated to the actual neural data. The correlation is the performance marker and is indicative of the prediction accuracy of the model. This approach was consistently applied over all analyses unless stated otherwise. A typical time lag window of [-100, 500] ms was used unless stated otherwise. Cross-validation included a lambda parameter search over values ranging from 10^{-4} to 10^4 in linear steps of 10.

To investigate the role of training data, we increased the available data and contrasted the bin IOS models against the single onset vector. First, we merged the datasets of the two conditions per participant. We then divided the concatenated data into 12 segments, each serving as a test set once.

8.2.8 *Analyses*

8.2.8.1 *Features*

Onsets: We are interested in the effect of the inter-onset interval of two consecutive sounds on the neural response. Unlike previous studies that investigated this effect with pure tones, we are interested here in whether we can replicate the effect in complex soundscapes. For this, we needed to identify sound onsets in the continuous soundscape. To obtain onsets, we used peak detection of acoustic novelty functions (Müller, 2021), which are defined by changes in the energy, spectral flux, and phase changes of the signal, respectively. Given that onset detection was performed on the raw audio signal, no distinction was made between specific sound categories (vocal instructions, conversation snippets, tones, or background noise). Therefore, all detected acoustic onsets were treated equally and weighted identically in the subsequent Tem-

poral Response Function (TRF) analyses. This approach intentionally disregards semantic or categorical aspects of the soundscape to focus purely on acoustic temporal structure. For a detailed discussion of how using purely acoustic rather than content-informed onsets impacts the estimation and interpretation of neural responses, see Haupt, Rosenkranz, and Bleichner (2024).

Energy Novelty: The underlying assumption of the first novelty function is that sound event onset leads to changes in the energy of the signal (x). To obtain this representation, the raw signal was squared, and a continuous measure of local energy was obtained by convolving it with a Hann windowing function. Next, the signal was downsampled to the EEG sampling rate at 100 Hz. Given that sound perception of different intensities is logarithmic in humans, we applied a logarithmic compression $\log(1+\gamma * x)$, where the compression is controlled by $\gamma = 10$. Finally, the rate of change of the signal was obtained by taking the derivative and half-wave rectifying it.

Spectral Novelty: The second novelty function we derived was the spectral flux. Instead of depicting changes in the broadband signal, where overlapping sounds could mask each other, spectral decomposition into different frequency bands can provide a more detailed account of acoustic changes in the signal. First, the signal was decomposed into its frequency components using the short-time Fourier transform (STFT). The magnitude in each frequency band was obtained by taking the absolute value and applying logarithmic compression ($\gamma = 10$). To determine the rate of change in each frequency band, the first derivative was taken, and the signal was half-wave rectified. At last, the signal was obtained by summing over frequency bands. Postprocessing involved removing small fluctuations by subtracting the local average of the signal. Negative values were set to zero.

Complex Novelty: The third novelty function extends the spectral flux function by considering changes in the phase of the signal's frequency components. To avoid chaotic noise-like phase fluctuations impairing the novelty estimate, the phase is weighted by the magnitude of the Fourier coefficient. That is, phase information becomes only relevant, given the magnitude of the Fourier coefficient. The novelty function was derived by determining the difference between the predicted and actual signal, where larger values refer to greater change. Here, the predicted signal was construed based on the assumption of local stationarity, implying that the phase and magnitude of the Fourier coefficients stay relatively constant over some time.

Similar to the spectral flux, the signal was decomposed into Fourier coefficients using the STFT, and besides the magnitude, phase values were extracted. The angle of the coefficients was derived and normalized by 2π. Afterwards, the rate of change was determined by taking the derivative of the phase values. The Fourier coefficient of the

next frame was predicted based on the magnitude, current phase, and rate of phase change. The difference between the actual and predicted coefficient was derived, and novelty values smaller than the previous one were set to 0. The novelty function was obtained by summing over frequencies. Local averaging and half-wave rectification were applied to obtain smoother results.

Onset Detection: Each novelty function represents a distinct measure of change in the signal, contributing unique information. To leverage the complementary representations of change, we normalized the novelty functions between 0 and 1 and averaged them together. This approach aimed to integrate the advantages of all novelty functions, capturing a more comprehensive depiction of changes in the auditory environment.

To detect sound event onsets, we applied an adaptive thresholding algorithm. Unlike global thresholding, which can overlook smaller, noise-like peaks, adaptive thresholding considers the local temporal structure. Specifically, we smoothed the combined novelty function using a Gaussian window ($\sigma = 4$) and applied an offset defined as $mean(x) + 0.05$. To further refine the signal, we employed a median filter with a window size of 1024 samples. The resulting signal represented the local average, and a peak was only selected if the novelty exceeded the local threshold. The temporal location of each detected peak was recorded as a sound event onset. Finally, the onset information was encoded into a binary feature vector for further analysis (Figure 22).

8.2.8.2 *IOI Analysis*

For the IOI, we calculated the time interval between successive onsets. Onsets separated by more than 10 seconds were excluded from further analysis, based on previous research findings suggesting that neural attenuation occurs within this time frame (Zacharias, König, and Heil, 2012).

Inspired by the method of Drennan and Lalor (2019), we applied a similar strategy based on the IOI between sound event onsets. First, we defined ranges for the IOI values and assigned each onset to its corresponding bin. To determine the bin edges, we analyzed the sample distribution of distance values (Figure 23). This distribution of distances of the sound onsets is non-normally distributed and is skewed to the lower distance values. Here, 80% of the sound event onsets follow another sound event within 3.63 seconds. The binning was designed to ensure a uniform distribution of onsets across bins, meaning each bin contained an equal number of onsets. Since no prior studies had applied this approach, we experimented with different bin numbers (ranging from 2 to 8), leading to seven distinct models with varying numbers of bins.

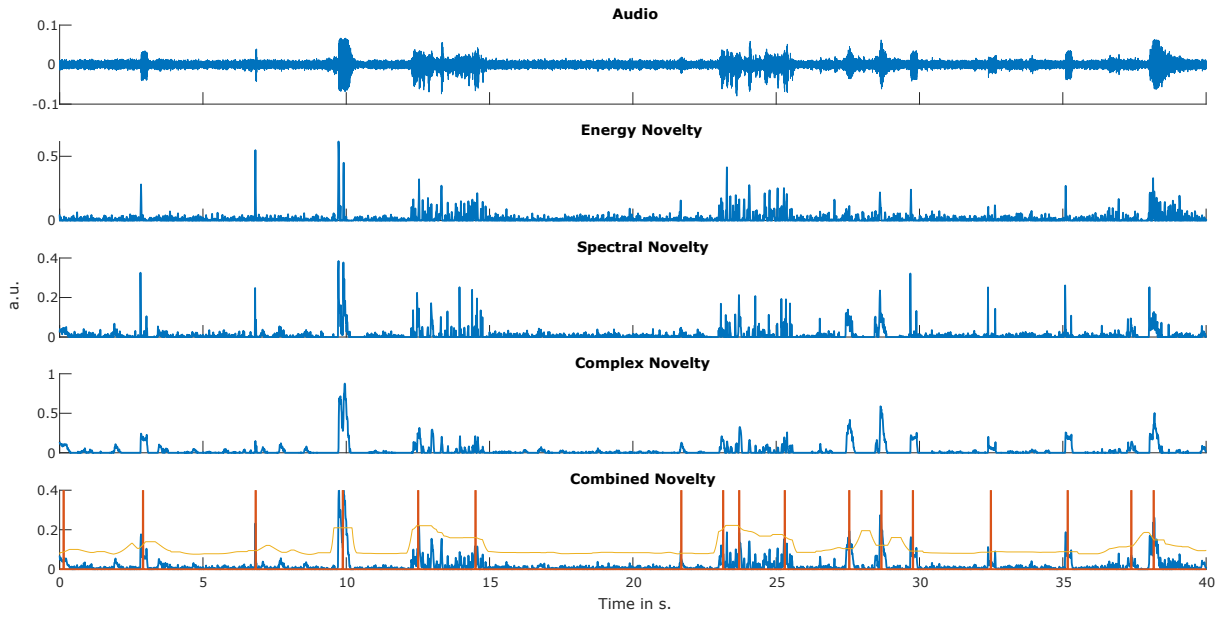

Figure 22: A representation of the novelty functions and the corresponding peak-picking algorithm. The top plot shows the audio signal played to Participant 1 during the narrow condition. Below are the corresponding energy, complex, and spectral novelty functions. The last plot shows the combined novelty functions in blue, the local average in yellow, and detected peaks in orange.

These models were then used to predict unseen data, and amplitude values were extracted.

To determine whether the peak values as a function of IOI were due to chance, we conducted a permutation analysis. Specifically, we randomly shuffled the allocation of sound onsets to their corresponding bins while preserving the overall distribution of onsets. This ensured that changes in the soundscape were still captured, but without a structured relationship to IOI. For each model, we summed the difference between each peak value, representing a rate of change score. The gradient of amplitude values over IOI for each bin model served as the aggregate score for comparison.

We applied the same analysis to 100 permuted model values, generating a distribution of 100 chance gradient values. Statistical significance was asserted at $p < 0.05$. To control for multiple comparisons, we applied a FDR correction.

8.2.9 *Acoustic Properties Beyond Sound Event Distance*

To investigate whether the observed neural amplitude differences between bins could be attributed to systematic acoustic properties of the stimuli, beyond IOI, we extracted

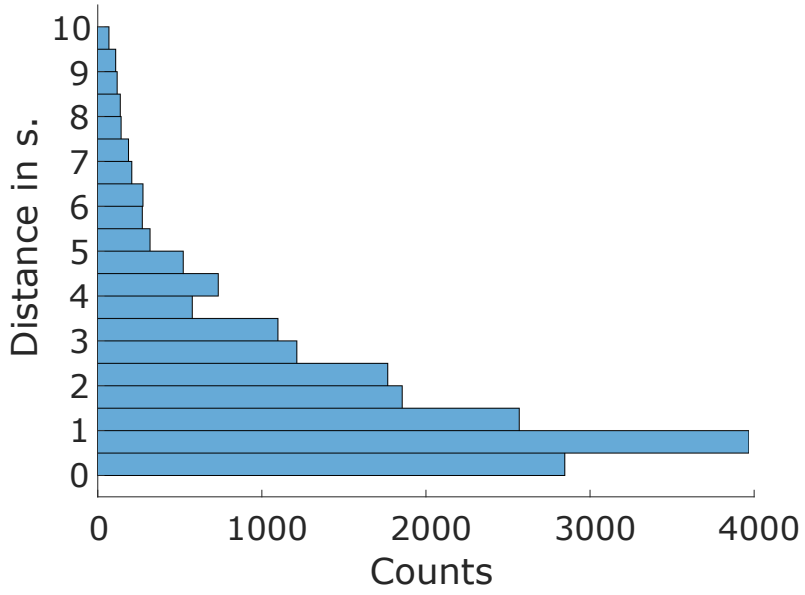

Figure 23: Histogram displaying the distance of onsets to the previous one over all participants.

two additional acoustic markers known to influence neural response magnitude. The first was the intensity (amplitude) of the sound event. Previous research has demonstrated that sound intensity is positively non-linearly related to neural response amplitude (Adler and Adler, 1989; Drennan and Lalor, 2019; López-Caballero et al., 2023). Interestingly, López-Caballero et al. (2023) examined both intensity and inter-stimulus interval (ISI) and found that both modulated the N1 and P2 components. Moreover, they reported a positive interaction between these factors, specifically, the modulatory effect of intensity on neural responses was more pronounced at longer ISIs, suggesting a dynamic interplay between temporal and intensity cues.

The second factor was the sharpness of the envelope onset. This characteristic has posed challenges in auditory research, as sound events with slow-rising envelopes complicate the accurate determination of perceptual onset (Rosenkranz et al., 2024). In such cases, automatic onset detection may not identify the optimal alignment point, potentially resulting in temporal smearing when responses are averaged. To quantify envelope sharpness, we calculated the gradient of the waveform within the first 50 ms following onset.

For each sound event, we derived values for these two markers: amplitude and sharpness, alongside the IOI to the preceding event. To assess their respective contributions, we employed a linear mixed-effects modeling approach. Due to the low signal-to-noise ratio associated with neural responses to rapidly successive events, we

did not model single-trial neural response amplitudes directly. Instead, we used IOI as the dependent variable to examine its relationship with the other acoustic predictors.

8.3 RESULTS

Previous research has established that for isolated tones, neural attenuation occurs when tones are played in close succession (López-Caballero et al., 2023; Wang et al., 2008a; Zacharias, König, and Heil, 2012). Here, we extend these findings by examining whether a similar modulation occurs in more complex, naturalistic auditory environments. Specifically, we examined whether the neural response to sound events is modulated by the inter-onset interval. To test whether accounting for varying IOI of sound onsets would show modulation of neural response in naturalistic soundscapes, we derived models by grouping sound event onsets in specific IOI ranges. We then tested whether the grouping of sound onsets into varying distance bins would also explain more neural variability.

8.3.1 *Modulating Acoustic Properties*

To evaluate the potential influence of systematic acoustic differences of the sound events of the different bins, we applied a linear mixed-effects modeling approach, with participants modeled as random intercepts. Before running the model, we assessed collinearity between predictors and found significant correlations between distance and intensity ($r = -0.15, p < 0.001$) and between sharpness and intensity ($r = 0.35, p < 0.001$), suggesting moderate interdependence among these variables. The correlation could impair the estimated coefficients.

The final model included sharpness, intensity, and their interaction (sharpness $*$ intensity) as fixed effects, with participants as random intercepts. The analysis revealed a significant main effect of intensity ($\beta = -195.92, SE = 9.75, t(18,327) = -20.09, p < 0.001$), indicating that higher sound intensity was reliably associated with shorter IOIs. In contrast, the main effect of sharpness was not significant ($\beta = 12.64, SE = 14.99, t(18,327) = 0.84, p = 0.399$), nor was the interaction between sharpness and intensity ($\beta = 17.41, SE = 21.33, t(18,327) = 0.82, p = 0.415$).

The estimated variance of the random intercept for participants was negligible ($4.36 * 10^{-14}$), suggesting minimal inter-individual variability in baseline inter-event distances. These results indicate that intensity is a robust negative predictor of IOI,

while sharpness and its interaction with intensity do not contribute significantly to explaining variability in distance.

8.3.2 *Peak Modulation*

Our results revealed amplitude modulation based on inter-onset interval. The larger the IOI, the larger the amplitude of the AEP. This finding was consistent across all variations of binning parameters. Neither the number of bins nor the specific constraints applied to the binning process significantly altered these findings. Specifically, we observed that the amplitude of the neural response was enhanced for sound events that followed a preceding sound at a greater IOI (Figure 24). To quantify this effect, we extracted peak values of the N1 and P2 components from group-averaged temporal response functions. The results demonstrated a clear trend in which neural response amplitude increased as a function of IOI between tones. Notably, we found that greater temporal spacing elicited a larger N1 peak and a stronger P2 peak across all binning variations (Figure 24).

To determine the relationship between the peak amplitudes and IOI, we fit a logistic, exponential, and polyfit model to the data. The first two models were inspired by existing literature (Herrmann et al., 2016; Zacharias, König, and Heil, 2012). The results showed that the optimal model to describe N1 was the polyfit model of second degree (R^2 = 0.84) and the logistic model for the P2 (R^2 = 0.75).

The results of the permutation testing revealed that the change of amplitude of the N1 and P2 for increasing IOI was significantly above the chance level ($p < 0.000$) (Figure 25). This effect was found for all models.

8.3.3 *Prediction Analysis*

Next, we examined whether incorporating IOI information into the neural model estimation improved the explained variability in the recorded signal. To do this, we compared the prediction accuracies of multiple models. First, we tested whether our IOI binned model outperformed chance-level predictions based on the results of the permutation testing (Figure 26). The results showed that all models outperformed the random models significantly: (**2**: $W = 202$, $Z = -3.62$, $p = 0.002$, $\rho = 0.57$; **3**: $W = 208$, $Z = -3.85$, $p = 0.001$, $\rho = 0.61$; **4**: $W = 204$, $Z = -3.7$, $p = 0.002$, $\rho = 0.58$; **5**: $W = 207$, $Z = -3.81$, $p = 0.001$, $\rho = 0.60$; **6**: $W = 210$, $Z = -3.92$, $p = 0.001$, $\rho = 0.62$; **7**: $W = 202$, $Z = -3.62$, $p = 0.002$, $\rho = 0.57$; **8**: $W = 198$, $Z = -3.47$, $p = 0.003$, $\rho = 0.55$). The

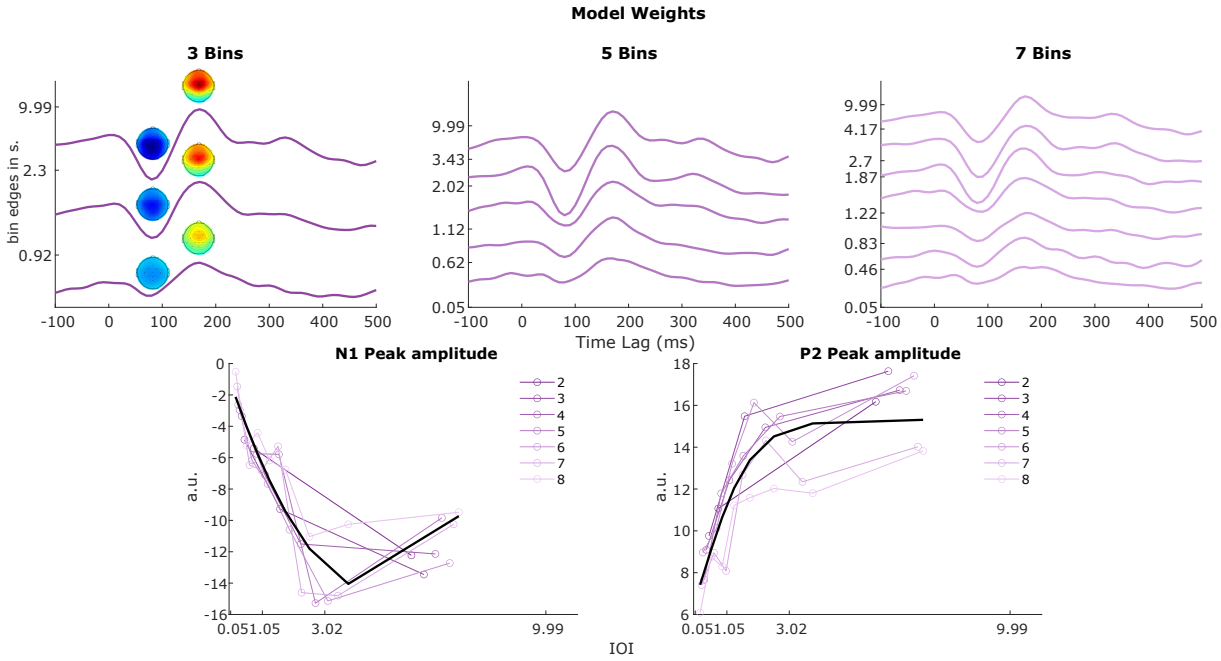

Figure 24: The top row shows model weights of three different bin models, i.e., 3, 5, and 7 bins. For the three-bin model, we also show the topographies at the N1 and P2 latencies for the different model weights. The Y-axis shows the upper edge of each bin. The bottom row shows the minimum and maximum magnitude of the N1 and P2 waves, respectively. The values are extracted for each bin of the seven different models. The models differ in their total number of bins, and bin edges vary to the uniform distribution constraint. On the bottom row, the left plot shows the distribution of N1 peak values as a function of bin edges. On the right, the same is displayed for the P2 values. The black line indicates the optimal model, fitted to all values.

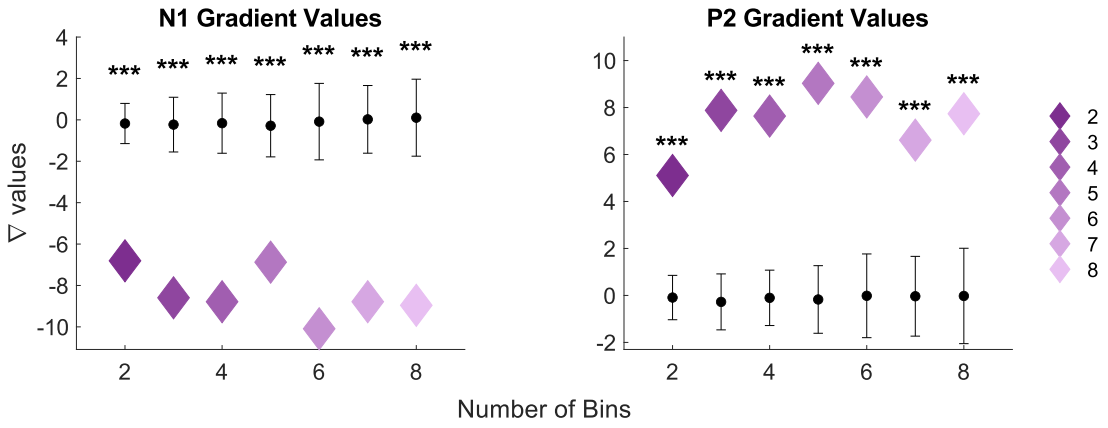

Figure 25: Displayed are the gradient amplitude values of the N1 and P2 for each binned model, respectively. The black bars indicate the standard deviation, and the dots the mean of the chance-level permutation values. The significance level here is indicated at *p $<$ 0.05, **p $<$ 0.01, and ***p $<$ 0.000.

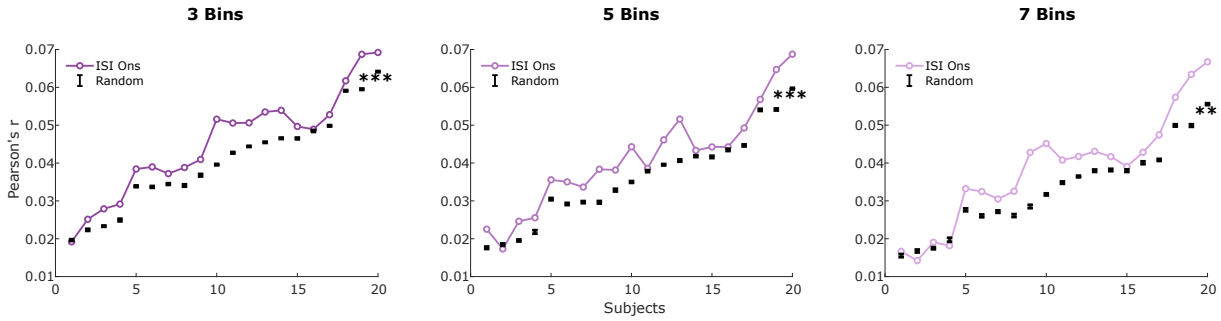

Figure 26: The plot shows the comparison of prediction accuracies between the binned models, i.e., 3, 5, and 7, and the random permutation model over participants. The significance level here is indicated at *p $<$ 0.05, **p $<$ 0.01, and ***p $<$ 0.000.

results indicated a non-linear relationship between model dimensionality and prediction accuracy. Specifically, the difference in prediction accuracy between structured and random models varied depending on the number of bins used. At extreme levels of model dimensionality, either very low or very high, the performance of the binned model only marginally exceeded the chance level. In contrast, intermediate levels of dimensionality produced the most pronounced differences in prediction accuracy. The greatest improvement occurred when using 3-6 bins, suggesting an optimal balance between information preservation and model complexity.

8.3.3.1 *Single Model Comparison*

Following our comparison of the binned IOI models to the random models, we aimed to determine whether the inclusion of binned temporal information would outper-

form a simpler model based on a single binary onset vector. Specifically, we contrasted our more complex model to the single-vector case, which does not incorporate binning information.

Despite observing amplitude modulation as a function of the temporal spacing of sound onsets, incorporating this information into model derivation did not yield higher prediction accuracies compared to using a single onset vector. In fact, the single onset model consistently outperformed the binned models across all cases (**3**: $W = 176$, $Z = 2.65$, $p = 0.039$, $\rho = 0.42$; **4**: $W = 210$, $Z = 3.92$, $p = 0.001$, $\rho = 0.62$; **5**: $W = 210$, $Z = 3.92$, $p = 0.001$, $\rho = 0.62$; **6**: $W = 208$, $Z = 3.85$, $p = 0.001$, $\rho = 0.61$; **7**: $W = 210$, $Z = 3.92$, $p = 0.001$, $\rho = 0.62$; **8**: $W = 210$, $Z = 3.92$, $p = 0.001$, $\rho = 0.62$). The only exception was the simplest binned model, which divided the data into two bins (**2**: $W = 128$, $Z = 0.86$, $p = 1$, $\rho = 0.14$) (Figure 27, middle). These results were stable regardless of whether binning was performed using linear or logarithmic spacing or when sample points per bin were uniformly distributed, as was the case here for the presented results. Additionally, accounting for condition differences did not significantly impact model performance.

Furthermore, we observed a deterioration of the prediction accuracy with an increasing number of dimensions. This suggests that data available for training may be a key factor driving the observed difference between the single and binned models. One potential concern is that the onset model and the binned IOI model may not be entirely comparable due to differences in the quantity of training data available for each set of feature weights.

To test whether, at comparable amounts of training data, our bin IOI model would capture neural data more optimally compared to the single onset model, we revisited the previous analysis. Specifically, we adjusted the number of onsets used for training in the single onset vector model to match the average number of onsets per bin for all binned models. For instance, in the case of the 2-bin model, roughly 200 onsets per bin are available. Thus, we randomly selected 200 onsets in the single onset model for training. This process was repeated 100 times for each model, for each participant, and condition.

The results, shown in Figure 27, contrast the performance of the adjusted single onset models with the full single onset vector model. Trivially, every adjusted onset vector was outperformed by the full single onset vector ($W = 210, Z = 3.92, p = 0.001, \rho = 0.62$). Notably, the statistics hold for all comparisons, since parametric tests do not consider the mean difference between distributions. Here, the respective statistics of the effect size, z value, W rank, and p value represent the upper bound.

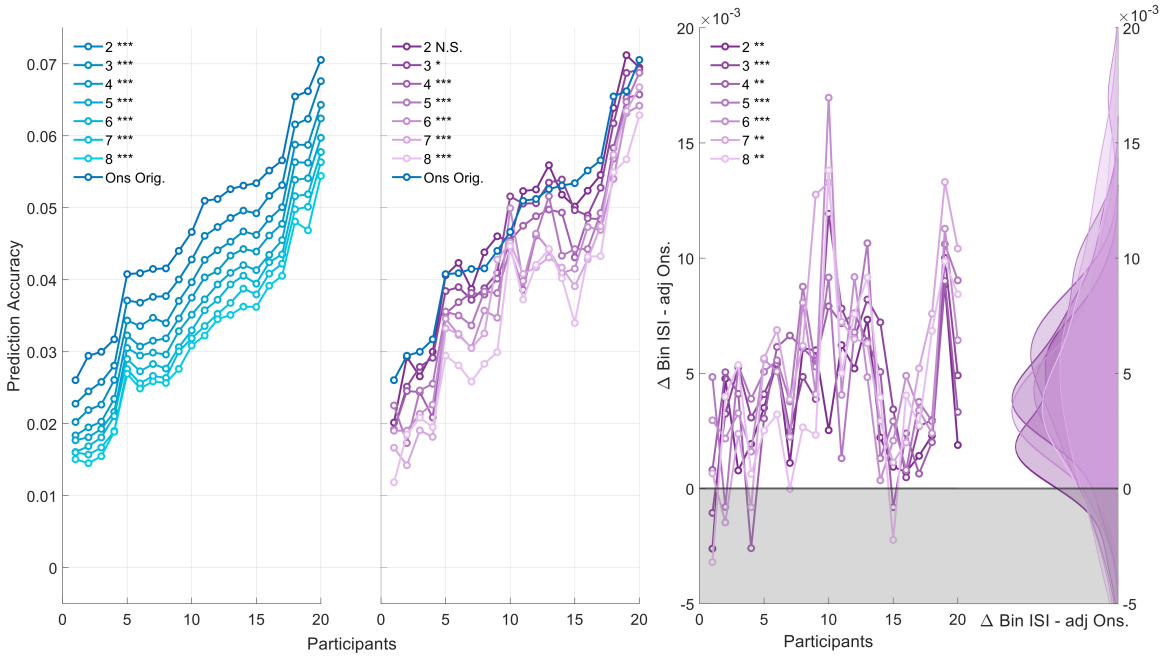

Figure 27: A shows the prediction accuracies over participants for different models. The left plot shows the prediction accuracies for the training adjusted onset model, which is based on the average number of onsets of the corresponding bin IOI model. The plot in the middle contrasts the prediction accuracy of the seven bin IOI models with the onset model. The plot on the right shows the difference between the bin IOI model and adjusted onset.

Given the impact of training data availability, we then revisited the comparison between the binned ISI models and the adjusted single onset vectors. The results indicate that the binned ISI models significantly outperformed their adjusted single onset counterparts across all bin variations (**2**: $W = 199$, $Z = -3.51$, $p = 0.003$, $\rho = 0.78$; **3**: $W = 208$, $Z = -3.85$, $p = 0.001$, $\rho = 0.86$; **4**: $W = 201$, $Z = -3.58$, $p = 0.002$, $\rho = 0.80$; **5**: $W = 208$, $Z = -3.85$, $p = 0.001$, $\rho = 0.86$; **6**: $W = 210$, $Z = -3.92$, $p = 0.001$, $\rho = 0.88$; **7**: $W = 204$, $Z = -3.7$, $p = 0.002$, $\rho = 0.83$; **8**: $W = 198$, $Z = -3.47$, $p = 0.003$, $\rho = 0.78$). These findings highlight the critical role of training data availability in the observed model performances for binary features.

8.3.3.2 *Extended Data Analysis*

When contrasting the prediction accuracy of the model containing both conditions with the single onset vector, we found that only the 2-bin and 3-bin models did not differ significantly from the single onset vector (**2**: $W = 103$, $Z = -0.075$, $p = 1$, $\rho = 0.017$; **3**: $W = 142$, $Z = 1.38$, $p = 0.51$, $\rho = 0.31$). For the more complex models

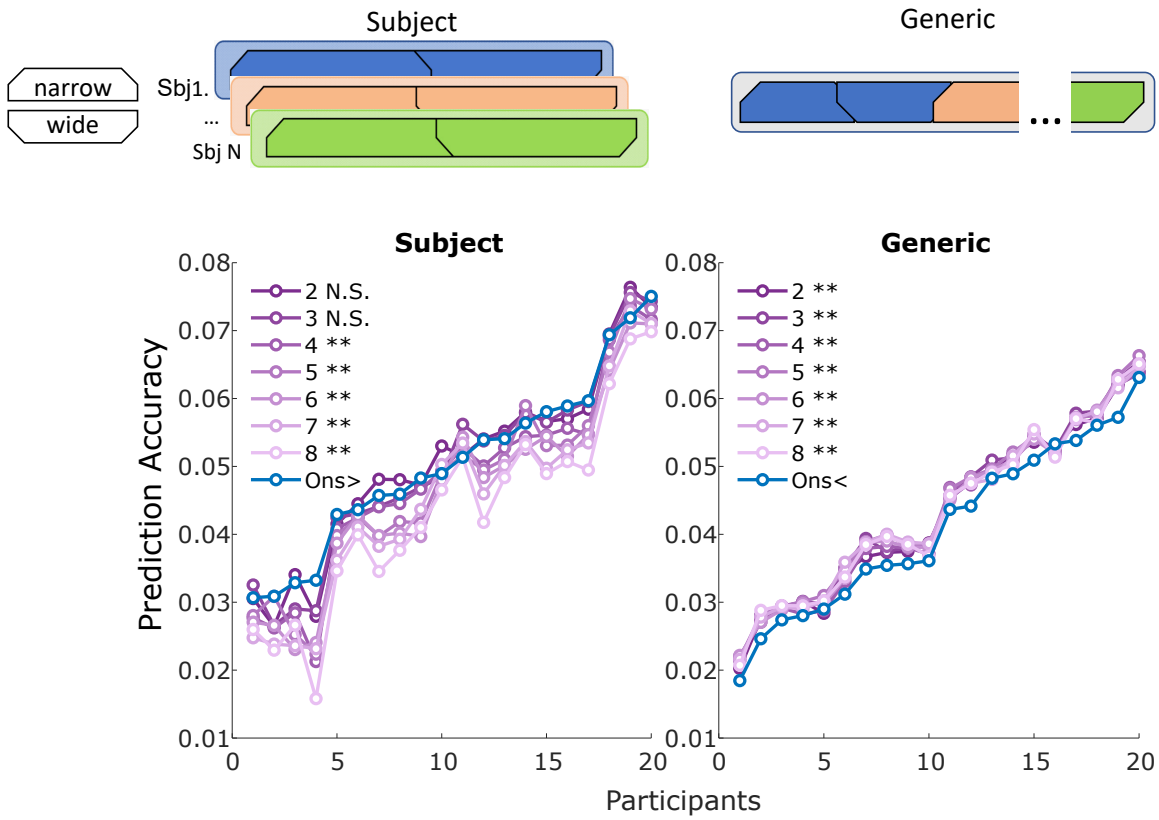

Figure 28: Shows the prediction accuracy of participants for different ways of training the data. The plot on the left shows the prediction accuracy over participants for merged condition data, where a model was trained on this longer dataset. The right side visualizes a generic model being trained. Here, each participant's merged condition dataset served as a held-out test set once. The $< / >$ next to the *Ons* model indicates the direction of significance. The significance level here is indicated at *$p < 0.05$, **$p < 0.01$, and ***$p < 0.000$.

(i.e., bin size > 3), prediction accuracy was significantly lower compared to the single onset vector (**4**: $W = 194$, $Z = -3.32$, $p = 0.005$, $\rho = 0.53$; **5**: $W = 190$, $Z = -3.17$, $p = 0.007$, $\rho = 0.50$; **6**: $W = 206$, $Z = -3.77$, $p = 0.001$, $\rho = 0.60$; **7**: $W = 204$, $Z = -3.7$, $p = 0.002$, $\rho = 0.58$; **8**: $W = 209$, $Z = -3.88$, $p = 0.001$, $\rho = 0.61$).

We then trained a generic model using all available participant data, leaving one participant out as a held-out test set (Figure 28). This was repeated for every participant once. The results showed that every binned model significantly outperformed the generic single onset model: (**2**: $W = 3$, $Z = -3.81$, $p = 0.001$, $\rho = 0.60$; **3**: $W = 2$, $Z = -3.85$, $p = 0.001$, $\rho = 0.61$; **4**: $W = 8$, $Z = -3.62$, $p = 0.002$, $\rho = 0.57$; **5**: $W = 1$, $Z = -3.88$, $p = 0.001$, $\rho = 0.61$; **6**: $W = 4$, $Z = -3.77$, $p = 0.001$, $\rho = 0.60$; **7**: $W = 6$, $Z = -3.7$, $p = 0.002$, $\rho = 0.58$; **8**: $W = 7$, $Z = -3.66$, $p = 0.002$, $\rho = 0.58$).

8.3.3.3 *Curse of Dimensionality*

The relation between training data and training data required is known as the curse of dimensionality, where more complex models require exponentially more training data. One way to mitigate this issue is to parameterize the binary onset vector by the normalized range of distance values. This approach is similar to weighting word onsets in speech processing based on meta-information, such as surprisal. However, adding a parameterized version of the onset vector to the model did not improve prediction accuracy compared to the single onset model ($p = 1$).

8.4 DISCUSSION

Neural response attenuation to acoustic properties has been investigated mostly on isolated pure tones (Beauducel et al., 2000; Herrmann et al., 2016; May and Tiitinen, 2010; Wang et al., 2008a; Wang et al., 2022). With this study, we extend those findings to naturalistic soundscapes.

This study provides evidence that neural attenuation is modulated by the IOI of sounds, generalizing validated lab findings to natural soundscapes. Here, the amplitude of the measured neural response was smallest when tones were close together and got larger the further tones were apart. This effect was robust over different numbers of pre-defined bins.

Based on these findings, we implemented the information into neural models to determine whether more neural variability could be explained. In summary, accounting for IOI information improves predictions of neural variability, provided that sufficient training data is available.

8.4.1 *Effects of IOI on Neural Data*

Our results replicate classic attenuation effects observed for tone sequences. (e.g., click trains or tone bursts) (Costa-Faidella et al., 2011; Herrmann et al., 2016; Lanting et al., 2013; Okamoto et al., 2004; Wang et al., 2008a; Zacharias, König, and Heil, 2012). Crucially, we extend these findings to naturalistic soundscapes, showing that neural responses remain sensitive to inter-event timing despite high variability in acoustic properties. This suggests that neural attenuation with IOI is a fundamental organizing principle in auditory processing.

Interestingly, this amplitude increase appears to plateau for IOIs exceeding three seconds and for more complex models in the case of the N1 peak to decrease again. Beyond these values, additional increases in IOI no longer resulted in further amplitude increase. However, it is important to interpret this plateau cautiously, since specific IOIs were not explicitly manipulated but emerged as a consequence of the binning constraint. Furthermore, due to the naturalistic nature of our stimuli, direct comparisons with studies using strictly controlled IOI intervals are limited. As a result, we cannot draw precise conclusions about the exact time point at which this asymptote occurs. Additionally, separate functions best model the N1 and P2 amplitude values, respectively. This suggests that different mechanisms of attenuation underlie the N1 and P2 and thus possibly reflect separate neural generators (Altmann et al., 2008; Herrmann et al., 2016; López-Caballero et al., 2023).

8.4.2 *Neural Mechanisms*

Neural attenuation is a fundamental principle of sensory processing, observed across all sensory modalities and at multiple levels of the neural hierarchy, from peripheral receptors to cortical areas. This general phenomenon allows organisms to remain sensitive to new and changing stimuli by dynamically adjusting neural responsiveness in the face of repeated or sustained input (Dean, Harper, and McAlpine, 2005; Hicks and McDermott, 2024; Ulanovsky, Las, and Nelken, 2003).

Previous research has proposed two primary frameworks to explain auditory neural attenuation: habituation (often framed within a predictive coding framework) and adaptation. Habituation accounts argue that repeated or predictable stimuli generate sensory expectations, leading to reduced neural responses when those predictions are confirmed and increased responses when they are violated (Costa-Faidella et al., 2011; Näätänen and Picton, 1987; Ruusuvirta, 2021; Silva, Melges, and Rothe-Neves, 2017; Wang et al., 2008a). In contrast, adaptation accounts propose a physiological explanation, suggesting that repeated stimulation leads to reduced neuronal responsiveness due to mechanisms such as synaptic fatigue or depletion, independent of stimulus predictability (Budd et al., 1998; López-Caballero, 2025; López-Caballero et al., 2023; May and Tiitinen, 2010; Rosburg and Mager, 2021; Rosburg, Weigl, and Mager, 2022).

Our study provides a unique opportunity to discuss these accounts because the naturalistic soundscape we used was highly variable spectrally and temporally. This random nature rules out the formation of stable sensory predictions, making a habituation or predictive coding explanation less likely. Additionally, the soundscape's

broad spectral variability suggests that the attenuation effects we observe are not simply the result of adaptation confined to narrowly tuned, organized auditory neurons. Interestingly, research suggests that temporal and spectral characteristics are adapted independently (Briley and Krumbholz, 2013). Thus, there may be separate neural processes underlying spectral and temporal adaptation.

In line, our findings point toward a general temporal sensitivity in auditory processing. We propose that the observed attenuation reflects broader adaptation mechanisms, such as synaptic depression or slow hyperpolarization, that operate independently of spectral content and are consistent with temporal recovery models described in previous work. The optimal parameters to describe the decay/ recovery and exact models to depict the attenuation are still under debate (Herrmann et al., 2016; Lanting et al., 2013; Regev et al., 2021; Wang et al., 2008a; Zacharias, König, and Heil, 2012).

A promising neural substrate for these effects is the extralemniscal auditory pathway, which has been linked to stimulus-specific adaptation, detection of sudden environmental changes, and supramodal modulation of global brain states (Carbajal and Malmierca, 2018; Shine et al., 2023; Somervail et al., 2021; Somervail et al., 2025; Willmore and King, 2023). Importantly, it lacks the strict tonotopic organization of the lemniscal pathway and has been proposed as the neural substrate of the MMN, reflecting an error or novelty signal. Given its broader tuning and role in orienting responses, the extralemniscal system may underlie the non-spectral, time-sensitive attenuation we observed when sound events occurred in close temporal succession.

Despite our promising results, it remains an ongoing challenge to determine the underlying neural mechanisms that are responsible for neural attenuation. Although some theories have been proposed, it remains to be shown whether their integration explains attenuation in response to real-world soundscapes. Future studies should continue to systematically vary both spectral and temporal aspects jointly to determine whether the same or different mechanisms underlie the neural attenuation.

Although unclear where the attenuation occurs, our findings suggest that the temporal sensitivity of the auditory system persists even in complex, unpredictable environments. Importantly, they highlight that attenuation processes extend beyond stimulus-specific mechanisms in the face of dynamic sensory input.

8.4.3 *Potential Confounding Factors*

Given the complexity of the soundscape under investigation, we examined whether systematic acoustic differences existed between sound events as a function of their IOI.

Specifically, we focused on two features previously implicated in modulating neural response amplitudes (Drennan and Lalor, 2019; López-Caballero et al., 2023) - sound intensity and envelope sharpness.

Correlational analyses and linear mixed-effects modelling revealed a systematic difference in intensity with IOI, where sound events occurring in close succession tended to have a higher intensity than those spaced further apart. This relationship might be partially biased by the adaptive threshold to select onsets. Here, novelty peaks need to surpass a threshold that is based on the context of the soundscape. Thus, successive onsets may need a larger amplitude to exceed the context-driven threshold, driving the negative relationship between IOI and intensity.

Given the well-established association between increased stimulus intensity and stronger neural responses (Adler and Adler, 1989; Beauducel et al., 2000; Drennan and Lalor, 2019; López-Caballero et al., 2023), the fact that closely spaced events were more intense should have amplified rather than diminished their evoked responses. However, our results show the opposite effect: sounds occurring with shorter IOIs elicited attenuated neural responses. This dissociation suggests that intensity differences did not drive the observed IOI-related amplitude modulation, which is in line with findings of López-Caballero et al. (2023), who also found a dissociation between these two factors. On the contrary, intensity-related enhancement may have masked part of the IOI effect, making our findings a conservative estimate of the actual modulation associated with IOI.

Besides accounting for the sharpness and intensity, the creation of the soundscape itself may have introduced a bias, specifically, through the RMS normalization of every sound to the average level. While normalization is standard practice to control for loudness-related confounds in neural analyses, this procedure may somewhat reduce ecological validity by artificially equalizing loudness levels that naturally vary. Consequently, participants' subjective perceptions and neural responses could have been slightly affected. However, given that individual sounds were individually adjusted prior to spatial separation using gain parameters (Kayser et al., 2009), we aimed to retain as much auditory realism as possible. Future studies could explicitly assess the impact of such loudness normalization procedures on subjective naturalness and corresponding neural dynamics.

Finally, we acknowledge the possibility that systematic motor-related artifacts or attentional biases could have influenced our neural findings. Specifically, motor responses associated with following verbal instructions or reacting to relevant tones could potentially reduce neural amplitudes (e.g., N1/P2). Conversely, increased attention towards verbal instructions might systematically enhance neural amplitudes for

those stimuli compared to less relevant background sounds, potentially confounding our observed effects of onset intervals. However, several factors mitigate these concerns in our experimental design: First, the condition-relevant tones were randomly embedded within the soundscape, minimizing any systematic temporal alignment between attention or motor responses and particular stimulus categories. Second, motor responses occurred significantly later than the neural responses analyzed (e.g., N1/P2), reducing the likelihood that motor execution systematically influenced these early neural signals. Additionally, to further control for motor artifacts, our EEG preprocessing explicitly identified and removed motor-related EEG activity via ICA, substantially reducing potential residual contamination. Lastly, because acoustic onsets were derived indiscriminately from the raw soundscape, systematic biases induced by increased attention to relevant speech compared to irrelevant background stimuli are unlikely to have influenced our findings.

8.4.4 *Unconsidered Factors Influencing Peak Amplitude Modulation*

Our analysis, focused primarily on the IOI as a key modulator of neural responses while accounting for sound intensity and sharpness. However, given the complexity of naturalistic soundscapes, other acoustic and contextual factors may have played a role, which were not systematically investigated in the present study.

One such factor is the duration of the preceding sound. Lanting et al. (2013) reported that longer adapter durations led to greater N1 suppression in a paired-click paradigm, highlighting duration as a potential modulator of adaptation. Although this effect was shown using simple tones, its role in complex soundscapes remains uncertain and warrants further investigation.

Contextual predictability is another critical factor. Previous work has demonstrated that neural attenuation depends on the stimulation history (Herrmann et al., 2016; Zacharias, König, and Heil, 2012). Notably, reduced attenuation effects are found under random IOI conditions compared to highly predictive sequences. How the general context impacts neural attenuation in autocorrelated soundscapes needs to be investigated by future studies. Evidence from a recent behavioural study suggests that response adaptation to stationary soundscapes occurs faster compared to those with increased spectral variability (Hicks and McDermott, 2024).

Finally, spectral similarity of successive sounds and global soundscape statistics has also been implicated in response attenuation. Herrmann et al. (2013) found stronger adaptation for spectrally similar tones. However, studies using more complex stim-

uli (e.g., vowels, animal vocalizations) have not observed such effects (Altmann et al., 2008; Silva, Melges, and Rothe-Neves, 2017). Given the broadband nature of our stimuli, the influence of spectral similarity remains ambiguous and was not directly tested here.

Taken together, these findings highlight that while IOI is a critical factor in auditory attenuation, other acoustic dimensions such as duration, spectral content, and stimulus context can also influence peak amplitude modulation. Future work should aim to incorporate these variables into more comprehensive models to better disentangle their individual and interactive contributions to auditory processing in naturalistic environments. As such, it would provide insights into how the brain processes complex soundscapes with greater detail.

8.4.5 *Prediction Accuracy*

8.4.5.1 *Model Performance*

We have shown that integrating IOI provides meaningful information, as shown by the comparison between those models with random onset-bin allocation. Since both models contain onsets at identical time points, but only one assigns onsets based on the IOI between successive sound events, while the other does so randomly, the increased accuracy in the informed model indicates that IOI serves as a meaningful feature for neural prediction.

8.4.5.2 *Trainings Data Availability*

When we compared the IOI model against a single onset predictor (i.e., the simple onset model), performance was worse. This result was unexpected, given that the previous analysis showed the model to contain meaningful information. These findings also contrast with the study by Drennan and Lalor (2019), who showed that deriving features that account for the non-linear response of the brain improves model estimation and consequently the amount of neural variability explained.

This discrepancy can be explained by reduced training data per predictor: dividing the onset vector into multiple IOI-based bins leads to fewer events per bin. This impairs the derivation of reliable weights. This observation is crucial in explaining the inferior performance of the bin IOI model to the simple onset model. Given that prediction accuracy is strongly influenced by the availability of training data (Desai, Field, and Hamilton, 2023; Mesik and Wojtczak, 2023). To verify this interpretation, we controlled

for data availability by reducing the simple onset model to match the number of events per bin in the IOI model. Under these conditions, the IOI model outperformed the reduced simple model, confirming that IOI carries predictive value.

We further tested this by increasing the available training data, either by pooling conditions within subjects or by training generic models across participants. In both cases, the IOI model benefited from increased data, often surpassing the simple model. Interestingly, the overall performance of generic models was lower compared to the individual models, and the difference between models was no longer visible. The reduced performance in the generic model compared to subject-specific models is likely due to the latter model better capturing individual nuances. This is in line with previous research showing that when sufficient training data is present, generic models underperform compared to subject-specific models (Mirkovic et al., 2015). Beyond this point, the subject-specific model is superior. The lack of difference between the bin IOI models suggests a ceiling effect of training data, indicating that further data would not yield additional performance gains. This underscores the need to balance feature complexity with the amount of available training data to avoid compromising model performance.

While increasing data availability helped address model complexity, we also examined whether simplifying the model could yield similar results. This approach was inspired by speech processing research, where word onset models are supplemented with parametric word surprisal scores to incorporate additional meaningful information into neural response estimation (Brodbeck, Hong, and Simon, 2018). Analogously, we weighted binary onsets by their respective distance to the previous onset. However, this approach did not yield significant improvements in prediction accuracy. Highlighting that the binned approach is capturing non-linear neural dynamics.

8.4.5.3 *Practical implications*

These findings highlight a key trade-off: while incorporating temporal context (e.g., inter-onset interval, IOI) improves neural response prediction, increased model complexity requires sufficient training data to avoid performance loss. While shorter lab-based recordings may not provide enough data for models to benefit meaningfully from IOI-based features, longer recordings, particularly those collected in real-world, non-laboratory settings, offer an opportunity to leverage temporal information effectively. As longitudinal, everyday-life recordings (Hölle and Bleichner, 2023; Hölle, Meekes, and Bleichner, 2021; Korte, Haupt, and Bleichner, 2025; Rosenkranz et al., 2024) become increasingly available, incorporating temporal structure such as IOI may

significantly enhance model performance and our ability to predict neural responses in complex, naturalistic contexts.

8.4.6 *Conclusion*

Our results provide important insights into how the brain processes complex soundscapes in everyday life. We showed that temporal structure, specifically, the timing between sound events, is a critical dimension that modulates neural responses, even in highly variable, naturalistic settings. By demonstrating that shorter IOIs attenuate auditory neural responses and that IOI-based models can outperform simpler onset models (when data availability allows), we highlight the importance of integrating temporal features into the study of auditory scene analysis. These findings lay the groundwork for future research linking neural processing of soundscapes to perceptual, cognitive, and behavioural outcomes, advancing our understanding of how the brain interprets and adapts to the acoustic complexity of the real world.

Funding Information

This work has been funded by the Deutsche Forschungsgemeinschaft: 10.13039/501100001659, ID: 490839860; 10.13039/501100001659, ID: 411333557

Acknowledgements

We would like to thank Manuela Jäger and Silvia Korte for the fruitful discussions throughout the development of the study.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Potential Declarations

During the preparation of this work, the author(s) used ChatGPT 4o and the free version of ChatGPT (mid 2024) in order to improve language and readability of selected sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author Contribution

TH, MR, and MB Designed Research, TH Performed Research, TH Analyzed Data, TH and MB Wrote the Paper

9

AUDITORY ATTENTION TO GO

"The trouble with having an open mind, of course,
is that people will insist on coming along
and trying to put things in it."

Pratchett (1996)

Auditory Attention Decoding to go with mobile and portable hardware

Thorge Haupt[1], Lisa Straetmans[1], Kamil Adiloglu[3], Martin G. Bleichner[4,5], Stefan Debener[1,2,5]

[1]Neuropsychology Lab, Department of Psychology, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

[2]Cluster of Excellence Hearing4all, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

[3]Sonova Consumer Hearing GmbH, Hanover, Germany.

[4]Translational Psychology Lab, Department of Psychology, Carl Von Ossietzky, Oldenburg, Germany

[5]Research Center for Neurosensory Science, University of Oldenburg, Oldenburg, Germany

This chapter is identical in content to the version submitted to:
Hearing Research

Abstract

Auditory Attention Decoding (AAD) has emerged as a promising approach for neuroadaptive hearing technology. However, its feasibility in naturalistic, mobile settings remains underexplored. In this study, we investigated AAD using a minimal and wearable setup comprising of around-the-ear EEG and portable hearing aid research hardware. Participants performed an auditory attention task under both controlled (seated) and naturalistic (walking) conditions with varying auditory scene complexity (single- and dual-speaker conditions, cafeteria background noise). We applied both backward (decoding) and forward (encoding) models to assess attention-dependent neural tracking of continuous speech.

Decoding results replicated established trends, with the highest reconstruction accuracy for the single speaker, followed by dual-speaker attend and then ignored. These trends were extended to the seated and walking contexts, initially demonstrating the potential of cEEGrids for mobile AAD. Post hoc forward modeling revealed an artifactual response in the walking condition, biasing the reconstruction and prediction accuracy. Further analysis indicated that this effect could reflect hardware-related interference, possibly from contact between hearing aid cables and electrodes. Importantly, the artifact, while speech-locked, was detectable only through the forward model.

Despite these constraints, our results demonstrate the feasibility of decoding auditory attention using a lightweight, unobtrusive EEG system in real-world scenarios. This work emphasizes the need for robust hardware integration, real-time validation, and user-centered design in future AAD applications. Our findings add a critical step towards practical brain-computer interface solutions for hearing support in everyday environments and highlight the importance of interpretable methods in application-driven research.

9.1 INTRODUCTION

A key cognitive function is the ability to selectively filter sensory input according to behavioral relevance. This is classically illustrated by the "cocktail party phenomenon," where a listener can focus on a single speaker while suppressing irrelevant background noise (Cherry, 1953; Haykin and Chen, 2005; McDermott, 2009). While this attentional filtering mechanism operates effectively in individuals with normal hearing, people who are hard of hearing report problems in situations where multiple sound sources are present (Festen and Plomp, 1990; Mirkovic et al., 2019; Reiss and Molis, 2021). A current issue is that assistive hearing technologies lack the ability to select sound streams that are relevant to a listener. This absence of selective amplification stems from the device's inability to access the user's listening intent, contributing to imperfect user satisfaction and listening fatigue in complex environments (Mustafa and Krishnamurthy, 2025).

To address this limitation, it has been proposed to bridge the gap between user intent and the hearing aid by integrating non-invasive neural recordings. EEG is particularly suitable as a neuroimaging modality, given its high temporal resolution (Winkler, Denham, and Escera, 2013). This temporal precision is a requirement for real-time decoding of user intent, as well as to monitor attentional shifts, which can occur within a few hundred milliseconds (Larson and Lee, 2013). A pivotal development nearly two decades ago revealed that abstract auditory representations, such as the amplitude envelope of speech, can be mapped onto EEG recordings using TRF (Lalor et al., 2009). These models can be used to explain neural variability by either encoding the envelope in the EEG (forward modelling) or decoding the envelope from the EEG (backward modelling) (Crosse et al., 2016). Interestingly, the filtering mechanism of user intent can be inferred by comparing how much neural variability is explained by competing speech streams. Attended speech streams consistently account for more neural variance than ignored ones, establishing a robust neural marker for AAD (Ding and Simon, 2012b; O'Sullivan et al., 2015). Over the past decade, AAD has been extensively validated across diverse experimental conditions, varying the number of speakers, types of background noise, decoding algorithms, recording hardware, and even movement contexts such as walking (Herrmann, 2024; Jaeger et al., 2020; Mirkovic et al., 2015, 2019; O'Sullivan et al., 2015; Straetmans, Adiloglu, and Debener, 2024; Tallus et al., 2015). While this growing body of work highlights AAD's potential, two major challenges remain before it can be deployed in real-life hearing support systems.

There are several challenges that remain, such as real-time implementations of AAD pipelines, where temporal mismatches between user intent and auditory feedback

needs to be minimized. Second, unobtrusive EEG acquisition hardware has to be comfortable enough for long-term daily use and tolerate natural movement patterns. Most prior studies achieving reliable decoding have relied on high-density, cap-based EEG systems wired up to bulky, immobile hardware. While these systems offer high spatial resolution, they are bulky, conspicuous, and impractical for everyday use (Bateson et al., 2017; Debener et al., 2012). Here, we aim to address the second challenge, using a minimal and unobtrusive setup.

Recent advances in wearable EEG technologies offer promising alternatives, ranging from low-density dry electrode caps (Ciccarelli et al., 2019), to around- (Bleichner and Debener, 2017), and in-ear solutions (Geirnaert, Kappel, and Kidmose, 2025; Thornton, Mandic, and Reichenbach, 2024). Among them, cEEGrids, flex-printed electrode arrays placed around the ear, stand out for their unobtrusiveness and ease of use (Debener et al., 2015). These systems are lightweight, require minimal preparation, and can be integrated with other wearable devices (Hölle and Bleichner, 2023; Hölle et al., 2022). Importantly, cEEGrids have been shown to capture key neural markers of auditory processing, including speech envelope tracking and AAD (Holtze et al., 2022; Mirkovic et al., 2015).

In two previous studies, we provided the first evidence of AAD being possible even in mobile, indoor (Straetmans et al., 2022) and outdoor (Straetmans, Adiloglu, and Debener, 2024) conditions, when participants walk freely. However, both studies used traditional EEG electrode caps, which are impractical for self-applied, assistive EEG technologies. The reduced number of electrodes used in ear-EEG systems and their peripheral positioning restricts the richness of the neural data and impacts reconstruction accuracy (Geirnaert, Kappel, and Kidmose, 2025; Mirkovic et al., 2016). This raises the question of whether sufficiently high AAD results can be achieved with ear-EEG recordings obtained in more ecologically valid, real-world contexts that involve participant movement, multitasking, and dynamic environmental noise.

The present study compares AAD in both laboratory-like and real-world scenarios. Specifically, we investigate neural speech tracking under single and dual-speaker conditions in which participants attend to one of two concurrent speech streams. We compare neural tracking performance during stationary (sitting) and mobile (outdoor walking) conditions to evaluate the impact of movement and environment on decoding reliability. To this end, we combine a highly portable ear-EEG system running on a smartphone with a closed-loop research hearing aid system, the portable hearing lab (Pavlovic et al., 2018). The current data set is part of a larger study where the same experimental paradigm was conducted using a traditional EEG cap (Straetmans, Adiloglu, and Debener, 2024). We expect to replicate the findings of the previous

study, which showed that single speakers were decoded most accurately. Furthermore, they showed that the AAD can be extended to the mobile walking condition. While we expect to replicate the trends, the overall effect size and reconstruction accuracies are expected to be lower, primarily due to the number and location of the EEG electrodes.

9.2 MATERIALS AND METHODS

The dataset reported in this paper is part of a larger study that involved two sessions per participant, one with a traditional electrode cap (reported in Straetmans, Adiloglu, and Debener (2024)) and one session with an ear-EEG system (reported here).

9.2.1 *Participants*

A total of 28 participants took part in the study. Due to participant drop-out (n=1), acute health problems (n=1), neurological conditions (n = 2), compromised hearing (n=1) and technical difficulties (n=2) data from N = 21 participants (10 female, mean age 25.2 years) were available for a BTL block and data from 20 of the same participants for Lab 1 and Lab 2 blocks (10 female, mean age 25 years). All participants reported normal hearing and normal or corrected to normal vision and were compensated for participation. The study was approved by the University of Oldenburg Ethics Committee (DRS.EK/2021/078).

9.2.2 *Paradigm*

Each participant took part in two experimental sessions, approximately three weeks apart. One session was conducted using a conventional 32-channel EEG cap system (results reported in Straetmans, Adiloglu, and Debener (2024)), and the other using two cEEGrid electrode arrays placed around the ears ((Debener et al., 2015); reported here). Both sessions followed the same experimental protocol and were conducted on weekdays between 09:00 a.m. and 12:30 p.m. Upon arrival, participants first completed an informed consent form and a brief questionnaire assessing general well-being and neurological health. During the first session, participants also underwent a pure-tone audiometry screening to confirm normal hearing thresholds.

Following preparation, participants were fitted with EEG and ECG electrode, the Portable Hearing Lab (PHL), and an Android smartphone for stimulus presentation and data recording. The technical setup is described in more detail in the Section 9.2.7.

The main experimental protocol consisted of three approximately 60-minute blocks: Lab 1, BTL, and Lab 2. Each block consisted of four narrative segments, each approximately nine minutes long. Each block included both single-speaker and dual-speaker attention paradigms (Figure 29). Participants were instructed to attentively listen to a target narrative read by either a male or female speaker. When multiple audio streams were presented, participants were asked to focus solely on a designated speaker and to ignore any other concurrent speech and/or background noise. After each story, participants answered four multiple-choice questions assessing their adherence to task instructions and comprehension. Subsequently, participants rated their perceived fatigue and task exhaustion on 7-point Likert scales (1 = not at all tired/exhausting; 7 = extremely tired/exhausting).

9.2.3 *Laboratory Blocks (Lab 1 and Lab 2)*

The Lab 1 and Lab 2 blocks took place at the Hörzentrum Oldenburg, Germany. Participants were seated in the center of a 16-loudspeaker array arranged in a circle. To systematically increase cognitive load and simulate real-world acoustic challenges, each of the four stories in the block featured more complex auditory scenes:

1. Single speaker: A single narrative played through one loudspeaker, with no background noise.
2. Dual speaker: A second narrative, spoken by a voice of the opposite sex, played concurrently through a different loudspeaker.
3. Single speaker + background: The target narrative was presented alongside a background noise recording of a busy cafeteria.
4. Dual speaker + background: The most complex condition included the target narrative, a distractor speaker, and the cafeteria background noise simultaneously.

The sequence of these conditions was fixed across all participants and the two lab blocks to allow for a controlled increase in listening effort and fatigue over time (Figure 29).

9.2.4 *Beyond-the-Lab Block (BTL)*

The BTL block was designed to test auditory attention decoding under mobile, real-world conditions. During this block, participants completed the single or dual speaker auditory condition, without the added cafeteria noise, due to the presence of concurrent ambient environment noise. This block took place immediately after Lab 1 and was conducted either in a quiet indoor hallway (sitting condition) or outdoors (walking condition) on a predefined walking route near the Hörzentrum. The walking route included both quieter streets and busier, traffic-heavy areas (introducing natural acoustic variability). The stories in the BTL block were presented with one or two speaker narratives, without added artificial background noise, but with real environmental noise being present. Unlike in the laboratory blocks, the order of scenarios (e.g., sitting vs. walking, single vs. dual speaker) was randomized across participants to minimize potential order effects (see Figure 29). Thus, participants started either walking or sitting, and either with the single or dual speaker condition. The movement conditions always switched (i.e., walk, sit, walk, sit), while the auditory conditions were presented in blocks (i.e., single, single, dual, dual). An example would be a participant starting with the walking condition and a single-speaker story. The next condition would be sitting and still listening to the single speaker. This would be followed again by walking, but then with the dual speaker auditory condition. Furthermore, the auditory streams were presented over inserted earphones, instead of over a speaker ring, to ensure free movement of the participant.

9.2.5 *Stimuli*

9.2.5.1 *Speech Materials*

All speech stimuli were audiobooks narrated in German and edited using Audacity (v2.3.0, https://www.audacity.de/). Preprocessing steps included the removal of DC offsets and normalization of the maximum amplitude to 1.0 dB to ensure consistent peak levels across both audio channels. This normalization helped maintain comparability across different narrative recordings. Additionally, prolonged silent pauses within the narratives were truncated to optimize the temporal structure and maintain listener engagement.

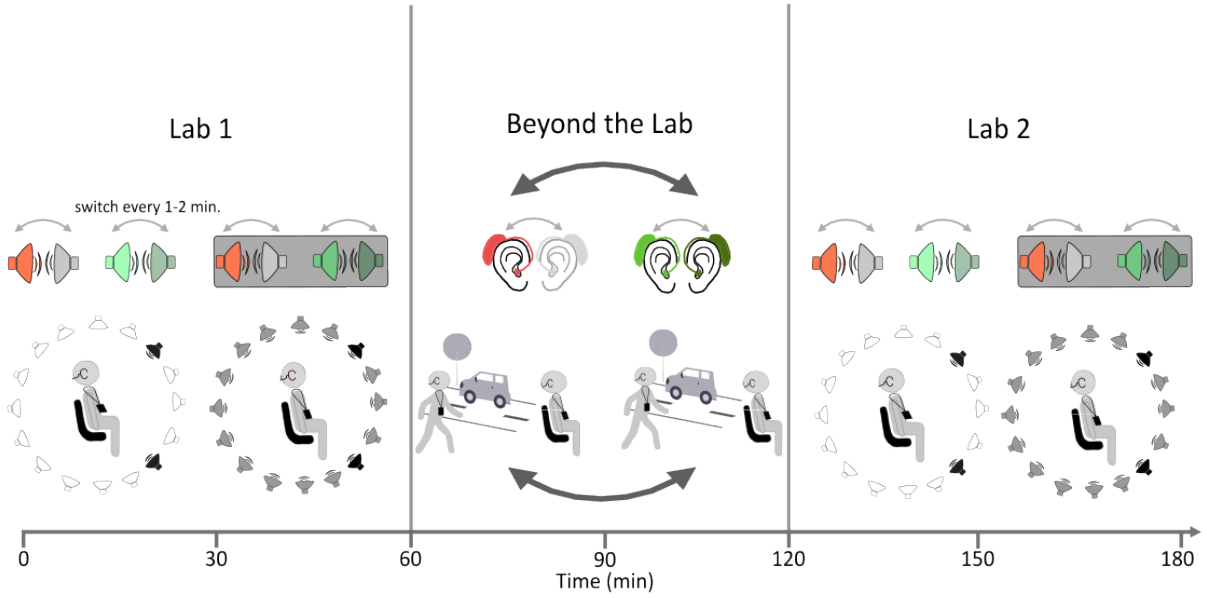

Figure 29: Paradigm of the experiment. In the first Lab session, participants were seated within a speaker ring. The experiment always started with a single speaker, then a dual speaker, then a single speaker with noise, and finally a dual speaker with noise. In the BTL session, participants were randomly assigned to either walking or sitting, and starting with either single or dual speaker. The speech was played over the inserted earphones of the hearing aids. The last session took place in the lab again and repeated the same order as that of Lab1. The color denotes the number of speakers: green = single speaker, orange = dual speaker, the grey box indicates the presence of cafeteria noise. The arrow at the bottom shows the order of the experimental blocks.

9.2.6 *Stimulus Presentation*

Participants were instructed to attend to a narrative spoken by either a male or female speaker and to maintain attention on that speaker's voice throughout the entire experiment. This was done to not introduce an additional variable of narrator gender and because the stories over the blocks formed together a coherent story. The assigned speaker's gender remained constant across both appointments for the participants; however, the actual speaker and narrative content varied between recording sessions. Auditory scenes were generated using the Toolbox for Acoustic Scene Creation and Rendering (TASCAR; Grimm, Luberadzka, and Hohmann (2019)). In single-speaker conditions, the target narrative was presented from either -30 (left) or $+30$ (right) relative to the participant's midline. In dual-speaker conditions, the to-be-ignored narrative was simultaneously presented from the opposite side (e.g., if the attended speaker was on the left at -30, the ignored speaker was on the right at $+30$). To maintain spatial engagement and reduce habituation, the spatial position of the attended speaker switched after intervals of 60 to 120 seconds. These switches occurred at semantically meaningful points within the stories. In single-speaker conditions, only the attended narrative was presented, and the side of presentation alternated similarly across time. Event markers for stimulus onset, attention switches, and condition changes were generated in MATLAB.

In the BTL block, participants listened to the stimuli through behind-the-ear hearing aids, each equipped with a receiver and two microphones. Audio was delivered via the open-fit receivers inside the ear canals. To simulate spatial separation of the audio streams, the raw speech signals were convolved with a general head-related impulse response according to Kayser et al. (2009). The spatial positioning of the narratives mirrored the lab-based setup: the attended stream was positioned at -30 or $+30$, and in dual-speaker conditions, the ignored stream was presented from the other direction. Stimuli in the BTL block were presented via Presentation® (Neurobehavioral Systems, Inc., Albany, CA, USA) running on an Android smartphone. Event markers for stimulus onset, attention switches, and condition changes were generated in the PHL.

At the start of Lab 1 and the BTL session, participants were able to adjust the playback volume to a comfortable level after listening to an audio snippet. The adjustment occurred either by instructing the experimenter (Lab 1) or by using the smartphone volume buttons themselves (BTL).

9.2.7 *Data Recording*

9.2.7.1 *Portable Hearing Lab (PHL)*

Data stream acquisition was performed using the PHL, a lightweight, wearable research platform designed for heating aid studies in realistic settings (Kayser et al., 2019). The PHL integrates multiple components, including a single-board computer (BeagleBone Black), battery, and audio interfaces, behind-the-ear hearing aid systems with a receiver in the ear canal (open fit) (Figure 30). During the experiment, participants were wearing the PHL around the neck. The system is capable of time-synchronizing multimodal recordings in real-world environments. All data streams, including EEG and relevant experimental event markers, were synchronized and recorded using the Lab Streaming Layer (LSL) framework (Kothe et al., 2025). LSL enabled real-time integration and timestamping of multiple input sources.

Recordings were stored as unified .xdf files using Lab Recorder software on the PHL device. Control of the recording process was managed via a custom smartphone-based graphical interface, which allowed the experimenter to initiate and monitor recordings in real time. Additional physiological and environmental data (e.g., ECG, acoustic scene recordings) were also collected during the experiment, but are not included in the present study.

9.2.7.2 *EEG*

The participants were fitted with flex-printed electrodes around the ear, called cEEGrids (Debener et al., 2015). The 10 electrodes (AG/AgCI) per cEEGrid are arranged in a C-shape and can be placed around the ear. The skin behind the ear was cleaned with alcohol, and an abrasive electrolyte gel (Abralyt HiCl, EasyCap GmbH, Herrsching, Germany) was applied. The cEEGrids were then placed behind both ears and were connected to a direct current (DC) amplifier (SMARTING Pro, mBrainTrain, Belgrade, Serbia). Data was recorded with a sampling rate of 250Hz. The electrode impedances were kept below 10Ω. The recorded EEG signal was transmitted wirelessly via Bluetooth to the android phone (Samsung S21). The phone was running the Smarting application compatible with the amplifier (v2, mBrainTrain, 2016, Fully Mobile EEG Devices). The recorded data stream was converted into an LSL stream, which was transmitted to the PHL Wi-Fi network.

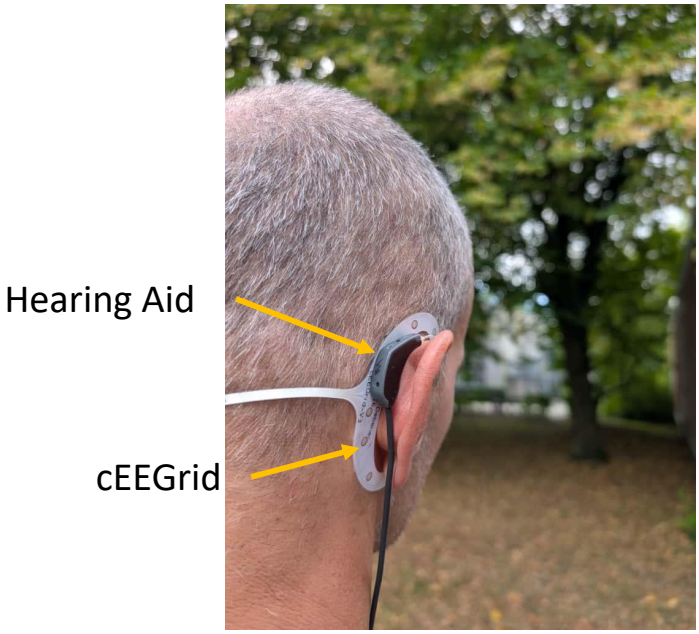

Figure 30: Hearing aid placement over the cEEGrid. The wires (white) of the cEEGrid connect to the amplifier mounted behind the head. The hearing aid cables (black) connect to the PHL worn around the neck.

9.2.8 *Analysis*

9.2.8.1 *Pre-Processing*

Similar to the pre-processing procedure of the cap data in the previous study, cEEGrid data were pre-processed and analyzed offline using MATLAB R2022b (MathWorks, Inc., Natick, USA) and EEGLAB v2024.0 (Delorme and Makeig, 2004). The EEG data were segmented for each story per condition. This resulted in 12 data sets (Lab 1, BTL, Lab 2) x four (four stories presented in each block). The duration of each recording was approximately nine minutes. Subsequently, the data underwent a high-pass filtering process at a cut-off frequency of 1 Hz (FIR, Hamming, filter order 828, pop_eegfiltnew), after which the RMS of each channel was calculated. Channels exceeding three standard deviations above the mean RMS were removed. Given the low channel count posing the risk of poor interpolation quality, bad channels were not replaced by interpolation. Subsequently, the data underwent low-pass filtering at a cut-off frequency of 15 Hz (FIR, Hamming, filter order 222, pop_eegfiltnew) and re-referencing to algebraically linked mastoids (see Debener et al. (2015)). Finally, the data were down-sampled to 128 Hz.

9.2.8.2 *Feature derivation*

In this study, we derived the acoustic envelope for AAD. For this, we used the mTRFenvelope function, which starts by first squaring the waveform to estimate instantaneous power. The signal was then smoothed and resampled using a 1-second moving average window via the mTRFresample function to 128Hz, matching the sampling rate of the EEG. To approximate auditory compression, the square root of the smoothed signal was raised to the power of $\log_{10}(2)$, reducing dynamic range while preserving envelope shape (Lalor and Foxe, 2010).

9.2.9 *Auditory Attention Decoding*

To estimate neural tracking of speech, we applied a linear multivariate decoding model using the TRF toolbox (Crosse et al., 2016), consistent with the analysis approach in the preceding study. In the backward (decoding) model, the goal is to reconstruct the stimulus feature (i.e., the speech envelope) from the neural response by learning a linear mapping from brain activity to stimulus dynamics. Decoder weights were estimated by minimizing the MSE between the actual stimulus, $s(t)$, and the reconstructed stimulus, $\hat{s}(t)$, which is obtained by convolving the lagged neural response with the decoder weights. This can be formulated as:

$$\min \varepsilon(t) = \sum \left[s(t) - \hat{s}(t) \right]^2$$

The optimization process is solved using reverse correlation (Boer and Kuyper, 1968), implemented via ordinary least squares linear regression:

$$w = \left(R^T R + \lambda M \right)^{-1} R^T s$$

Here, R is the lagged neural response matrix, containing time-shifted EEG signals over a defined range of temporal lags, and s is the speech envelope vector. The resulting weight matrix w contains the decoder coefficients for each channel and time lag, reflecting how neural signals at different times contribute to reconstructing the acoustic envelope.

The model was trained to reconstruct the attended speech envelope using EEG data within a temporal window spanning 0 to 500 milliseconds post-stimulus. A 45-millisecond moving window with a 30-millisecond overlap was applied to capture the temporal dynamics of the neural response.

For each subject and condition, the data were segmented into nine non-overlapping 60-second folds. A leave-one-out cross-validation procedure was used: in each iteration, eight folds were used as the training set, and the remaining fold served as the test set. Within the training set, an inner loop cross-validation was performed to determine the optimal regularization parameter (λ). The lambda value that yielded the highest correlation between the reconstructed and actual attended envelope was selected. The model was then trained with this lambda and tested on the held-out fold. This process was repeated until each fold had served as the test set once.

Importantly, the model trained on the attended envelope was used to reconstruct both the attended and the ignored envelope from the held-out test set. The correlation between the reconstructed envelope and the actual attended or ignored envelope was used as a measure of neural tracking strength. This allowed us to quantify how accurately participants' EEG responses aligned with the speech stream they were instructed to attend to, across different listening and movement conditions.

Due to violations of normality in the correlation data, we used the Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test, to compare neural tracking estimates across conditions. Specifically, we used a right-tailed test due to the previous study indicating that decoding accuracy follows the order: single>dual>ignore. Where multiple comparisons were made, results were corrected using the FDR procedure to control for type I error inflation (Benjamini and Yekutieli, 2001).

9.2.9.1 *Auditory Attention Encoding*

In addition to the decoding analysis, we also employed a post hoc encoding analysis. Rather than reconstructing the stimulus from the neural data, we mapped the envelope onto the neural data. This approach, albeit being mass univariate, has the advantage that the resulting model weights from the linear modeling can be interpreted. Here, the same minimization solution is applied; merely the order of the matrices is changed.

$$w = \left(S^T S + \lambda M\right)^{-1} S^T r$$

Here, S is the lagged stimulus matrix, containing time-shifted versions of the envelope feature across a specified window of time lags, and r is the neural response vector. The solution yields a set of model weights for each channel and time lag, reflecting how strongly each stimulus feature at each lag contributes to the observed neural signal.

The derivation of the weights followed the same cross-validation procedure as outlined for the decoding analysis. The only difference, besides the direction of the mapping, was that time lags from -100 to 500ms were used. These lags were chosen to capture the auditory response (Crosse et al., 2021; Haupt, Rosenkranz, and Bleichner, 2024; Winkler, Denham, and Escera, 2013). Again, the attended model was used to predict both the neural response to the attended and ignored envelopes. Nonetheless, we trained a model on the ignored envelope to derive the respective model weights for later comparison to the attended model weights. The predicted neural data were correlated with the actual neural data of the held-out test set. Multiple comparison correction was applied accordingly.

9.3 RESULT

In order to test the reliability and endurance of the setup after the walking condition, we compared the Lab1 and Lab2 reconstruction accuracies for the different auditory conditions. We found no significant difference between Lab1 and Lab2, and thus averaged the results together for the following analyses.

9.3.1 *Decoding*

In the Lab no-noise condition, reconstruction accuracy was significantly higher for envelope of the single speaker condition compared to the attended dual speaker condition ($W = 182, Z = 3.48, p = .003, r = 0.79$). For the Lab dual-speaker condition, attended decoding did not differ significantly from the ignored speech stream. Numerically, however, the mean of each distribution followed the expected trend, where the single speaker reconstruction accuracy ($\mu = 0.04; std = 0.016$) was larger than that of the dual speaker ($\mu = 0.019; std = 0.011$), followed by the ignored stream ($\mu = 0.014; std = 0.006$) (Figure 31A, top). The reconstruction accuracies over time show fluctuations for the single and dual attend model. The maximal decoding accuracy for both models was reached around 200ms. For the ignored model, no discernible fluctuations were observed (Figure 31A, bottom).

In the Lab noise condition, the single speaker's envelope reconstruction accuracy did not differ significantly from the attended dual speaker, but reached near statistical significance ($W = 150, Z = 2.19, p = .062, r = 0.50$). For the noise Lab dual speaker condition showed a significant advantage for attended over ignored speech envelope reconstruction ($W = 173, Z = 3.12, p = .007, r = 0.72$). Numerical mean differences

of the auditory conditions followed the expected trend ($single : \mu = 0.026; std = 0.014; dual : \mu = 0.018; std = 0.007; Ign : \mu = 0.009; std = 0.005$) (Figure 31B, top). The reconstruction accuracy over time for the different auditory conditions showed fluctuations that appeared comparable to the no noise condition. The single and attended dual speaker showed trajectories with increasing reconstruction accuracies that peaked around 240ms. The ignored speaker, however, showed no discernible fluctuations (Figure 31b, bottom).

For the sitting BTL condition, reconstruction accuracy for the single speaker envelope did not differ significantly from that of the attended dual speaker envelope. The attended versus ignored speech in the dual condition was significantly different during sitting ($W = 190, Z = 3.80, p = .002, r = 0.87$). The mean of the reconstruction accuracy distribution followed the expected pattern for the sitting ($single : \mu = 0.032, std = 0.013; dual : \mu = 0.028, std = 0.012; Ign : \mu = 0.015, std = 0.009$) (Figure 31C, top). The reconstruction accuracy over time for the BTL sitting condition shows an early peak around 100ms for all auditory models. Where the ignored models' reconstruction accuracy drops shortly after and fluctuates around zero, the attended and single models' reconstruction accuracy increases further and peaks around 250ms (Figure 31C, bottom).

For the walking BTL condition, envelope reconstruction of the single compared to the dual speaker attend did not differ significantly. The attended speaker envelope reconstruction accuracy differed significantly from that of the ignored speech envelope ($W = 163, Z = 2.72, p = .018, r = 0.62$). The numerical mean reconstruction accuracies in the BTL walking conditions followed the expected trend ($single : \mu = 0.028, std = 0.015; dual : \mu = 0.026, std = 0.011; Ign : \mu = 0.016, std = 0.011$) (Figure 31D, top). The reconstruction accuracy in the BTL walking condition for the attend and single model peaks around 100ms and then drops towards 0. The ignored model shows a very early peak and then steadily declines towards 0 as well (Figure 31D, bottom).

9.3.2 *Encoding*

The previous analysis corroborated previous results that the reconstruction accuracy for the attended stream is greater than that for the ignored stream. The decoding analysis, however, does not lend itself to interpreting the underlying neural response. Furthermore, the reconstruction accuracy of the ignored stream is determined by the attended model. While relevant for practical applications, such as neuro-steered hearing aids, it does not provide insights into the underlying brain mechanisms. Thus, to

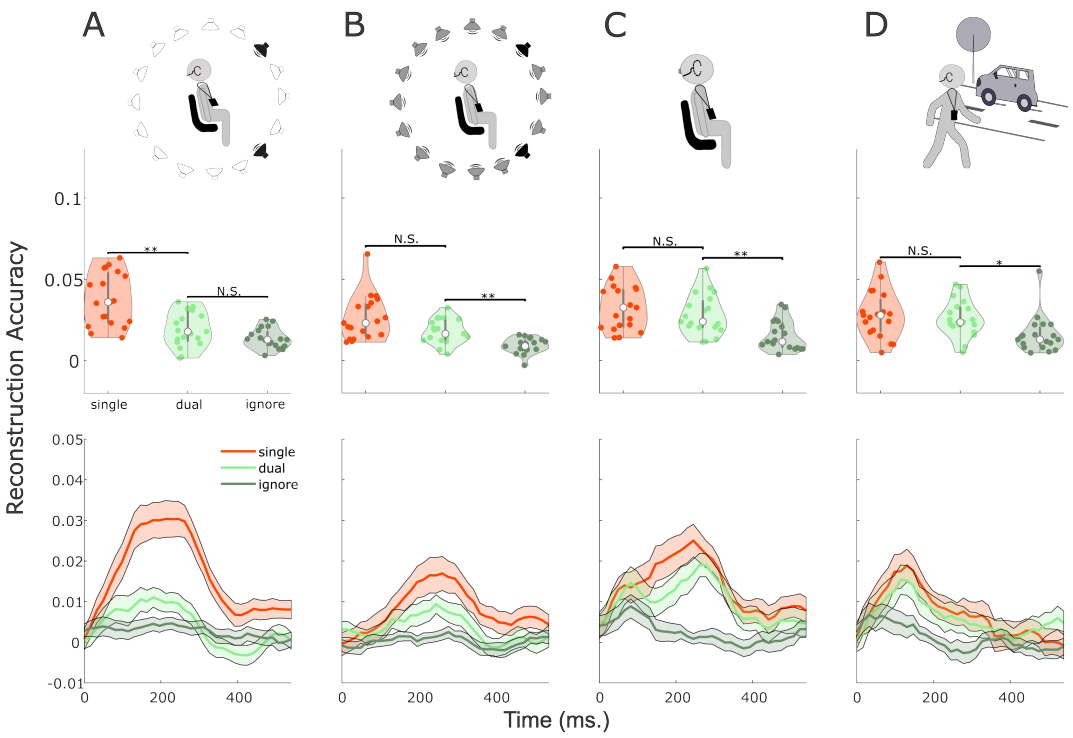

Figure 31: The reconstruction accuracies for the different auditory and movement conditions. The color scheme denotes the auditory condition: orange = single, light green = dual attend, dark green dual ignore. The top panel shows the maximal reconstruction accuracy per participant. The lower panel displays the reconstruction accuracies over time. The bold line is the average over participants, and the shaded area shows the standard error of the mean. **A:** shows the results for the no noise condition. **B:** displays the results for the cafeteria background noise condition. **C,D:**. showcases the reconstruction accuracies for the BTL recordings, sitting and walking, respectively. Statistical comparisons are denoted at N.S. = non-significant, * = $p<0.05$, ** = $p<0.01$, ***$p<0.001$.

gain insights into how the neural response differs between sitting and walking conditions, we decided post hoc to map the envelope onto the neural data. The resulting model weights represent this mapping function and can be interpreted (Diedrichsen and Kriegeskorte, 2017; Holdgraf et al., 2017). Furthermore, we also used these models to predict neural data and compared them to the actual data.

For the forward models' prediction accuracy, one contrast survived the multiple comparison correction. In the Lab no-noise condition, the single speaker prediction accuracy did not differ significantly from the attend dual speaker. The prediction accuracy for the dual condition attended was significantly higher than that of the ignored model. ($W = 166, Z = 2.84, p = .0495, r = 0.65$). However, the mean level prediction accuracy followed the expected trend ($single : \mu = 0.028, std = 0.009; dual :$

$\mu = 0.023, std = 0.008; Ign : \mu = 0.015, std = 0.006$) (Figure 32A, top). When inspecting the response model weights, a clear triphasic pattern for the single-speaker model was observed relative to baseline activity. For the attend and ignore model, this pattern was less obvious. Albeit not as discernible as for the single speaker, the attend dual model also shows a triphasic pattern that is in synch with the single speaker model. In the context of the triphasic pattern of the single model, the ignore model only shows the initial positive peak. We will return to this specific pattern in a later section (Figure 32A, bottom).

In the noise condition, none of the prediction accuracies reached statistical significance after multiple comparison correction. The neural prediction accuracy was numerically higher in the dual condition for attended speech compared to ignored speech, and almost reached statistical significance ($W = 156, Z = 2.42, p = .08, r = 0.56$). Furthermore, the mean level prediction accuracy followed the expected trend ($single : \mu = 0.022, std = 0.011; dual : \mu = 0.018, std = 0.006; Ign : \mu = 0.013, std = 0.007$) (Figure 32B). For the TRF model weight trajectories, we again only observed a trajectory for the single model. The trajectory of the model looked highly similar to that of the no-noise single model (Figure 32B, bottom).

For the sitting BTL condition, none of the contrasts reached statistical significance. The mean of the prediction accuracy distribution did not follow the expected pattern for the single and attend dual distribution for the sitting ($single : \mu = 0.024, std = 0.011; dual : \mu = 0.024, std = 0.012; Ign : \mu = 0.018, std = 0.014$) (Figure 32C, top). The TRF model weights for the sitting condition show oscillatory activity with no distinct pattern similar to those of the stationary laboratory recordings (Figure 32C, bottom).

For the BTL walking condition, no significant differences for the prediction accuracies of the different auditory conditions was found. The numerical mean, also did not follow the expected trend in the walking condition ($single : \mu = 0.027, std = 0.016; dual : \mu = 0.029, std = 0.013; Ign : \mu = 0.02, std = 0.011$) (Figure 32D, top). Regarding the TRF patterns, we saw a very distinct triphasic pattern peaking at 63, 102, and 150ms, respectively, in all of the auditory conditions (Figure 32D, bottom).

9.3.3 *Model Weigths*

Post hoc analysis of the forward model TRF showed a very large response in the walking condition, which is surprising when compared to the stationary conditions. Therefore, we decided to investigate the forward models more closely. Specifically,

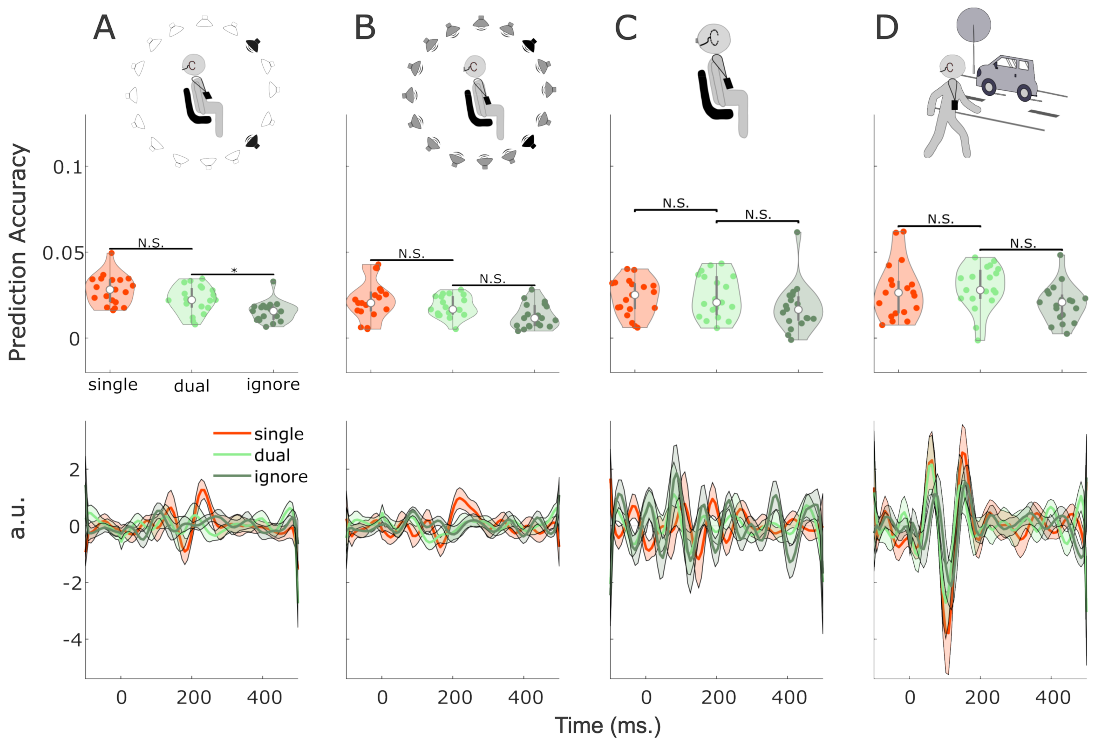

Figure 32: The prediction accuracies and model weights for the different auditory and movement conditions. The color scheme denotes the auditory condition: orange = single, light green = dual attend, dark green dual ignore. The top panel shows the maximal prediction accuracy per participant per condition. The lower panel displays the model weights over time. The bold line is the average over participants, and the shaded area shows the standard error of the mean. **A:** shows the results for the no noise condition. B: displays the results for the cafeteria background noise condition. **C,D:** showcases the output for the BTL recordings, sitting and walking, respectively. Statistical comparisons are denoted at N.S. = non-significant, * = $p<0.05$, ** = $p< 0.01$, ***$p< 0.001$.

we opted to establish a baseline of response models from the controlled laboratory recordings where no experimental noise was present.

The resulting trajectory of the attend speaker model in the no noise condition showed small amplitude fluctuations, characterized by a positive peak at 125ms, followed by a negative deflection around 170ms and a second positive peak at approximately 210ms (Figure 33A).

In contrast, the model weights for the ignored stream exhibited less fluctuation and did not display a triphasic profile. When aligned to the first positive peak of the attended stream, a single positive peak was visible around 100ms. However, no additional deflections followed. Due to the placement of the cEEGrids electrodes, the removed channels, and low SNR, we were unable to generate reliable topographies,

which limited our ability to directly compare these results with those from prior studies using traditional high-density EEG.

For the walking condition, a triphasic trajectory was observed for both the dual-speaker and ignore models. In both cases, the first positive peak occurred around 62 ms, followed by a negative peak at approximately 100 ms and a second positive peak around 150 ms. Compared to the stationary, no-noise lab condition, these model weights had larger amplitudes and showed improved signal-to-noise ratios (Figure 33B). This unexpected result raised several concerns: First, the early peak latencies aligned closely with established auditory response timings, unlike those peaks in the no-noise condition, with a clear peak at the P1 (50 ms) and N1 (100 ms) latency (Winkler, Denham, and Escera, 2013). Second, the improved SNR was unexpected under mobile conditions, where typically more noise can be expected (Mirkovic et al., 2015). Third, a similar triphasic response was evident even in the ignore condition, which is absent in all other conditions. Given these concerns, we considered the possibility that the observed responses in the walking condition may be driven by some kind of artifact rather than neural activity. As detailed below, we therefore examined whether artifacts may have biased prediction performance.

9.3.4 *Artifact Investigation*

Due to the time-locked nature of TRF estimation, any non-neural artifact locked to the speech could directly affect prediction accuracy. To test whether the observed pattern represented such an artifact, we calculated, for each participant, the sum of the absolute model weights and correlated this value with prediction accuracy. In both the single- and dual-speaker conditions, a strong positive correlation ($dual : r = 0.56, p = 0.01; single : r = 0.61, p = 0.001$) emerged between these summed weights and prediction performance (see Figure 34A,B). This suggests that participants exhibiting stronger expressions of the presumed artifactual response also showed higher predictive accuracy.

We also examined decoding performance across conditions. In typical auditory decoding studies, accuracy tends to peak around 200 ms, which was also observed in our sitting condition. However, in the walking condition, the decoding peak shifted to around 100 ms. this is the latency where the suspected artifact appeared prominently in the model weights (Figure 34C). This temporal shift supports our concern that prediction and reconstruction accuracy in the walking condition could have been influenced by non-neural, speech-locked artifacts.

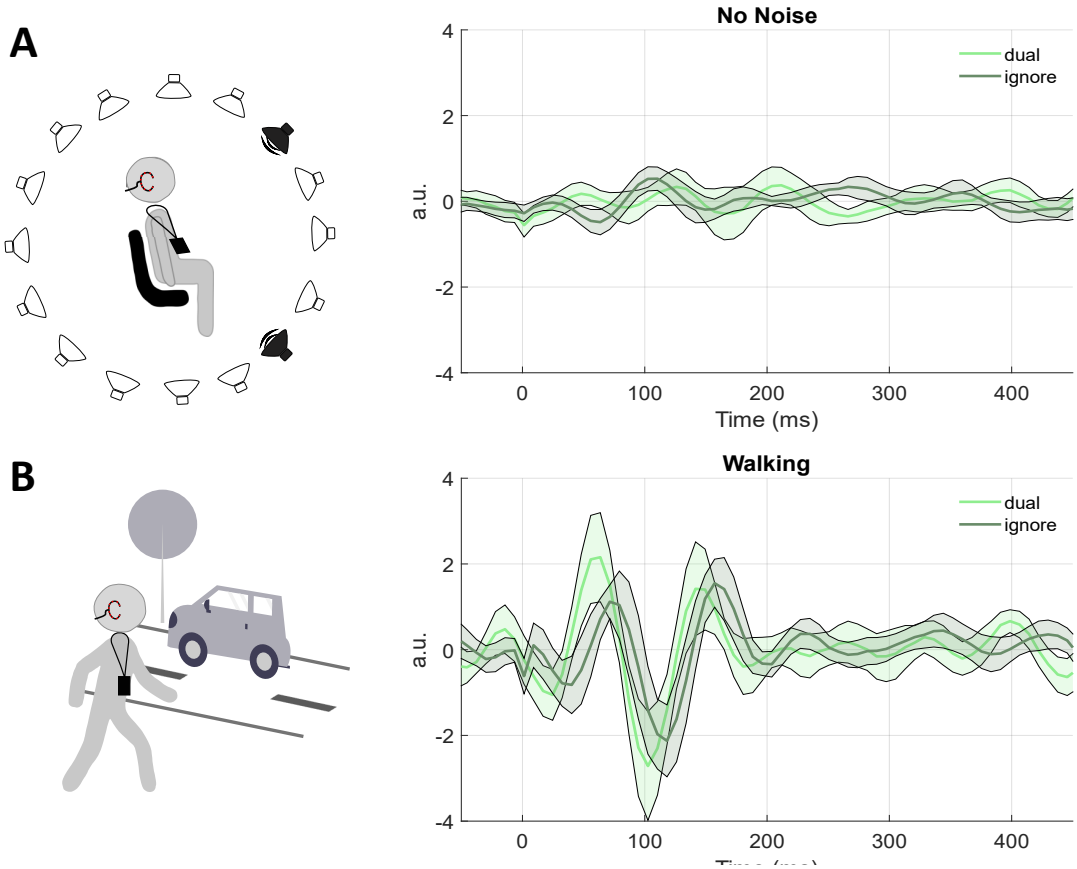

Figure 33: The model weights of the forward modelling for the no noise and walking condition. The color scheme denotes the auditory condition: light green = dual attend model, dark green = dual ignore model. The solid line represents the average, the shaded area the standard error over participants. **A:** shows the trajectory of the no noise condition for the best performing channels, averaged over participants. **B:** The model weights over time lags for the walking condition.

Considering the complex setup of the experiment, including mobile EEG and movement-related noise, we speculated that the artifact may be linked to gait.

If gait artifacts indeed aligned with the speech envelope, we would expect a regressor based on gait event markers to show a similar peak structure. To test this hypothesis, we extracted vertical accelerometer data, which captured individual stepping events, and used it in two ways. We opted for this channel based on previous studies investigating the effects of gait in EEG data (Jacobsen et al., 2022). First, we used it as a regressor on the speech envelope to assess potential alignment. Second, we used it as an additional regressor in the TRF model to investigate whether movement-related signals contributed to the observed pattern. Results revealed that, while the neural model again displayed the triphasic artifact, the gait-only model did not reveal any

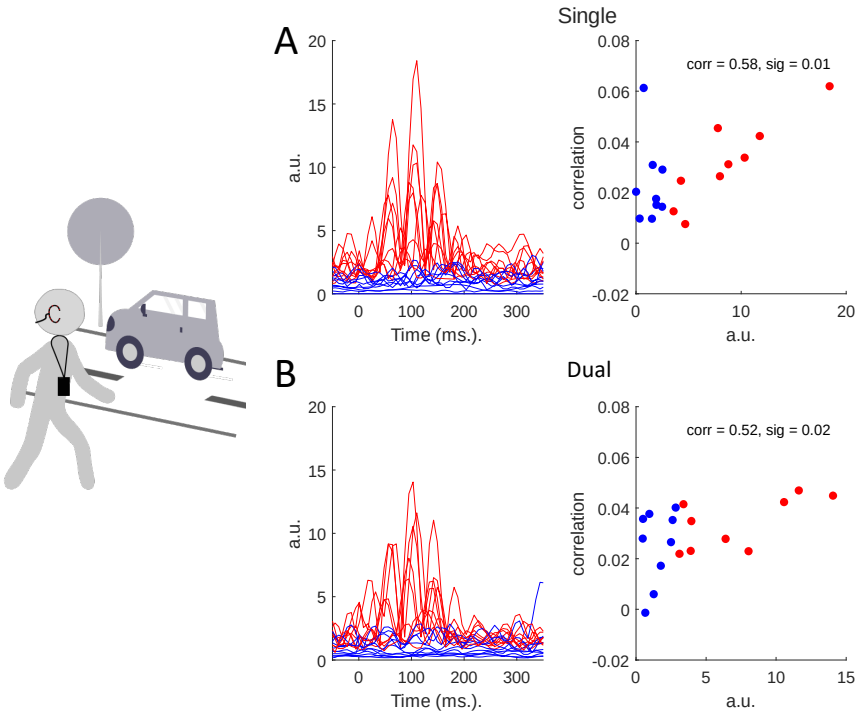

Figure 34: We investigated whether the artifactual response biases prediction accuracy in the walking condition. The left plot shows the absolute model weights for each participant. In the right plot, the y-axis shows the reconstruction accuracy, and the x-axis shows the summed value of the model weights in arbitrary units. The color coding represents the split of participants based on the median of the absolute values over participants: blue = participants lying below the median, red = participants whose values are above the median. **A:** shows the investigation of the artifact's correlation with the prediction accuracies for the single model. **B:** shows the results for the dual attend model.

clear peaks. When both neural and gait data were combined in a single model, this resulted in a small attenuation of the artifact (Figure 35).

9.4 DISCUSSION

This study evaluated the feasibility of performing auditory attention decoding (AAD) using a highly mobile, minimal EEG setup in naturalistic settings. Specifically, we tested whether AAD is feasible during outdoor walking, a common everyday activity, and thus highly relevant for practical applications such as neuro-steered hearing aids. Our initial results indicated that attention decoding appears to be possible while walking. However, closer investigation with follow-up encoding model analysis revealed that these promising results may be driven by an artifact that was not present in stationary recordings. We cannot exclude the possibility that this artifact originates from

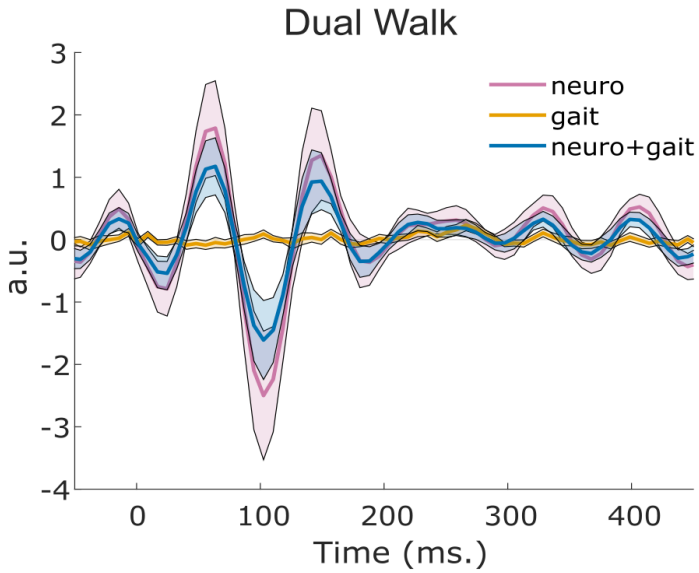

Figure 35: Post-hoc analysis on whether vertical acceleration locks to speech. Shown are the model weights for the model containing only the neuronal data (purple), only the gait data (yellow), and the combined model i.e., neuronal and gait channel data (blue).

the interaction between hearing aid components and cEEGrid electrodes. Although the precise cause of this artifact remains unclear, a useful lesson that can be learned is to treat corroborating results of a singular analysis at face value with care when testing a novel setup.

9.4.1 *AAD using an unobtrusive, mobile setup*

In general, we that that the auditory attention decoding was feasible using a compact, mobile EEG setup. We found the highest reconstruction accuracy for single-speaker conditions, followed by dual-speaker and then ignored-stream conditions. While not all pairwise comparisons reached statistical significance after multiple comparison correction, the mean trend was consistently found across conditions. It is important to note that the reconstruction accuracy of the ignored envelope is based on the attended speech model. The reasoning for using the attend speech model to reconstruct the ignored envelope is that the model is assumed to represent attention-based pro-

cessing of sound streams. Hence, lower reconstruction accuracy is indicative that this particular auditory stream is not being attended to. The increased reconstruction accuracy of the attended compared to the unattended stream likely reflects a neural signature of an active filtering mechanism. Whether this means suppression of the unattended or enhancement of the attended stream remains to be shown (Fiedler et al., 2019; Hausfeld et al., 2021; Obleser and Kayser, 2019).

Besides the increased reconstruction accuracy of the attended stream, we did observe an effect of noise, where the reconstruction accuracy is generally lower when noise is present. This reduction can be explained by the additional load to process the increased complexity of the sound streams. One theory to explain this is load theory (Murphy, Spence, and Dalton, 2017; Rudner, Rönnberg, and Lunner, 2011). When presented with additional streams of information, fewer neuronal resources are available to process the attended speaker stream in the context of noise sources, which results in a reduction in reconstruction accuracy (Muncke, Kuruvila, and Hoppe, 2022; Vanthornhout, Decruy, and Francart, 2019). This effect is probably of relevance, since unattended channels still receive semantic processing (e.g. Holtze et al., 2021). However, it has to be noted that this effect seems to be nonlinear, as a minimal noise floor has been shown to have the reverse effect, where the neural tracking of the attended speaker increases due to stochastic resonance (Herrmann, 2024). In our case, the background noise of the cafeteria posed sufficiently challenging conditions to reduce the reconstruction accuracy.

The experimental design directly replicates partially the results of the same study conducted with a conventional 32-channel EEG cap system, allowing for a meaningful comparison (Straetmans, Adiloglu, and Debener, 2024). Despite the substantially reduced spatial coverage of the cEEGrids, we were able to reproduce the general decoding trends observed in cap-based data (single>dual>ignore). However, as expected, there were notable differences compared to the cap results. In contrast to the results reported by Straetmans, Adiloglu, and Debener (2024), we observed generally lower reconstruction accuracies, smaller effect sizes, and reduced statistical power. These differences are probably attributable to the minimal EEG setup around the ear (Fuglsang, Dau, and Hjortkjær, 2017; Geirnaert, Kappel, and Kidmose, 2025). Prior work by Mirkovic et al. (2015) demonstrated that reconstruction accuracy scales with the number of EEG channels, identifying a minimum of 25 spatially separated electrodes as necessary for achieving stable decoding performance. Using fewer channels or less spatial coverage resulted in a systematic decline in accuracy. This is due to the multivariate nature of the decoding approach, which benefits from higher spatial coverage, as more information is available (Crosse et al., 2016, 2021; Holdgraf et al., 2017).

Our findings are consistent with that observation. Nevertheless, the fact that decoding remains possible with such our minimal setup is encouraging. Interestingly, that also appeared to be the case for the walking condition, which was expected to have the lowest SNR. While promising at first glance, the cEEGrid results in the walking condition seem to be biased, as will be discussed later on. Nonetheless, for the stationary conditions, it suggests that meaningful neural tracking of auditory attention can be achieved without full cap coverage, which is particularly valuable for real-world applications. High-density EEG systems are often impractical for daily use due to long setup times, discomfort, and restrictions on user mobility. In contrast, flex-printed EEG electrodes offer a lightweight, unobtrusive, and potentially self-applicable (Da Silva Souto et al., 2022) alternative that may facilitate more accessible and user-friendly neurotechnology.

Taken together, these findings highlight both the promise and pitfalls of ear-centered mobile EEG for AAD. Replicating general mean-level trends of decoding validates that we were able to measure meaningful information using a highly mobile, minimal setup. While performance may not yet match that of high-density laboratory systems or the full-cap setup in mobile situations, the ability to replicate key effects with a minimal setup represents a significant step toward practical deployment in everyday listening environments. While we also found the expected trend of the reconstruction accuracy for the mobile walking condition, a post hoc encoding analysis revealed a confounding artifactual response, calling for a cautious interpretation of the results.

9.4.2 *Forward Modeling*

In contrast to decoding approaches, forward modeling offers an interpretable framework by mapping stimulus features, such as the speech envelope, onto the neural response while accounting for the temporal delays introduced by auditory processing (Crosse et al., 2021; Diedrichsen and Kriegeskorte, 2017; Fiedler et al., 2019; Hamilton et al., 2021; Haupt, Rosenkranz, and Bleichner, 2024). TRFs offer insights into the time course of auditory processing and the relative contribution of individual stimulus features, as reflected in the model weights and prediction accuracy (Brodbeck, Hong, and Simon, 2018; Holdgraf et al., 2017).

In our study, prediction accuracy from the forward models closely mirrored the trends observed in decoding, that is, highest accuracy for the single speaker, followed by the dual-speaker attend and then the ignored model. However, overall prediction performance was lower than that of the decoding. This is to be expected as encoding

models treat channels as univariate variables. In this case, information unrelated to the stimulus is not regressed out, as is the case in the multivariate decoding approach. Here, cross-channel information is leveraged implicitly, and redundant information is removed in the weight optimization process (Diedrichsen, 2020; Haufe et al., 2014; Holdgraf et al., 2017). This distinction of approaches leads to reduced performance, both in terms of accuracy and effect size, for the encoding compared to the decoding approach (Wong et al., 2018). Nevertheless, the similarity in trends across both frameworks (encoding and decoding) suggests that the encoding models captured core aspects of the attentional modulation present in the neural signal, justifying the interpretation of TRFs in relation to decoding outcomes.

9.4.3 *Interpretation of Model Weights*

To determine whether cEEGrids captured physiologically plausible auditory responses, we first examined TRFs from the no-noise laboratory condition. Here, the model for the attended stream exhibited a very weak, triphasic pattern. For the attended stream, the model weights displayed a weak triphasic trajectory, characterized by peaks resembling canonical auditory components: P1 (~ 50 ms), N1 (~ 100 ms), P2 (~200 ms) (Winkler, Denham, and Escera, 2013). These components are thought to reflect successive stages of auditory processing, associated with early sensory encoding (P1) (Puvvada and Simon, 2017), attentional selection (N1) (Ding and Simon, 2012a; Thornton, Harmer, and Lavoie, 2007), and higher-order integration (P2) (Brodbeck, Hong, and Simon, 2018; Di Liberto, O'Sullivan, and Lalor, 2015).

In contrast, the ignored stream showed no clear triphasic pattern, but shared the initial P1 peak, a finding consistent with Jaeger et al. (2020), who reported that only the first component was present for unattended stimuli. While these observations align qualitatively with prior work (Fiedler et al., 2019; Fuglsang, Dau, and Hjortkjær, 2017), our data do not fully replicate the expected timing and spacing of peaks. Specifically, the latency of P1 and the N1–P2 interval deviated from established values. While we cannot exclude that some latency variation may have resulted from hardware-related timing issues, the magnitude and structure of the observed deviation suggest that these differences cannot be fully attributed to noise. Thus, we cannot conclusively establish that the derived response in the no-noise condition reflects neural processing consistent with the literature.

This discrepancy may stem from the reduced spatial coverage and SNR of cEEGrids compared to full-cap EEG setups. The close proximity of cEEGrid electrodes limits

the sensitivity to dipolar fields from distant sources, resulting in attenuated amplitudes (Meiser et al., 2020). Furthermore, the lack of topographic maps precludes a comparison of the current weight distributions with prior cap-based studies to assess whether a similar spatial pattern occurs. In conclusion, although the prediction accuracies not only match the decoding results but are also consistent with other studies, the current weight trajectories do not allow us to conclude with certainty that the derived response models reflect the auditory response in line with previous results. While this is the case for the forward approach, the results of the decoding for the stationary auditory conditions suggest that recording devices with low SNR benefit from the multivariate decoding approach. Therefore, the forward models may not have reflected the auditory response to speech stream processing, but the backward models highlight the cEEGrids capacity to measure neural responses. Note that alternative ear-EEG systems, which position electrodes in the outer ear canal and the concha, cover much less space than the cEEGrids and therefore suffer from the same problem, likely to a larger extent (Geirnaert, Kappel, and Kidmose, 2025).

Given the increased noise that can be expected in mobile recording conditions, we had anticipated noisier and less structured trajectories of the model weights for this condition (Straetmans et al., 2022). Contrary to expectations, the TRFs derived from the walking condition exhibited very pronounced triphasic patterns in both attended and ignored streams. Three distinct properties led us to conclude that, rather than representing a neural response, the derived weights are the result of an artifact. The first being the latency shift of the triphasic peaks compared to the other conditions. It is unlikely that this effect was stimulus-driven, as we used identical envelope features across all conditions. Furthermore, potential differences in hardware-related trigger timing delay were accounted for by timing tests. Therefore, this latency shift cannot be attributed to stimulus-related factors. The second aspect is the susceptibility of EEG recordings to artifacts in mobile conditions, that is, when participants move during data acquisition. Thus, one would expect the SNR to be lower in mobile conditions. The fact that we observed stronger, more defined TRFs in the walking condition than in the seated lab condition is inconsistent with our previous studies comparing EEG signals from mobile to stationary recording conditions (e.g. Scanlon et al. (2021)). Finally, the TRFs for attended and ignored streams were nearly identical in the walking condition. This is atypical, as attention-related modulation is a robust finding in auditory tracking studies (Fiedler et al., 2019; Jaeger et al., 2020).

Together, these findings suggest that the TRFs in the walking condition reflect artifactual activity that is temporally locked to the speech envelope.

The artifact interpretation raises significant concerns regarding the interpretability and reliability of the results. First, prediction accuracy was strongly correlated with model weight amplitude, indicating that the artifact inflated predictive performance. Second, reconstruction accuracy peaked around 100 ms, exactly where the artifactual component emerged. This deviates from the typical 200 ms decoding peak reported in most auditory attention studies (Jaeger et al., 2020; Mirkovic et al., 2015; O'Sullivan et al., 2015; Straetmans, Adiloglu, and Debener, 2024). Since the artifact was confined to the walking condition and time-locked to the stimulus, we considered whether it could have a movement-related origin. To test this, we examined correlations between the speech envelope and vertical acceleration signals. However, no consistent mapping was observed, making it unlikely that gait-related artifact causes this effect.

An alternative interpretation is that hardware interference may have occurred. In some individuals' cases, the hearing aid cable or housing may have come in direct physical contact with cEEGrid electrodes, as both were positioned behind the ears. Unfortunately, exact documentation of how devices were positioned on individual participants was unavailable, but for some participants the artefact was present, while it was absent for others. This limits our ability to isolate the issue definitively. Nevertheless, the findings underscore the need for future studies to systematically investigate hardware interactions. This is particularly important, since the benefit of a direct integration of ear-EEG and hearing aid technologies requires prior investigation with separate research systems, such as cEEGrid and PHL.

Importantly, the artifact issue was only detectable through the forward model approach. This highlights the value of interpretable models in translational research, mostly concerned with practical applications. Investigating only decoding results that corroborate expectations at face value may lead to complications in applied settings, such as neuro-steered hearing aids, where robust signal interpretation is essential. Despite this limitation, reliable performance was observed with the cEEGrids in the seated and dual-speaker conditions. These results support their potential as a lightweight, mobile EEG solution. Moving forward, more thorough hardware validation will be necessary to ensure the integrity of neural recordings in mobile recording conditions.

9.4.4 *Relevance of the dual speaker paradigm*

While our study contributes to the growing body of work on auditory attention decoding (Ciccarelli et al., 2019; Geirnaert, Kappel, and Kidmose, 2025; Jaeger et al., 2020;

Mirkovic et al., 2015, 2016; Straetmans et al., 2022; Thornton, Mandic, and Reichenbach, 2024), several limitations of the dual-speaker paradigm deserve broader consideration. Importantly, the concerns discussed below are not unique to the present study but represent general methodological challenges that affect the ecological and practical validity of AAD research more widely.

Although auditory attention decoding offers a compelling approach to interpreting selective attention in neural data, the ecological validity of the dual-speaker paradigm warrants critical evaluation. In everyday communication, individuals rarely must sustain attention on one of two concurrently active speakers over long durations (Ryck et al., 2025). While our study incorporated more realistic elements, such as background cafeteria noise, mobile conditions, and an unobtrusive recording setup, there remains a disconnect between experimental designs and real-world communicative scenarios.

Another key limitation lies in the temporal resolution of current decoding approaches. Peak decoding performance typically occurs around 200 ms post-stimulus, introducing a lower bound for response latency in neuroadaptive systems. Furthermore, achieving reliable classification often requires at least several seconds of data; most studies report above-chance decoding only when averaging across windows of at least five seconds or more (Fuglsang, Dau, and Hjortkjær, 2017; Mirkovic et al., 2019). This temporal lag complicates real-time responsiveness, particularly during attentional shifts.

Whether users can adapt to such delays, or how switching costs (e.g., changing focus from one speaker to another) interact with system responsiveness, remains to be tested. An online experiment conducted by Hjortkjær et al. (2024) reported participants describing misclassification as disturbing. These findings highlight unresolved challenges regarding switching costs, responsiveness, and the trade-off between speed and accuracy in closed-loop systems. Designing neuro-steered applications that can accommodate these constraints, without overwhelming the user, remains a key challenge for the field.

9.4.5 *Conclusion*

This study is the first to investigate AAD using a mobile, minimal ear-EEG configuration paired with hearing aid research hardware. The results suggest that such configurations are feasible to record neural responses to AAD in stable conditions, where movement is minimal. For the actual success of decoding attention in realistic walking settings, we found mixed results. While decoding trends observed in

the stationary conditions align with expectations, we identified artifacts in the walking condition that biased model weights and prediction accuracies. Artifacts were likely introduced through interactions between hardware components. Despite this limitation, our study highlights the potential of forward modeling as an explainable and diagnostic framework, capable of revealing issues that may be masked in purely performance-driven decoding approaches. These insights are particularly relevant for application-focused research, such as neuro-steered hearing aids, where model interpretability and signal reliability are crucial. Future studies should pursue fully online implementations with optimized minimal setups. Identifying the optimal balance between decoding speed, comfort, and interpretability will be essential for real-world applications. Taken together, these results represent a meaningful step forward in bringing brain-based technologies closer to real-world, practical application, whether in the form of hearing aids, assistive technologies, or broader mobile neuroimaging systems.

Funding Information:

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286 and by the German Federal Ministry of Education and Research (BMBF, 16SV8596). Also, it was funded by the DFG Emmy Noether ID: 490839860.

Acknowledgements

We would like to thank Jennifer Decker for her help during data collection and Hendrik Kayser and Paul Maanen for their assistance during software development.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Potential Declarations

During the preparation of this work, the author(s) used ChatGPT 4o and the free version of ChatGPT (mid 2024) in order to improve language and readability of selected sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author Contributions

Conceptualization, S.D., T.H., K.A., M.B., and L.S.; Methodology, L.S., K.A.; Software, K.A.; Formal Analysis, T.H.; Investigation, T.H.; Resources, S.D.; Data Curation, L.S., K.A; Writing – Original Draft Preparation, T.H.; Writing – Review & Editing, T.H., L.S., M.B., K.A., and S.D.; Visualization, T.H.; Supervision, M.B., S.D. and L.S.; Project Administration, S.D. ; Funding Acquisition, S.D.

Part III

DISCUSSION

This is the final part of thesis and discusses the findings of my empirical work in a broader context.

10

SUMMARY OF FINDINGS

"They both savored the strange warm glow
of being much more ignorant than ordinary people,
who were ignorant of only ordinary things."

Pratchett (2009)

I set out in this thesis to advance the understanding of naturalistic soundscape perception. This was driven by the motivation to understand the neural underpinnings of everyday life cognition and behavior. Before taking this research beyond the laboratory, it was essential to identify and evaluate the key factors driving soundscape perception under controlled conditions. This approach provided the foundation for future work aimed at capturing auditory processes in real-world environments.

The discussion begins by summarizing the key findings of the empirical studies (Figure 36) and situating them within the broader framework of soundscape perception. I then contextualize how these findings advance our understanding of both the proximal and perceptual soundscapes (Chapter 11), and evaluate the suitability of the hardware used for Beyond the Lab (BTL) measurements (Chapter 12). Finally, I reflect on three central conceptual questions: whether extending experimentation beyond the laboratory resolves the lab-dilemma, whether we truly measure information processing in the brain, and whether the models derived here provide meaningful neural interpretations (Chapter 13). The chapter concludes with an outlook on how these findings may inform future research and methodological development (Chapter 14).

The first study (Chapter 7) established the empirical foundation for analyzing the neural underpinnings of naturalistic soundscape perception. Specifically, it tested how different acoustic and meta-information features, and their corresponding Temporal Response Functions (TRF)s, explain variability in EEG responses (Figure 36A, top). Acoustic Features (AC) captured physical aspects of the proximal soundscape (e. g., onsets, envelope, and mel-spectrogram). This study systematically compared models

based on different levels of acoustic detail to determine how soundscape features contribute to explaining neural variability. More detailed AC, such as mel-spectrograms, improved model performance compared to simpler descriptors like acoustic onsets (Figure 36A, bottom). Interestingly, of the variance explained by the more complex models, a large proportion was captured by acoustic onsets. This suggests that simpler models account for much of the activity measured, and more complex models only marginally improved the variance explained. Besides the AC, meta-information reflected perceptual soundscape cues, such as Sound Identity (SI) and Cognitive Priors (CP). Adding meta-information features, particularly SI, further enhanced model accuracy. This demonstrates that perceptual soundscape information improves the prediction of neural activity beyond purely acoustic cues. Overall, these findings suggest that simple acoustic transients are a key driving factor of the neural representation of naturalistic soundscape perception captured by Electroencephalography (EEG). Higher-level proximal and perceptual information refines the representation when available. This underscores the importance of combining proximal and perceptual features in future models of everyday auditory perception.

Building on the first study, which focused on acoustic features, the second study (Chapter 8) investigated how temporal dynamics contribute to explaining neural responses during naturalistic soundscape perception. Because real-world auditory scenes are defined not only by what sounds occur but also by when they occur, temporal context was modeled through the Inter Onset Interval (IOI) (Figure 36B, top). Here, we grouped onsets according to their temporal proximity. The analysis revealed that the N1 and P2 peaks of the TRFs attenuated nonlinearly as a function of onset proximity: shorter inter-onset intervals led to stronger adaptation than longer intervals (Figure 36B, bottom). This nonlinearity indicates that auditory cortical responses dynamically adjust to temporal density, reflecting context-dependent adaptation mechanisms. These findings extend laboratory results on tone adaptation to complex, naturalistic soundscapes, demonstrating that adaptation to temporal context can be measured by EEG. Importantly, they reveal that neural adaptation introduces nonlinearities into TRF-based analyses, challenging the assumption of linearity. Accounting for these temporal effects thus enhances model accuracy and enables the characterization of dynamic auditory processing directly from the proximal soundscape.

The final study (Chapter 9) addressed a critical step in advancing naturalistic soundscape research: testing whether minimal and mobile EEG hardware can reliably capture neural activity beyond the lab. Using an existing dataset, participants performed an Auditory Attention Decoding (AAD) task across two environmental scenarios: one in a controlled laboratory setting (seated) and one in naturalistic environments (seated

and walking) (Figure 36C, top). This allowed a direct assessment of the measurement system's integrity under movement compared to seated controlled laboratory settings. The results showed that AAD performance remained stable across the two laboratory sessions, confirming the system's temporal reliability. In the seated BTL condition, decoding accuracy matched that of the laboratory recordings, indicating that minimal hardware can record neural data in stationary natural environments sufficiently. However, during walking, encoding analyses revealed an apparently artifactual response that biased both decoding and model estimation (Figure 36C, bottom). Although the exact source of this signal remains unclear, it underscores the need for careful artifact characterization when extending EEG recordings to mobile conditions, where additional uncontrollable factors impact the measurement. Importantly, the application of explainable encoding models enabled the identification of this artifact, demonstrating how physiologically interpretable modeling can safeguard the validity of BTL neuroscience from erroneous conclusions. These findings not only highlight the potential of mobile EEG systems for real-world applications but also identify the critical role of transparent, explainable models in ensuring the interpretability and reliability of neural data collected BTL.

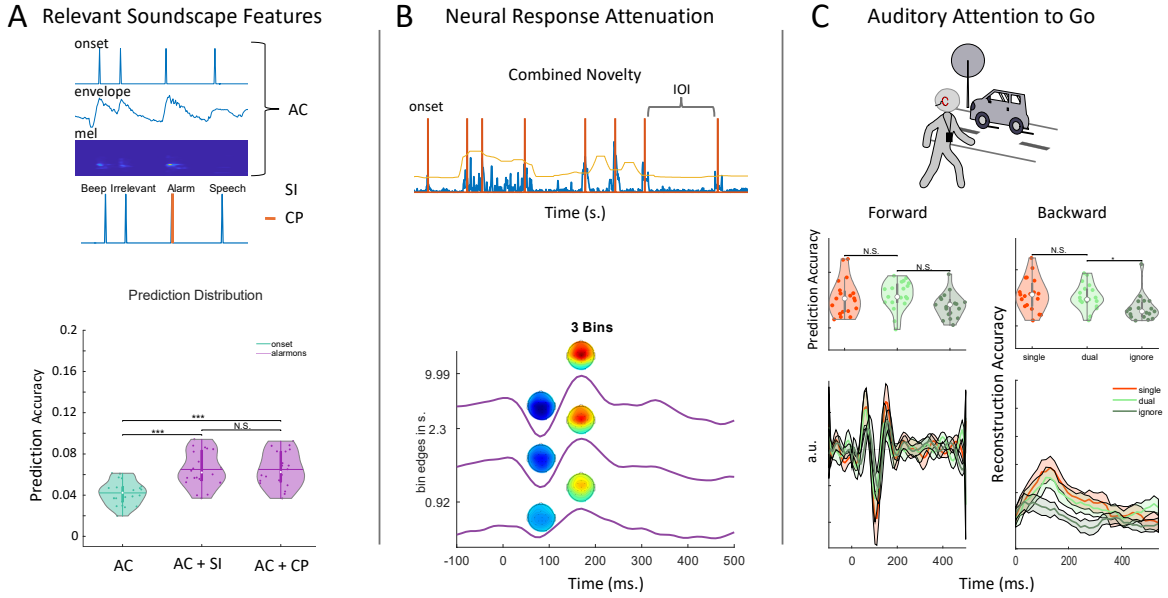

Figure 36: Provides an overview of the key findings of the three studies presented in this thesis. Displayed are results from study one (Chapter 7), two (Chapter 8), and three (Chapter 9) from left to right, respectively. For each section, the top part shows the methods used, while the bottom part shows the key results. * p< 0.05, **p< 0.01 ***p<0.000, N.S.=non-significant. **A:** The first study investigated different features in terms of the neural variability that the respective TRF models can explain. Here we distinguish between AC, SI, and CP. **B:** The second study explored the impact of temporal context, modeled by IOI, on neural response amplitude. **C:** The last study investigated AAD using a mobile and minimal setup in laboratory and BTL settings.

11

NATURALISTIC SOUNDSCAPE PERCEPTION

"Being adjacent to that much beauty—
more than adjacent; immersed in, pierced by it—
was the point. The physical risks were footnotes."

Finnegan (2015)

Key Takeaways

- The proximal soundscape benefits from additional descriptive detail.
- Future studies may consider using bio-informed priors or deep neural networks to enhance the description of the proximal soundscape.
- The perceptual soundscape can merely be approximated, but cognitive prior information enhances the model estimation.
- Internal models (alpha fluctuations) and additional sensors besides EEG (i. e.pupillometry) may be used in the future to better contextualize the perceptual state.

What you will learn:
In how far this thesis has managed to depict the proximal and perceptual soundscape, whether it is relevant, as well as other ways to improve the description further.

A central theme in this thesis is the distinction between the proximal and perceptual soundscape, two complementary perspectives essential for understanding how humans experience complex auditory environments. The proximal soundscape refers to the measurable acoustic environment, as captured, for instance, by a microphone, whereas the perceptual soundscape represents how this environment is subjectively experienced by the listener. Within the context of investigating naturalistic soundscape perception using EEG, this distinction is crucial. The proximal soundscape offers a quantifiable, physical signal, while the perceptual soundscape can only be

inferred indirectly through neural and behavioral measures. Understanding the interplay between these two soundscapes is fundamental for linking physical sound environments to neural and perceptual processes. This section, therefore, evaluates to what extent the work presented in this thesis succeeded in operationalizing both perspectives and identifying neural markers that bridge them. Furthermore, I explore potential approaches to improve their characterization.

11.1 PROXIMAL SOUNDSCAPE

To analyze neural activity underlying soundscape perception, one needs to determine relevant features to guide model derivation. As has been shown, the type of feature not only determines the model's ability to explain neural variability but also determines which aspects of neural processing can be explained. In laboratory settings, often artificial soundscapes are presented, consisting of click trains. Here, careful variation along one dimension, such as IOI, isolates the effect of interest, and a full description of the proximal soundscape is possible. Naturalistic soundscapes, in contrast, can be complex, consisting of polyphonic sounds, each with its own temporal dynamic. Obtaining the proximal soundscape thus requires careful description of relevant aspects. This study series approached this challenge by progressively testing how different feature types, acoustic, semantic, and temporal, capture neural variability.

In light of the results presented throughout this thesis thus far, a question that emerges is whether we have determined how the proximal soundscape is best described. The short answer is: no. Having investigated different acoustic features in Chapter 7 and the impact of temporal dynamics in Chapter 8, we have showcased the diversity and complexity of aspects that need to be considered, rather than determining the exact features to be derived. Therefore, a much more nuanced question would be to ask whether the thesis provides insights into how the proximal soundscape can be investigated. To this, I would argue yes. Specifically, I highlighted how the processing of distinct acoustic aspects can be analyzed and contrasted in naturalistic soundscapes (why features explored in isolation do not provide meaningful insights will be discussed in Section 13.2). For instance, a key finding was that acoustic onsets/ transients make up much of the variability explained by the more detailed envelope. Here, the acoustic features derived share much of the information represented in the signal captured by EEG. Additionally, the second study showcased the versatility of simple acoustic features in depicting the temporal context, which modulates the neural response. These results not only highlight the different ways the

proximal soundscape can be characterized, but also the informational wealth that can be extracted beyond the characteristics the features depict.

It is important to note, however, that this thesis was not the first to derive the ACs. Specifically, the AC explored in this thesis, acoustic onsets, envelope, mel-spectrogram, and IOI, have been investigated extensively in earlier work (Brodbeck, Presacco, and Simon, 2018; Deoisres et al., 2023; Desai et al., 2021; Heer et al., 2017; Oganian et al., 2023). While the results in this thesis corroborate previous findings, where acoustic detail improves the amount of neural variability explained, they expand these to soundscapes that besides speech, also include non-speech sounds. Such generalizability is essential, as everyday life soundscapes contain several other sources of sounds besides speech. This underscores the ability of EEG to capture general acoustic processing of the auditory cortex. For BTL recordings, these results indicate that simple acoustic models are a promising feature to capture auditory processing and adaptation.

Returning to the initial goal of determining which aspects of the soundscape are relevant to capture, one has to acknowledge that soundscapes recorded BTL can be considerably more complex. While the preceding section demonstrated that simple acoustic models capture core aspects of cortical auditory processing, extending these findings to real-world, BTL conditions requires refining how the proximal soundscape is described. Thus, the question arises whether additional methods are required to improve the proximal soundscape description. This can be addressed in two ways: 1. bio-informed features 2. improved soundscape description.

The bio-informed approach draws on the analogy of the TRF as a window into neural processing, where information at a distinct spatial location in time is measured. For example, the mel-spectrogram reflects frequency selectivity of human hearing psychoacoustically and thus aligns well with EEG recordings. This can be advanced via bio-informed models of subcortical processing, given that these are the stages prior to cortical acoustic processing. For instance, models of IC activity have been used as input features for TRF estimation, achieving improved variance explained compared to acoustic models alone. Optimal performance, however, was obtained when both acoustic and subcortical features were combined (Lindboom et al., 2023). Similarly, subcortical models have been used to explain EEG differences between music and speech perception (Shan, Cappelloni, and Maddox, 2024). These findings suggest that biologically inspired features, incorporating lower-level auditory transformations, may enhance EEG analyses of complex soundscapes. Nonetheless, invasive work in animal models indicates that simple acoustic features can yield comparable predictive

power (Rahman et al., 2020), suggesting that the added benefit of bio-informed models may depend on the recording modality and experimental context.

A second approach to improving proximal soundscape description involves the use of DNN. Unlike bio-informed models, which rely on assumptions about auditory physiology, DNNs learn perceptual-level representations directly from labeled sound data. By classifying sound sources and events, these models derive semantic labels that provide contextual information about the likely composition of a soundscape. Such labels can be combined with acoustic features to enhance neural response modeling, as demonstrated with the SI in Chapter 7. One viable solution is the YAMNET, which was trained to recognize 521 audio event classes from the AudioSet corpus (Gemmeke et al., 2017; Hershey et al., 2017). The model can be implemented online and applied to label the proximal soundscape, thereby providing immediate contextual information about environmental sounds (Haupt et al., 2025). Beyond classification outputs, intermediate layers within DNNs, where information becomes compressed into abstract, high-density representations, also constitute valuable features for model derivation. These layers often capture the latent structure in the soundscape that parallels perceptual organization, where early layers correlate with acoustic properties and later layers with order concepts, such as saliency (Huang, Slaney, and Elhilali, 2018). Likewise, Tuckute et al. (2023) found that intermediate layers of audio networks correlate best with primary auditory cortex activity, whereas deeper layers align with non-primary cortical responses. Thus, these inner layers may provide a more detailed description of the soundscape in a higher-dimensional feature space to improve neural model estimation. However, practical constraints of using DNN for soundscape characterization remain: the lack of large annotated training datasets (Mesaros, Heittola, and Virtanen, 2016), the quality of the labeling (Song), poor generalization to unseen contexts (Stowell et al., 2019), and difficulties in representing polyphonic mixtures (Mesaros, Heittola, and Virtanen, 2016; Nourifard, 2025) all limit current applicability in BTL settings. Moreover, as feature dimensionality increases, so does the required training data (see Figure 27 for visualization of this effect). Thus, DNN-based features should be evaluated in controlled conditions before they are applied to complex, real-world soundscapes.

An interesting case for rather simple yet powerful acoustic models can be made with a recent study published by our group. Korte, Haupt, and Bleichner (2025) showed that the neural responses to acoustic onsets, binned by their auditory novelty, were strongest for highly novel sounds. Interestingly, the responses showed comparable amplitudes across active and passive listening, suggesting that novelty processing occurs automatically. While sound novelty has previously been associated with large

neural responses in controlled paradigms (Debener et al., 2005; Escera and Malmierca, 2014; Garrido et al., 2009; Rosburg, Weigl, and Mager, 2022), our findings extend this effect to continuous, naturalistic soundscapes recorded over multiple hours. This suggests that novelty, as derived from the proximal soundscape, represents a promising feature for BTL studies when detailed acoustic annotation is not feasible. While simple acoustic models provide a powerful tool to investigate auditory perception beyond the lab, a necessary level of description detail is required. For example, self-generated sounds may be acoustically novel yet elicit reduced neural responses due to predictive coding mechanisms (Näätänen, 1992; Sanmiguel, Todd, and Schröger, 2013; Timm et al., 2013). In this case, a simplistic novelty model may be able to explain neural variability, given that they are contextualized, i. e., by self-produced onset markers.

The optimal description of the proximal soundscape remains an ongoing challenge due to the inherent complexity of naturalistic environments. While a more detailed description is generally desirable, this thesis demonstrates that simple acoustic models can often capture the essential dynamics of natural soundscape perception, measured by EEG. Thus far, we have only discussed the proximal soundscape; the findings regarding the perceptual soundscape will be discussed next.

11.2 PERCEPTUAL SOUNDSCAPE

The perceptual soundscape reflects the subjective, qualitative experience of an individual perceiving the proximal soundscape. Although direct access to qualia of perception is impossible (Kanai and Tsuchiya, 2012), indirect measures such as neural activity provide a means to approximate the processes that underlie the subjective experience. Describing the perceptual soundscape is considerably more complex than characterizing the proximal one, as cognitive states that shape perception are not directly observable (Gruijters, 2022; Nastase, Goldstein, and Hasson, 2020). Experimental manipulations, therefore, serve as a practical approach to elicit specific perceptual states, allowing researchers to infer aspects of subjective experience indirectly.

In this thesis, I examined how such perceptual markers can inform neural modeling under controlled experimental conditions. Specifically, in Chapter 7, we included SI and CP as additional features to test whether perceptual context improves model performance. While CP did not lead to a significant improvement, markers of sound identity did, suggesting that auditory object information provides meaningful information for model estimation. The SI provides an interesting case, showing the overlap between the proximal and perceptual soundscape. While currently derived through

manual labeling, an enhanced description of the proximal soundscape through i. e.DNN could shift this class of features to be derivable from the proximal soundscape. However, whether SI markers are merely a more accurate acoustic model or provide information beyond acoustic properties cannot be determined based on the current results. A clear distinction of information regarding the perceptual sound was presented in Chapter 9. Here, CP were incorporated by training neural models on the distinction between attended and ignored speech envelopes. The feature distinction is critical to derive accurate neural models that not only explain more neural variability, but also allow for investigating the different neural processing underlying the perceptual mechanisms. In controlled experimental paradigms, this information may be present; however, it becomes difficult to assess in recording scenarios, where attentional focus cannot be directly observed. The absence of perceptual information inevitably constrains model performance and limits the scope of questions that can be addressed in naturalistic settings.

Going back to the example of the study conducted by Hölle and Bleichner (2023), it becomes apparent that knowledge regarding the sound identity, or even the sounds that were attended to, would have provided a framework to estimate more accurate neural representations of soundscape processing. However, while the results of this thesis make this point apparent, it does not provide a framework to solve this issue. Here, I would like to argue that it is impossible to obtain the ground truth without developing a feedback loop that requires active input from the perceiver. There are, however, ways in which the perceptual state of the perceiver can be approximated. In the next section, I will explore whether these options are suitable for BTL recordings of sound perception.

This thesis has demonstrated the importance of including perceptual markers such as SI and CP to explain neural variability. However, such information is typically unavailable in BTL recordings. Consequently, improving perceptual soundscape estimation requires alternative strategies. This could be addressed by either deriving neural markers indicating the internal state or by incorporating other sensors, besides EEG, to monitor the perceptual state indirectly.

Beyond externally controlled manipulations, endogenous neural dynamics can offer insight into the listener's perceptual state. For instance, alpha band fluctuations have been associated with variation in cortical excitability (Klimesch, 2012; Klimesch, Sauseng, and Hanslmayr, 2007). This has been shown to modulate sensory perception across sensory modalities, altering the perceptibility of stimuli (Busch, Dubois, and VanRullen, 2009; Craddock et al., 2017; Kayser, McNair, and Kayser, 2016; Mathewson et al., 2009). In the case of continuous auditory perception, Kasten, Busson,

and Zoefel (2023) showed that neural entrainment to speech perception fluctuated inversely to alpha power. That is, when speech entrainment was high, alpha power was low, and vice versa. Moreover, pre-stimulus alpha power differentiated between detected and missed deviant syllables, with successful detections associated with lower alpha power. These findings suggest that ongoing alpha fluctuations provide a neural signature of perceptual readiness, offering an internal measure of the listener's perceptual state. Conversely, these effects vanished when participants closed their eyes. It may be that the increase in alpha when closing the eyes (Barry et al., 2007), overshadowed the alpha fluctuations associated with auditory processing. This highlights an important caveat: multiple alpha generators may be active concurrently, which do not reflect auditory perceptual readiness. Thus, although alpha as neural marker of the perceptual state can be readily derived BTL, the existence of multiple alpha generators may complicate the application of alpha fluctuations as a perceptual marker BTL. Furthermore, so far, this has been shown predominantly using isolated and artificial stimuli; therefore, it remains to be tested whether alpha fluctuations are a suitable model of the perceptual soundscape for longitudinal recordings using more complex stimuli.

In addition to internal neural markers, other measurement modalities besides EEG have been shown to capture aspects of the perceptual state. Pupillometry, for instance, provides an accessible peripheral index of cognitive load and arousal, reflecting activity in the locus coeruleus–norepinephrine system (Aston-Jones and Cohen, 2005). Baseline pupil diameter predicts auditory deviant detection performance (Gilzenrat et al., 2010), and for complex stimuli such as speech, pupil dilation increases with reduced intelligibility (Peelle, 2018). Furthermore, pupil responses typically decrease over the course of an experiment, reflecting task-induced fatigue and reduced cognitive engagement (Zekveld, Koelewijn, and Kramer, 2018). These findings suggest that pupil dynamics can serve as a sensitive, non-invasive measure of perceptual effort and listening state. However, application of pupillometry in BTL settings faces their own unique set of issues. For instance, the shift of baseline pupil dilation when the luminescence changes in the environment, or the cancellation of the infrared signals needed to measure the pupil response when recording outside in the sunlight. Furthermore, the sluggish nature of the pupil response may be insensitive to rapid changes in attention, when integrated over a longer window of analysis (Cohen Hoffing and Thurman, 2024). These factors complicate linking pupil dilation to a specific perceptual state and thus need to be closely evaluated for a potential application BTL. Besides pupillometry, which is an indirect measure of neural activity (Aston-Jones and Cohen, 2005), respiration has been shown to modulate the perceptual state, specifically alpha

fluctuations (Kluger et al., 2021). They found that the accuracy of stimulus detection, presented near-threshold, was more accurate when respiration-induced baseline alpha power was suppressed. Together, pupillometry and respiration offer two complementary data streams that can be captured with wearable systems. Integrated with EEG, these measures could provide a more comprehensive account of the listener's perceptual state in natural environments.

The distinction between perceptual and proximal soundscapes underscores the central challenge of studying auditory perception. Methodologically, the results show that refined acoustic modeling and the integration of perceptual markers substantially improve the amount of neural variability explained, yet they cannot fully capture the experiential dimensions of perception. Recognizing this boundary is essential: it defines the epistemic limits of what can be inferred from neural data. Yet, careful feature design and multimodal integration significantly advance our understanding of the neural mechanisms underlying auditory perception.

12

THE SUITABILITY OF EEG

"But when the heart grows too full, it overflows.
And mine, inevitably, overflows onto a page."

Shannon (2020)

Key Takeaways

- EEG, as an indirect measure of neural activity, inherently constrains the claims that we can make about the brain.
- ECoG and invasive EEG (iEEG) are invasive alternatives to EEG, which offer the same temporal precision, but enhance the SNR measured.
- For BTL research, EEG is the most suitable imaging modality for the broad population to capture the temporal dynamics of auditory perception.

What you will learn:
Why EEG, despite its inherent limitations, is the most suitable imaging modality to capture neural processes underlying sound perception in everyday life.

When considering the results of the present thesis in the context of their goal to further our understanding of naturalistic soundscape perception to move measurement beyond the lab, one has to question if EEG as the imaging modality is suitable, and whether interpretative limits have been reached.

Before exploring the suitability of EEG is imaging modalities, I would like to point out one aspect I have always found confusing when reading research papers. There, statements regarding core principles of auditory processing, or regarding higher-order cognitive functions (i. e., attention), are sometimes made in isolation and never specified in the context of the imaging modality. Maybe this was due to the implicit assumption that there is a common understanding of the inevitable limitations of any

imaging modality. But the two statements: "here we have shown that auditory processing ..." and "here we have shown, using EEG, that auditory processing" have two fundamentally different implications. The former implies an unmediated access to neural mechanisms, whereas the latter confines interpretation to what EEG can reveal: large-scale, temporally precise, but spatially diffuse neural activity. In the context of the imaging modality, EEG, one has to be acutely aware that brain data is viewed through a muddy and spatially imprecise lens (see Section 4.3.1 and Section 4.3.2). Thus, any type of differential activity reflects auditory processing as measured by EEG, not auditory processing in the absolute sense.

Apart from the considerations of the inherent limitations of EEG, its susceptibility to artifacts of non-neural origins complicates the analysis of BTL recordings, where many uncontrolled factors are present. Artifacts that are not correlated to the signal of interest merely decrease the SNR and can be addressed through careful pre-processing (Jacobsen et al., 2021). Those that correlate with the signal of interest, however, bias interpretation. In Chapter 9, the artifact in the BTL walking condition did not lead to a performance decrease compared to the other conditions, but severely impaired the ability to investigate underlying neural mechanisms. Given the limited research so far on BTL measurement using several streams of data, the detection of artifact sources is underexplored and thus requires a careful approach, controlling for potential complications. These can be alleviated through additional sensors that capture a-prior assumed artifact sources (i.e. motion sensors). For instance, Studnicki, Downey, and Ferris (2022) placed a second cap, inverted, on the measurement cap during active table tennis play. The second cap only measures the artifacts produced by natural motions (motion, cable sway, electrode movement). The artifactual data is then used to project the neural and artifact cap data into an artifact subspace, to regress out the artifacts. Another study used motion sensors on the feet to determine relevant gait events and which served as reference points to determine the impact of gait on the neural recordings (Jacobsen et al., 2021). While in conflict to Bateson et al. (2017) mobility ratings, this approach shows a unique way of using additional sensors to account for artifacts. The artifact detection in naturalistic tasks can also be achieved without additional sensors. For instance, using ICA, artifactual components can be determined and even used to explain neural variability and behavior (Holtze et al., 2023; Wascher et al., 2022). It is important to note here that while artifacts are not unique to EEG as an imaging modality, its applicability BTL poses a set of unique and novel artifacts that have to be accounted for. While most of the work on ICA artifact detection has been done in stationary settings, recent studies are aiming to extending it to mobile settings (Klug et al., 2022; Klug and Gramann, 2021). In how far these identi-

fied sources remain stable over hours of recordings remains to be shown. However, if addressed correctly, EEG enables a wealth of data that can be obtained to investigate sound perception BTL.

Given the inherent limitations of EEG, the type of activity it measures, the achievable SNR, and the interpretive boundaries these impose, it is natural to question whether EEG is the appropriate choice for BTL measurement of auditory perception. For instance, alternative modalities that achieve the same high temporal solution and even higher SNR, and longer recordings as EEG, are invasive solutions such as ECoG and iEEG. iEEG has been used to decode naturalistic affective behavior from mesolimbic activity in patients undergoing seizure monitoring, using 24-hour audiovisual and neural recordings (Bijanzadeh et al., 2022). Likewise, ECoG has been used to elucidate the temporal processing of speech and sound features in the auditory cortex (Hamilton et al., 2021; Oganian and Chang, 2019). In a more impressive real-world application, ECoG has been used to decode speech production (Makin, Moses, and Chang, 2020) in the auditory cortex to enable patients to communicate in real-time (Metzger et al., 2023). ECoG and iEEG thus provide the same temporal precision as EEG, but increase the SNR and are less susceptible to non-neural artifacts. Although ECoG and iEEG are promising methods to investigate neural underpinnings of real-life cognition and auditory perception, their obvious limitation is their invasive nature. This renders them unsuitable for widespread use in healthy participants, as of now. Other non-invasive modalities such as fMRI or fNIRS may provide a higher spatial resolution, or can even be applied in BTL settings in the case of fNIRS. However, due to the sluggish nature of the hemodynamic response that is measured by both modalities, they generally lack the temporal resolution to detect rapid auditory processing that is of interest in this thesis. Thus, these imaging modalities are more suited to answer different kinds of research questions.

Given the requirement of high-level temporal precision, as well as the portability required, EEG is currently the most suitable non-invasive candidate for the general population, as argued in Section 4.6. While EEG does not provide the spatial specificity or internal representational format of neural information (i. e.single cell recordings) compared to other imaging modalities, it excels at tracing the temporal evolution of how complex auditory information is represented. Furthermore, the temporal precision allows us to observe how the auditory system continuously recalibrates its responses to a dynamic environment. Additionally, the ability to record for multiple hours continuously allows experiments to monitor the previously mentioned dynamics over long periods of time. Therefore, EEG provides a suitable imaging modality for under-

standing perception in naturalistic contexts, if the inherent limitations are addressed correctly.

13

LESSONS FOR SCIENTIFIC INVESTIGATION

"The brain is not an information-absorbing, perpetual coding device, as it is often portrayed, but a venture-seeking explorer, an action-obsessed agent constantly controlling the body's actuators and sensors to test its hypotheses."

Buzsáki (2019)

Key Takeaways

- The lab-dilemma, which refers to the constraints of generalizing laboratory findings, can be improved by BTL recordings, through careful operationalization of environmental factors.
- Information processing in the brain needs to be considered from a brain perspective (cortex-as-receiver) rather than from the perspective of the experimenter (experimenter-as-receiver).
- Interpretation of feature sets being encoded in the brain is a misconception, as a theoretical infinite amount of features could explain the same neural variability.

What you will learn:
A theoretical basis to question neuroscientific concepts that appear intuitively clear and what we can learn about the brain.

Throughout this thesis, I have argued that going BTL enables the investigation of how auditory information processing can be interpreted using TRF models. These findings would allow to advance our understanding of sound perception in everyday life. This statement rests on three central assumptions that have not yet been critically examined: 1. that going beyond the lab resolves the constraints of artificial experimental settings, 2. that we can meaningfully measure information processing in the brain,

and 3. that TRF models provide interpretable insights into neural processing. In the following sections, I will critically evaluate whether these assumptions are justified and can be scientifically supported.

13.1 ADDRESSING THE LAB-DILEMMA IN THE ROOM

Throughout this thesis, I have mentioned that the concept of BTL measurements and usage of naturalistic soundscapes are vital to understand real-world sound perception. This is based on the assumption that recordings conducted in the laboratory do not represent the complex and real-world dynamics, where the immersion of participants is constrained. Thus, it is questionable whether laboratory findings in isolated settings and artificial stimuli generalize to real-world processes. This conflict has been coined the Lab-dilemma.

The lab-dilemma emerged as a response to the growing tendency of researchers adopting a reductionist approach in the 20th century. As real-world behavior appeared too complex and confounded for systematic scientific study, experimental control was sought by isolating the concept of interest in highly controlled settings (Vallet and Van Wassenhove, 2023). Brunswik (1943, p. 262) was among the first to challenge this view and pointed out the issue with *"narrow-spanning problems of artificially isolated proximal or peripheral technicality of mediation which are not representative of the larger patterns of life"*. In other words, the study of a concept outside the context in which it is encountered in real life does not provide meaningful insights into explaining it in the first place. This concept has been termed ecological validity (Brunswik, 1952).

This issue persists in modern neuroscience. For example, in the field of neuroergonomics, which studies the neural substrates of decision-making, risk is tied to the expected payout. In how far this reflects decision-making processes in real-life, where personal finances are involved, is highly questionable (Johnson, Stopka, and Bell, 2002). Apart from the tasks themselves, the artificial setting of the laboratory may already induce confounding effects, altering the behavior of participants due to unfamiliarity (Orne, 1962). These concerns can be extended to auditory research, where controlled stimuli such as isolated tones or synthetic speech fail to capture the acoustic and contextual richness of natural soundscapes (Hamilton and Huth, 2020; Lorenzi et al., 2023). Empirical comparisons support this critique: studies contrasting laboratory and naturalistic conditions have revealed substantial differences in both neural activation and behavioral performance (Bohbot et al., 2017; Jeung et al., 2023; Krakauer et al., 2017). These findings highlight that while reductionist approaches

enable precision and reproducibility, they may not generalize to the dynamic complexity of everyday perception. Consequently, researchers have called for increasing ecological validity by taking experimentation beyond the lab. Yet, the critical question remains whether doing so truly resolves the lab-dilemma, or simply shifts it to a new level of methodological and interpretive complexity.

Going BTL, however, does not automatically solve the lab-dilemma. Several researchers have argued that the concept of ecological validity is often ill-posed or inconsistently applied. In many cases, the defining characteristics of real-world situations are underspecified, making it unclear what aspects of the experiment make it 'realistic' (Holleman et al., 2020). Specifically, Holleman et al. (2020) critiques that what is considered complex and naturalistic often depends on the researcher applying the paradigm. Without a clear definition of what constitutes complexity or naturality, these concepts lose analytical value. This lack of factorization, that is, the operationalization of the complex, naturalistic environments into measurable variables, undermines the ability to generalize findings to everyday cognition (Vallet and Van Wassenhove, 2023). A field in which this is especially evident is Virtual Reality (VR) research. VR immerses participants in simulated environments that aim to reproduce real-world experiences (Schöne et al., 2023). However, in order for VR to successfully enhance ecological validity, it is necessary to understand which environmental features actually define the experience. Without such theoretical grounding, VR studies risk reproducing superficial aspects of realism while missing the contextual and affective dimensions that shape real-world perception (Stangl, Maoz, and Suthana, 2023; Vallet and Van Wassenhove, 2023).

In order to solve this, Schmuckler (2001) proposed to evaluate ecological validity along three dimensions: 1. the stimuli, 2. the task, 3. the context. These three dimensions can be grouped into simplicity-complexity and artificiality-naturality. The first refers to the stimulus nature, where isolated trains of click tones would be considered simplistic compared to a complex proximal soundscape (refer to Section 3.1 how I defined complexity). The latter dimension refers to the setting, where being seated in a soundproof booth in front of a computer screen would be considered artificial, compared to being in a real-life setting (i. e., a cafe). This provides a useful distinction when contrasting stimuli and settings in relation to each other within the same experimental paradigm. While any manipulation that disrupts natural behavior diminishes the naturality of the experience of the situation, it presents a step towards real-life investigation (Shamay-Tsoory and Mendelsohn, 2019). A compelling example of this bridging approach is offered by the in situ experiments outlined by Vallet and Van Wassenhove (2023) (for a list of guiding principles for in situ experiments, re-

fer to the original paper). These designs explicitly factorize environmental variables, for example, investigating optical flow in train passengers sitting in versus against the direction of travel, to examine how contextual factors modulate neural activity in real-life situations. Such approaches retain ecological realism while preserving experimental interpretability, demonstrating how real-world research can progress beyond mere immersion toward systematic explanation.

Rather than attempting to fully replicate real-world conditions, I approached the lab-dilemma by using more complex auditory stimuli and contrasting different environmental settings. For instance, the Chapter 7 and Chapter 8 used a soundscape emulating an operating room environment. This soundscape included both task-relevant and irrelevant speech and non-speech sounds, reflecting the auditory richness of an actual operating room. Investigating auditory perception in this more complex soundscape, compared to isolated and repeated click tones, provides a strong basis to suggest that findings are more likely to reflect real-world settings. While the stimuli can be considered more complex, the setting would be classified as being artificial. Thus, the possibility that sound perception in a real-life operating room may differ substantially remains possible. This distinction illustrates that increasing stimulus complexity alone cannot fully substitute for contextual realism. Concomitantly, the Chapter 9 used a stepwise transition from laboratory to real-world environments, enabling assessment of environmental influences on auditory processing. Here, the setting went from artificial to more naturalistic. The task, as discussed in Section 9.4.4, may not be the most suitable to investigate auditory perception. However, compared to performing a BTL oddball paradigm (Scanlon et al., 2019), the task may be regarded as being more relevant to real-life behavior. While not fully resolving the lab-dilemma, these studies demonstrate how ecological validity can be systematically increased through controlled variation in stimulus complexity, setting, and task design, thereby bridging the gap between controlled laboratory research and beyond-the-lab investigation.

13.2 INFORMATION MODELING

Having addressed whether going beyond the lab will solve the lab-dilemma, the next aspect to consider is whether I am actually measuring information processing in the brain. While ecological validity determines the context of measurement, the concept of "information" itself governs what we claim to be measuring.

In their work, Wit et al. (2016) make an important point of questioning whether neuroscience in general is measuring information in the brain. They start by pointing out

that neuroscience frequently assumes a shared understanding of what 'information' means without ever defining it. They conceptualize their arguments by using Shannon (1948) framework of information processing, which consists of three components: the transmitter, the channel, and the receiver. Here, information is encoded by the transmitter, transferred over the channel, and decoded by the receiver. In this definition, information refers to how activity is used by the receiver and formulated as *"the difference that makes a difference"* (Wit et al., 2016, p. 1416). The authors highlight a critical misconception in neuroscience, as differential activity in area X is interpreted as representing information Y. In doing so, the experimenter, not the brain, assumes the role of the receiver. The authors argue that this stance restricts interpretation to correlational inference and obscures how neural activity is actually used by other brain regions.

In the context of EEG, this issue becomes apparent in how we interpret ERP, which are viewed as reflecting information processing (Näätänen, 1990). This can be extended to the auditory domain, where it remains unclear whether AEP reflects genuine evoked activity or phase reorganization (Burgess, 2012; Obleser and Kayser, 2019; Oganian et al., 2023). Interpreting ERP peaks as direct markers of information processing, therefore, overlooks how the brain itself may utilize the underlying neural dynamics. This issue extends to forward modeling approaches, such as the TRFs used throughout this thesis. Looking back at the example of TRFs representing a window into the neural activity to provide an intuitive understanding, aptly mirrors the issue highlighted by Wit et al. (2016). TRF models offer snapshots of cortical activity correlated with specific stimulus features, but do not reveal how these activations are functionally used within the brain. Thus, according to Wit et al. (2016) using TRFs does not measure information processing in the brain.

Conversely, there are circumstances in which it is desirable to have the experimenter as receiver. This is the case when the neural activation is used to inform the experimenter or some other system (BCI). As evidenced by the approach of AAD showcased in Chapter 9, where the explicit goal is to determine whether recorded brain activity can be used to reconstruct a perceived sound envelope. Thus, the explicit goal is to decode the brain activity, rather than to understand how it is encoded functionally. In this context, information is operationalized as utility rather than mechanism.

An example of how the brain can be conceptualized as a perceiver is provided in Chapter 2, where the decomposition of sound is traced from the spectral decomposition in the cochlea (Section 2.2.1) to the formation of the auditory objects in the cortex (Section 2.1.5). The progressive transformation of neural activity along the auditory pathway exemplifies information processing as a hierarchical interpretive pro-

cess. Therefore, rather than analyzing regions of interest showing variable activity in response to stimulation, research should focus on building models of neural activation (i. e., network, computational neuroscience) (Barabási et al., 2023), identify the different coding channels involved (i. e., phase, amplitude, population codes) (Buzsáki, 2010; Panzeri et al., 2015), and link it to observable behavioral variations (Krakauer et al., 2017).

Although Shannon's framework provides a valuable metaphor for thinking about information transmission, its direct application to the brain is limited. The brain is not a simple linear channel with fixed boundaries between transmitter, receiver, and decoder; rather, it is a dynamic, recursive network with extensive feedback loops (see Section 2.1.4). Nonetheless, the model highlights the importance of critically examining our assumptions on the concept of information processing, particularly the tendency to view the brain as a transmitter of information rather than as an active receiver engaged in interpretation. As emphasized in this section, especially in forward modeling, the experimenter-as-perceiver stance persists. While this stance in itself is not invalid, one has to be aware of the role of the experimenter when drawing conclusions regarding neural processing. How information in neural data can be identified, given the experimenter-as-perceiver constraint, and the additional pitfalls of feature and single model interpretation, will be discussed next.

Despite the clearly delineated view of how to regard information processing in the brain, there are explicit uses of encoding and decoding models in the context of information processing in the brain. As a necessary step for functional interpretation of information processing, that is, how the information is used by other regions, is to determine whether information is represented in a region of interest at all (Diedrichsen, Yokoi, and Arbuckle, 2018). Encoding and decoding models, as used in this thesis, provide complementary tools to explore this question. Decoding models can only be used to determine the content of a neural recording by mapping from neural activity back to the stimulus; thus, they do not provide a computational brain model. Encoding models, however, can be used to model computational processes by predicting raw neural activity from some feature. However, one must be cautious when interpreting the results (Crosse et al., 2021; Holdgraf et al., 2017; Kriegeskorte and Douglas, 2019).

The activity in a region, regarded as an activity profile, can be modeled by the linear combination of different features (Diedrichsen, 2020). Here, the features can be considered as the basis vectors that span the subspace where the activity profiles are represented. It is tempting to interpret a specific set of features, explaining a significant amount of variance, as being encoded in the recorded area. However, such an inference is flawed, since any number of linear combinations of features may explain

comparable amounts of variance without necessarily reflecting the same underlying neural computation. This is a problem known as the feature fallacy (Diedrichsen, 2020). Besides the feature fallacy, another pitfall is to interpret a single encoding model explaining significant variance to reflect the computations underlying the encoding. This is termed the single-model-significance fallacy (Kriegeskorte and Douglas, 2019). These fallacies are particularly relevant in studies (i. e., Chapter 7 and Chapter 8) presented here, where multiple acoustic and perceptual features may explain overlapping variance in EEG. Especially, where specific sets of features are tested and compared to each other. In both cases, the feature and single model fallacy, a more defensible conclusion would be to infer that the presented stimulus information is represented in the recorded neural activity, rather than reflecting an active process of encoding. Therefore, a more robust approach is to compare multiple feature models, evaluating their relative predictive performance rather than treating any single feature as uniquely represented (Desai et al., 2021; Heer et al., 2017; Kriegeskorte and Douglas, 2019).

Having discussed the challenges of ecological validity, the limits of measuring information in the brain, and the interpretational pitfalls of feature-based modeling, we can now return to a central question: what, ultimately, have we learned about the brain?

13.3 LESSONS ABOUT THE BRAIN

In this thesis, I presented the hierarchical processing of sound information along the auditory pathway (Chapter 2) and a series of empirical studies investigating distinct aspects of naturalistic auditory processing using mobile EEG. A natural question, then, is: what have I learned about the brain's response to naturalistic soundscapes?

To answer this, it is important to acknowledge what I cannot claim. None of the studies presented here directly measures information processing in the brain from the perspective of the cortex as a perceiver, since the analyses do not reveal how information is transmitted between regions. Additionally, despite the numerous features that I have derived, compared, and tested, I cannot infer how they are encoded in the brain, given the methods that I have used. These constraints, however, are informative; they delineate the epistemic boundaries within which meaningful inference about brain function can be made.

Within these boundaries, the results of this thesis contribute unique insights. Specifically, it is one of the first that shows how the brain represents different type of features (i. e., proximal and perceptual soundscape features) of naturalistic soundscapes. The

findings that more detailed acoustic features (e. g., spectrally resolved representations) explain more neural variability reveal that detailed information of the soundscape is represented in the measured EEG signal. Additionally, the finding that acoustic onsets account for much of the envelope's predictive power highlights that the EEG captures the brain's tendency to react to change over constancy. This suggests that the auditory system, as measured by EEG, represents salient acoustic transitions rather than static physical features, a hallmark of predictive and adaptive processing (Willmore and King, 2023). Similarly, this is one of the first academic works that shows that the attenuation of neural responses, as a function of IOI, is found in response to complex, ecologically valid stimuli, rather than isolated pure tones. This shows that temporal context dynamically modulates auditory responsiveness, even in acoustically rich environments, including both speech and non-speech sounds. It is important to consider these findings in the light of the imaging modality chosen. EEG, given its limited spatial resolution and sensitivity to large-scale cortical fields, constrains interpretation to surface-level correlates of neural population activity. Thus, conclusions regarding principles of auditory processing are constrained by the synchronous activity that is detected by EEG (see Section 4.3.1 and Chapter 12 for a discussion). So whether these findings are true core principles of auditory processing, or merely the only activity that is detected by the EEG, cannot be answered by this thesis. It does, however, provide evidence, that meaningful activity of the brain can be captured in response to naturalistic soundscape perception.

In sum, this thesis does not claim to have uncovered how information is processed in the brain in a mechanistic sense. Instead, it demonstrates how EEG can reveal when and to what extent acoustic and perceptual features are represented in neural activity. Furthermore, I have shown how these representations adapt to the temporal structure of the auditory context. Together, these findings establish a methodological and conceptual foundation for future research aiming to bridge the gap between measurable acoustic environment and the lived experience of auditory perception.

14

OUTLOOK

"Coming back to where you started
is not the same as never leaving."

Pratchett (2005)

Going forward, an interesting question is to explore how the aspects that were previously discussed can be implemented to guide future research projects. For this, I would propose to distinguish two lines of research aiming at 1. Understanding the capacities of EEG in detecting sound perception in response to naturalistic soundscapes using a cross-model approach and enhanced information processing methods 2. Implementing and testing various additional sensors in BTL settings to improve the estimation of the proximal and perceptual soundscape. These two approaches complement each other and mirror the approach chosen in this thesis. By critically evaluating what can be measured with EEG in controlled settings, findings can be extrapolated to BTL recordings. For this, not only the stimuli in controlled settings need to more closely mimic soundscapes encountered BTL, but also the ability of different sensors to adequately depict the proximal and perceptual soundscape needs to be evaluated.

Regarding the first point, testing the capacities of EEG in understanding auditory perception would require modeling information processing in the brain using network analysis. Here, source localization analysis may be used to determine relevant network structures involved in sound processing (Kasten, Busson, and Zoefel, 2024). This would elucidate the involvement of different brain structures and more adequately depict how information is processed in the brain. Rather than determining differential activity in AEPs, network analysis takes the interdependencies of several brain areas into consideration and thus more closely models how the brain operates (Barabási et al., 2023). Furthermore, I would advocate combining EEG with fMRI. Given the higher spatial resolution and the advances in auditory object representation in fMRI research (Formisano, 2025; Giordano et al., 2023; Santoro et al., 2014) would help

to set a reference in determining what signals can be captured by EEG. Although these two modalities measure different signals, the assumption that these reflect underlying neural activity would provide a frame of reference for what can be measured with EEG. That is, given that the new set of artifacts i. e., scanner noise, gradient artifacts, that is introduced by the combination of these two modalities are sufficiently addressed. Nonetheless, the combination of network analysis cross-imaging modalities elucidates an improved estimation of how information is processed in the brain and what can be detected with EEG.

Simultaneously, I would investigate how additional sensors may contribute to the estimation of the proximal and perceptual soundscape in BTL recordings. For this, I propose to use an In Situ experimental setup (see Section 13.1 for a definition of In Situ experiments), emulating a situation where laboratory-level information (room recording, orange) is contrasted to that available to a closed-loop (person-centric, blue) BTL setup (Figure 37A). Here, an office scenario could be recreated, where laboratory-grade information can be determined through several microphones, placed around the room, picking up the different sound sources. Furthermore, information regarding CP can be extracted through experimental manipulation information (i. e. *"respond to the sound of email notifications of your colleagues"* vs. *"ignore those sounds"*). This information availability can then be contrasted to an emulated BTL recording setup, where merely the microphone in the nEEGlace (Bleichner and Emkes, 2020) records the mixture of sound sources, and no CP information is available. The sensitivity of EEG recordings in response to the different sets of information spaces can be contrasted using the variance explained of the different TRF (Figure 37B). In the scope of this experiment, the application of different DNN models and of bio-informed acoustic modeling could be tested to determine the precision with which the auditory soundscape could be described. Additional sensors, such as gyrometers in the amplifier, as well as an additional high-density cap, worn by one of the co-workers, could elucidate the effects of head movement on the recorded EEG, as well as allow for contrasting high-density vs. low-density setups.

14.1 CONCLUSION

Together, the three studies presented in this thesis form a coherent progression toward understanding how the brain processes complex, naturalistic soundscapes. The first study identified which features of the proximal soundscape, ranging from simple acoustic onsets to semantically meaningful sound identities, predict neural variability.

The second study expanded this framework by showing that temporal context, captured through inter-onset intervals, systematically shapes neural adaptation, revealing non-linear properties of auditory processing in natural environments. The third study moves towards a BTL setting using mobile EEG, while also highlighting the challenges of maintaining signal interpretability in real-world conditions. Collectively, these findings establish both the conceptual and methodological groundwork for linking measurable acoustic structure to perceptual experience, forming the foundation for future investigations into everyday auditory perception using mobile neuroimaging.

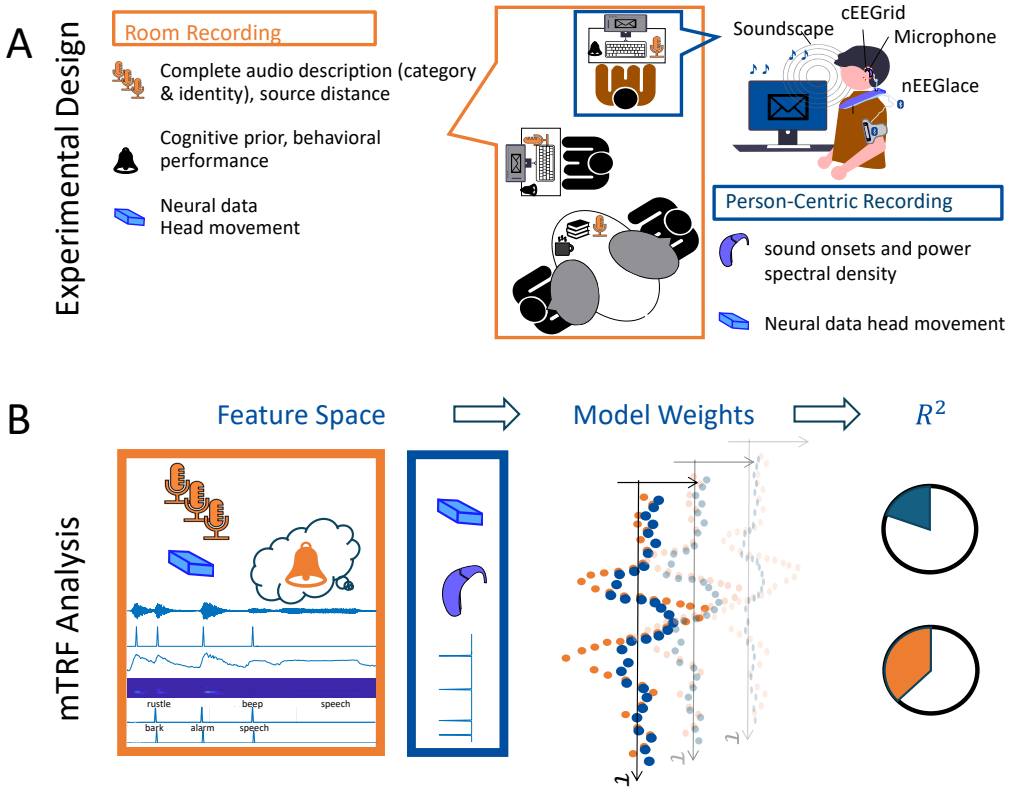

Figure 37: Given the results of the current thesis, an In Situ experiment could provide the next step in investigating auditory perception BTL. **A:** shows a theoretical experimental paradigm in an office. Two information spaces can be obtained. The person-centric recording (blue) is limited to the sensors worn by the participant, including cEEGrids, an amplifier, and microphones. The room recording (orange) contains, besides all the sensor streams from the person-centric recording, several different microphones placed next to potential sound sources. Furthermore, CP and experimental markers can be obtained through experimental manipulation. **B:** proposes a multivariate TRF analysis to determine the degree of neural variability explained by features derived from the respective information spaces. This setup provides a point of reference to determine the type of information that can be extracted and in how far they are able to explain neural variability.

BIBLIOGRAPHY

Adler, G. and J. Adler (1989). "Influence of Stimulus Intensity on AEP Components in the 80- to 200-Millisecond Latency Range." In: *Audiology* 28.6, pp. 316–324. DOI: 10.3109/00206098909081638 (cit. on pp. 36, 104, 115).

Agmon, Galit et al. (2023). ""Um. . . , It's Really Difficult to. . . Um. . . Speak Fluently": Neural Tracking of Spontaneous Speech." In: *Neurobiology of Language* 4.3, pp. 435–454. DOI: 10.1162/nol_a_00109 (cit. on pp. 34, 93).

Alain, Claude and István Winkler (2012). "Recording Event-Related Brain Potentials: Application to Study Auditory Perception." en. In: *The Human Auditory Cortex*. Ed. by David Poeppel et al. New York, NY: Springer, pp. 69–96. DOI: 10.1007/978-1-4614-2314-0_4 (cit. on p. 93).

Altmann, C. F. et al. (2008). "Temporal Dynamics of Adaptation to Natural Sounds in the Human Auditory Cortex." en. In: *Cerebral Cortex* 18.6, pp. 1350–1360. DOI: 10.1093/cercor/bhm166 (cit. on pp. 113, 117).

Amzica, Florin and Fernando H. Lopes da Silva (2017). "C2Cellular Substrates of Brain Rhythms." In: *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Ed. by Donald L. Schomer et al. Oxford University Press, p. 0. DOI: 10.1093/med/9780190228484.003.0002 (cit. on pp. 26, 27, 33).

Ansari, Sam et al. (2023). "A survey of artificial intelligence approaches in blind source separation." In: *Neurocomputing* 561, p. 126895. DOI: 10.1016/j.neucom.2023.126895 (cit. on p. 23).

Antunes, Flora M. and Manuel S. Malmierca (2014). "An Overview of Stimulus-Specific Adaptation in the Auditory Thalamus." en. In: *Brain Topography* 27.4, pp. 480–499. DOI: 10.1007/s10548-013-0342-6 (cit. on p. 10).

Antunes, Flora M. and Manuel S. Malmierca (2021). "Corticothalamic Pathways in Auditory Processing: Recent Advances and Insights From Other Sensory Systems." English. In: *Frontiers in Neural Circuits* 15. DOI: 10.3389/fncir.2021.721186 (cit. on pp. 8, 10).

Aston-Jones, Gary and Jonathan D. Cohen (2005). "An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance." eng. In: *Annual Review of Neuroscience* 28, pp. 403–450. DOI: 10.1146/annurev.neuro.28.061604.135709 (cit. on p. 163).

Axelsson, Östen, Mats E. Nilsson, and Birgitta Berglund (2010). "A principal components model of soundscape perception." en. In: *The Journal of the Acoustical Society of America* 128.5, pp. 2836–2846. DOI: 10.1121/1.3493436 (cit. on p. 23).

Barabási, Dániel L. et al. (2023). "Neuroscience Needs Network Science." en. In: *Journal of Neuroscience* 43.34, pp. 5989–5995. DOI: 10.1523/JNEUROSCI.1014-23.2023 (cit. on pp. 174, 177).

Barry, Robert J. et al. (2007). "EEG differences between eyes-closed and eyes-open resting conditions." In: *Clinical Neurophysiology* 118.12, pp. 2765–2773. DOI: 10.1016/j.clinph.2007.07.028 (cit. on p. 163).

Bartlett, Edward L. (2013). "The organization and physiology of the auditory thalamus and its role in processing acoustic features important for speech perception." In: *Brain and Language* 126.1, pp. 29–48. DOI: 10.1016/j.bandl.2013.03.003 (cit. on p. 11).

Bateson, Anthony D. et al. (2017). "Categorisation of Mobile EEG: A Researcher's Perspective." In: *BioMed Research International* 2017.1, p. 5496196. DOI: 10.1155/2017/5496196 (cit. on pp. 39, 57, 166).

Beauducel, André et al. (2000). "On the reliability of augmenting/reducing: Peak amplitudes and principal component analysis of auditory evoked potentials." In: *Journal of Psychophysiology* 14.4, pp. 226–240. DOI: 10.1027/0269-8803.14.4.226 (cit. on pp. 37, 112, 115).

Benjamini, Yoav and Daniel Yekutieli (2001). "The Control of the False Discovery Rate in Multiple Testing under Dependency." In: *The Annals of Statistics* 29.4, pp. 1165–1188 (cit. on pp. 68, 133).

Bijanzadeh, Maryam et al. (2022). "Decoding naturalistic affective behaviour from spectro-spatial features in multiday human iEEG." en. In: *Nature Human Behaviour* 6.6, pp. 823–836. DOI: 10.1038/s41562-022-01310-0 (cit. on p. 167).

Bleichner, Martin G. and Stefan Debener (2017). "Concealed, Unobtrusive Ear-Centered EEG Acquisition: cEEGrids for Transparent EEG." In: *Frontiers in Human Neuroscience* 11. DOI: 10.3389/fnhum.2017.00163 (cit. on pp. 31, 39, 57, 124).

Bleichner, Martin G. and Reiner Emkes (2020). "Building an Ear-EEG System by Hacking a Commercial Neck Speaker and a Commercial EEG Amplifier to Record Brain Activity Beyond the Lab." en. In: *Journal of Open Hardware* 4.1, p. 5. DOI: 10.5334/joh.25 (cit. on pp. 39, 178).

Boer, R. de and P. Kuyper (1968). "Triggered correlation." eng. In: *IEEE transactions on bio-medical engineering* 15.3, pp. 169–179. DOI: 10.1109/tbme.1968.4502561 (cit. on p. 132).

Bohbot, Véronique D. et al. (2017). "Low-frequency theta oscillations in the human hippocampus during real-world and virtual navigation." en. In: *Nature Communications* 8.1, p. 14415. DOI: 10.1038/ncomms14415 (cit. on p. 170).

Bosnyak, Daniel J., Robert A. Eaton, and Larry E. Roberts (2004). "Distributed auditory cortical representations are modified when non-musicians are trained at pitch discrimination with 40 Hz amplitude modulated tones." eng. In: *Cerebral Cortex (New York, N.Y.: 1991)* 14.10, pp. 1088–1099. DOI: 10.1093/cercor/bhh068 (cit. on p. 38).

Breska, Assaf and Leon Y. Deouell (2017). "Neural mechanisms of rhythm-based temporal prediction: Delta phase-locking reflects temporal predictability but not rhythmic entrainment." en. In: *PLOS Biology* 15.2, e2001665. DOI: 10.1371/journal.pbio.2001665 (cit. on p. 35).

Briley, Paul M. and Katrin Krumbholz (2013). "The specificity of stimulus-specific adaptation in human auditory cortex increases with repeated exposure to the adapting stimulus." In: *Journal of Neurophysiology* 110.12, pp. 2679–2688. DOI: 10.1152/jn.01015.2012 (cit. on p. 114).

Brodbeck, Christian, L. Elliot Hong, and Jonathan Z. Simon (2018). "Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech." eng. In: *Current biology: CB* 28.24, 3976–3983.e5. DOI: 10.1016/j.cub.2018.10.042 (cit. on pp. 62, 85, 118, 144, 145).

Brodbeck, Christian, Alessandro Presacco, and Jonathan Z. Simon (2018). "Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension." In: *NeuroImage* 172, pp. 162–174. DOI: 10.1016/j.neuroimage.2018.01.042 (cit. on pp. 12, 34, 48, 159).

Brodbeck, Christian et al. (2023). "Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions." In: *eLife* 12. Ed. by Andrea E Martin, Barbara G Shinn-Cunningham, and Sophie Slaats, e85012. DOI: 10.7554/eLife.85012 (cit. on pp. 84, 93).

Broek, S. P van den et al. (1998). "Volume conduction effects in EEG and MEG." In: *Electroencephalography and Clinical Neurophysiology* 106.6, pp. 522–534. DOI: 10.1016/S0013-4694(97)00147-8 (cit. on p. 28).

Bruns, Patrick (2019). "The Ventriloquist Illusion as a Tool to Study Multisensory Processing: An Update." English. In: *Frontiers in Integrative Neuroscience* 13. DOI: 10.3389/fnint.2019.00051 (cit. on p. 23).

Brunswik, E. (1943). "Organismic achievement and environmental probability." In: *Psychological Review* 50.3, pp. 255–272. DOI: 10.1037/h0060889 (cit. on p. 170).

Brunswik, Egon (1952). *The conceptual framework of psychology. (Int. Encycl. unified Sci., v. 1, no. 10.)* The conceptual framework of psychology. (Int. Encycl. unified Sci., v. 1, no. 10.) Oxford, England: Univ. Chicago Press (cit. on p. 170).

Budd, T.W et al. (1998). "Decrement of the N1 auditory event-related potential with stimulus repetition: habituation vs. refractoriness." en. In: *International Journal of Psychophysiology* 31.1, pp. 51–68. DOI: 10.1016/S0167-8760(98)00040-3 (cit. on p. 113).

Burgess, Adrian P. (2012). "Towards a Unified Understanding of Event-Related Changes in the EEG: The Firefly Model of Synchronization through Cross-Frequency Phase Modulation." en. In: *PLOS ONE* 7.9, e45630. DOI: 10.1371/journal.pone.0045630 (cit. on pp. 33, 173).

Busch, Niko A., Julien Dubois, and Rufin VanRullen (2009). "The Phase of Ongoing EEG Oscillations Predicts Visual Perception." en. In: *Journal of Neuroscience* 29.24, pp. 7869–7876. DOI: 10.1523/JNEUROSCI.0113-09.2009 (cit. on p. 162).

Buzsaki, György (2006). *Rhythms of the brain*. eng. New York, NY: Oxford Univeristy Press (cit. on p. 33).

Buzsáki, G. (2019). *The brain from inside out*. New York, NY: Oxford University Press (cit. on p. 169).

Buzsáki, György (2010). "Neural syntax: cell assemblies, synapsembles, and readers." eng. In: *Neuron* 68.3, pp. 362–385. DOI: 10.1016/j.neuron.2010.09.023 (cit. on p. 174).

Buzsáki, György (2020). "The Brain–Cognitive Behavior Problem: A Retrospective." en. In: *eNeuro* 7.4. DOI: 10.1523/ENEURO.0069-20.2020 (cit. on p. 42).

Buzsáki, György, Costas A. Anastassiou, and Christof Koch (2012). "The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes." en. In: *Nature Reviews Neuroscience* 13.6, pp. 407–420. DOI: 10.1038/nrn3241 (cit. on pp. 25–27, 33).

Buzsáki, György and Kenji Mizuseki (2014). "The log-dynamic brain: how skewed distributions affect network operations." en. In: *Nature Reviews Neuroscience* 15.4, pp. 264–278. DOI: 10.1038/nrn3687 (cit. on pp. 48, 94).

Buzsáki, György and Mihály Vöröslakos (2023). "Brain rhythms have come of age." en. In: *Neuron* 111.7, pp. 922–926. DOI: 10.1016/j.neuron.2023.03.018 (cit. on p. 33).

Bünau, Paul Von (2012). "Stationary Subspace Analysis: Towards understanding non-stationary data." en. In: DOI: 10.14279/DEPOSITONCE-3357 (cit. on p. 42).

Cant, Nell Beatty (2005). "Projections from the Cochlear Nuclear Complex to the Inferior Colliculus." en. In: *The Inferior Colliculus*. Ed. by Jeffery A. Winer and

Christoph E. Schreiner. New York, NY: Springer, pp. 115–131. DOI: 10.1007/0-387-27083-3_3 (cit. on p. 9).

Carbajal, Guillermo V. and Manuel S. Malmierca (2018). "The Neuronal Basis of Predictive Coding Along the Auditory Pathway: From the Subcortical Roots to Cortical Deviance Detection." In: *Trends in Hearing* 22, p. 2331216518784822. DOI: 10.1177/2331216518784822 (cit. on pp. 9, 10, 14, 16, 17, 114).

Celesia, Gastone G. and Gregory Hickok (2015). *The human auditory system: fundamental organization and clinical disorders*. eng. Handbook of clinical neurology volume 129, 3rd series. Amsterdam, Netherlands: Elsevier B.V (cit. on p. 11).

Chalas, Nikos et al. (2023). "Speech onsets and sustained speech contribute differentially to delta and theta speech tracking in auditory cortex." In: *Cerebral Cortex* 33.10, pp. 6273–6281. DOI: 10.1093/cercor/bhac502 (cit. on p. 33).

Cherry, E. Colin (1953). "Some experiments on the recognition of speech, with one and with two ears." In: *Journal of the Acoustical Society of America* 25, pp. 975–979. DOI: 10.1121/1.1907229 (cit. on p. 123).

Cheveigné, Alain de and Israel Nelken (2019). "Filters: When, Why, and How (Not) to Use Them." en. In: *Neuron* 102.2, pp. 280–293. DOI: 10.1016/j.neuron.2019.02.039 (cit. on p. 35).

Ciccarelli, Gregory et al. (2019). "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods." en. In: *Scientific Reports* 9.1, p. 11538. DOI: 10.1038/s41598-019-47795-0 (cit. on pp. 124, 147).

Cohen Hoffing, Russell A. and Steven M. Thurman (2024). "What's a Pupil Worth? The Promise and Challenges of Cognitive Pupillometry in the Wild." en. In: *Modern Pupillometry: Cognition, Neuroscience, and Practical Applications*. Ed. by Megan H. Papesh and Stephen D. Goldinger. Cham: Springer International Publishing, pp. 259–282. DOI: 10.1007/978-3-031-54896-3_9 (cit. on p. 163).

Costa-Faidella, Jordi et al. (2011). "Interactions between "What" and "When" in the Auditory System: Temporal Predictability Enhances Repetition Suppression." en. In: *Journal of Neuroscience* 31.50, pp. 18590–18597. DOI: 10.1523/JNEUROSCI.2599-11.2011 (cit. on pp. 112, 113).

Craddock, Matt et al. (2017). "Pre-stimulus alpha oscillations over somatosensory cortex predict tactile misperceptions." In: *Neuropsychologia* 96, pp. 9–18. DOI: 10.1016/j.neuropsychologia.2016.12.030 (cit. on p. 162).

Crosse, Michael J. et al. (2016). "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli." In: *Frontiers in Human Neuroscience* 10 (cit. on pp. 47, 59, 62, 67, 93, 94, 99, 123, 132, 143).

Crosse, Michael J. et al. (2021). "Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research." In: *Frontiers in Neuroscience* 15, p. 705621. DOI: 10.3389/fnins.2021.705621 (cit. on pp. 42, 48, 62, 63, 70, 93, 94, 99, 134, 143, 144, 174).

Da Silva Souto, Carlos F. et al. (2022). "Pre-gelled Electrode Grid for Self-Applied EEG Sleep Monitoring at Home." In: *Frontiers in Neuroscience* 16, p. 883966. DOI: 10.3389/fnins.2022.883966 (cit. on p. 144).

Darwin, Christopher J. (2005). "Pitch and Auditory Grouping." en. In: *Pitch: Neural Coding and Perception*. Ed. by Christopher J. Plack et al. New York, NY: Springer, pp. 278–305. DOI: 10.1007/0-387-28958-5_8 (cit. on p. 3).

Daube, Christoph, Robin A. A. Ince, and Joachim Gross (2019). "Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech." en. In: *Current Biology* 29.12, 1924–1937.e9. DOI: 10.1016/j.cub.2019.04.067 (cit. on pp. 34, 62, 84, 86).

Davis, P. A. (1939). "Effects of acoustic stimuli on the waking human brain." In: *Journal of Neurophysiology* 2.6, pp. 494–499. DOI: 10.1152/jn.1939.2.6.494 (cit. on p. 37).

Dean, Isabel, Nicol S. Harper, and David McAlpine (2005). "Neural population coding of sound level adapts to stimulus statistics." en. In: *Nature Neuroscience* 8.12, pp. 1684–1689. DOI: 10.1038/nn1541 (cit. on pp. 15, 113).

Debener, S et al. (2002). "Auditory novelty oddball allows reliable distinction of top–down and bottom–up processes of attention." In: *International Journal of Psychophysiology* 46.1, pp. 77–84. DOI: 10.1016/S0167-8760(02)00072-7 (cit. on p. 37).

Debener, Stefan et al. (2005). "What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis." In: *Cognitive Brain Research* 22.3, pp. 309–321. DOI: 10.1016/j.cogbrainres.2004.09.006 (cit. on p. 161).

Debener, Stefan et al. (2012). "How about taking a low-cost, small, and wireless EEG for a walk?: EEG to go." en. In: *Psychophysiology* 49.11, pp. 1617–1621. DOI: 10.1111/j.1469-8986.2012.01471.x (cit. on p. 39).

Debener, Stefan et al. (2015). "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear." en. In: *Scientific Reports* 5.1, p. 16743. DOI: 10.1038/srep16743 (cit. on pp. 31, 124, 125, 130, 131).

Delorme, Arnaud and Scott Makeig (2004). "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." eng. In: *Journal of Neuroscience Methods* 134.1, pp. 9–21. DOI: 10.1016/j.jneumeth.2003.10.009 (cit. on p. 131).

Dennett, D. C. (2006). *Sweet dreams: philosophical obstacles to a science of consciousness*. eng. 1. paperback ed., 4. print. The Jean Nicod Lectures. Cambridge, Mass. London: MIT Press (cit. on p. 5).

Deoisres, Suwijak et al. (2023). "Continuous speech with pauses inserted between words increases cortical tracking of speech envelope." eng. In: *PloS One* 18.7, e0289288. DOI: 10.1371/journal.pone.0289288 (cit. on pp. 35, 64, 85, 159).

Desai, Maansi, Alyssa M. Field, and Liberty S. Hamilton (2023). "Dataset size considerations for robust acoustic and phonetic speech encoding models in EEG." English. In: *Frontiers in Human Neuroscience* 16. DOI: 10.3389/fnhum.2022.1001171 (cit. on p. 117).

Desai, Maansi et al. (2021). "Generalizable EEG Encoding Models with Naturalistic Audiovisual Stimuli." en. In: *The Journal of Neuroscience* 41.43, pp. 8946–8962. DOI: 10.1523/JNEUROSCI.2891-20.2021 (cit. on pp. 62, 68, 70, 84, 93, 159, 175).

Di Liberto, Giovanni M., James A. O'Sullivan, and Edmund C. Lalor (2015). "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing." eng. In: *Current biology: CB* 25.19, pp. 2457–2465. DOI: 10.1016/j.cub.2015.08.030 (cit. on pp. 62, 84, 86, 99, 145).

Diedrichsen, Jörn (2020). "Representational Models and the Feature Fallacy." en. In: *The Cognitive Neurosciences*. Ed. by David Poeppel, George R. Mangun, and Michael S. Gazzaniga. 6th ed. The MIT Press, pp. 669–678. DOI: 10.7551/mitpress/11442.003.0074 (cit. on pp. 86, 145, 174, 175).

Diedrichsen, Jörn and Nikolaus Kriegeskorte (2017). "Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis." en. In: *PLOS Computational Biology* 13.4, e1005508. DOI: 10.1371/journal.pcbi.1005508 (cit. on pp. 47, 136, 144).

Diedrichsen, Jörn, Atsushi Yokoi, and Spencer A. Arbuckle (2018). "Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns." In: *NeuroImage*. New advances in encoding and decoding of brain signals 180, pp. 119–133. DOI: 10.1016/j.neuroimage.2017.08.051 (cit. on p. 174).

Ding, Nai and Jonathan Z. Simon (2012a). "Emergence of neural encoding of auditory objects while listening to competing speakers." In: *Proceedings of the National Academy of Sciences* 109.29, pp. 11854–11859. DOI: 10.1073/pnas.1205381109 (cit. on pp. 33, 145).

Ding, Nai and Jonathan Z. Simon (2012b). "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening." en. In: *Journal of Neurophysiology* 107.1, pp. 78–89. DOI: 10.1152/jn.00297.2011 (cit. on pp. 59, 123).

Ding, Nai and Jonathan Z. Simon (2014). "Cortical entrainment to continuous speech: functional roles and interpretations." In: *Frontiers in Human Neuroscience* 8. DOI: 10.3389/fnhum.2014.00311 (cit. on pp. 23, 33, 93).

Ding, Nai et al. (2016). "Cortical tracking of hierarchical linguistic structures in connected speech." en. In: *Nature Neuroscience* 19.1, pp. 158–164. DOI: 10.1038/nn.4186 (cit. on pp. 22, 34).

Diogenes Laertius and Charles Duke Yonge (2006). *The lives and opinions of eminent philosophers*. eng. U.S.A.: Kessinger (cit. on p. iii).

Drennan, Denis P. and Edmund C. Lalor (2019). "Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity." en. In: *eneuro* 6.3, ENEURO.0082–19.2019. DOI: 10.1523/ENEURO.0082-19.2019 (cit. on pp. 37, 42, 48, 64, 79, 84, 94, 102, 104, 115, 117).

Du, Xinyu et al. (2025). "The multifaceted role of the inferior colliculus in sensory prediction, reward processing, and decision-making." In: *eLife* 13. Ed. by Peng Cao and Huan Luo, RP101142. DOI: 10.7554/eLife.101142 (cit. on p. 9).

Ehrhardt, Nina M. et al. (2024). "Comparison of dry and wet electroencephalography for the assessment of cognitive evoked potentials and sensor-level connectivity." English. In: *Frontiers in Neuroscience* 18. DOI: 10.3389/fnins.2024.1441799 (cit. on p. 31).

Eliades, Steven J. and Xiaoqin Wang (2008). "Neural substrates of vocalization feedback monitoring in primate auditory cortex." en. In: *Nature* 453.7198, pp. 1102–1106. DOI: 10.1038/nature06910 (cit. on p. 17).

Escera, Carles and Manuel S. Malmierca (2014). "The auditory novelty system: An attempt to integrate human and animal research." en. In: *Psychophysiology* 51.2, pp. 111–123. DOI: 10.1111/psyp.12156 (cit. on p. 161).

Farina, Almo and Nadia Pieretti (2012). "The soundscape ecology: A new frontier of landscape research and its application to islands and coastal systems." en. In: *Journal of Marine and Island Cultures* 1.1, pp. 21–26. DOI: 10.1016/j.imic.2012.04.002 (cit. on p. 21).

Festen, J. M. and R. Plomp (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing." eng. In: *The Journal of the Acoustical Society of America* 88.4, pp. 1725–1736. DOI: 10.1121/1.400247 (cit. on p. 123).

Fettiplace, Robert and Kyunghee X. Kim (2014). "The Physiology of Mechanoelectrical Transduction Channels in Hearing." en. In: *Physiological Reviews* 94.3, pp. 951–986. DOI: 10.1152/physrev.00038.2013 (cit. on pp. 6, 15).

Fiedler, Lorenz et al. (2019). "Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions." en. In: *NeuroImage* 186, pp. 33–42. DOI: 10.1016/j.neuroimage.2018.10.057 (cit. on pp. 143–146).

Finnegan, William (2015). *Barbarian days: a surfing life*. New York: Penguin Press (cit. on p. 157).

Formisano, Elia (2025). "Understanding real-world audition with computational fMRI." en. In: *Encyclopedia of the Human Brain*. Elsevier, pp. 563–579. DOI: 10.1016/B978-0-12-820480-1.00214-X (cit. on p. 177).

Formisano, Elia et al. (2003). "Mirror-Symmetric Tonotopic Maps in Human Primary Auditory Cortex." In: *Neuron* 40.4, pp. 859–869. DOI: 10.1016/S0896-6273(03)00669-X (cit. on p. 11).

Frescura, Alessia et al. (2025). "Affective responses to upstairs neighbours footsteps sound in a simulated living room: The role of temporal, spatial, and non-acoustic factors." In: *Journal of Building Engineering* 104, p. 112340. DOI: 10.1016/j.jobe.2025.112340 (cit. on p. 23).

Fuglsang, Søren Asp, Torsten Dau, and Jens Hjortkjær (2017). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes." en. In: *NeuroImage* 156, pp. 435–444. DOI: 10.1016/j.neuroimage.2017.04.026 (cit. on pp. 143, 145, 148).

Garrido, Marta I. et al. (2009). "The mismatch negativity: A review of underlying mechanisms." In: *Clinical Neurophysiology* 120.3, pp. 453–463. DOI: 10.1016/j.clinph.2008.11.029 (cit. on pp. 18, 161).

Geirnaert, Simon, Simon L. Kappel, and Preben Kidmose (2025). *A Direct Comparison of Simultaneously Recorded Scalp, Around-Ear, and In-Ear EEG for Neural Selective Auditory Attention Decoding to Speech*. DOI: 10.48550/arXiv.2505.14478 (cit. on pp. 31, 124, 143, 146, 147).

Gemmeke, Jort F. et al. (2017). "Audio Set: An ontology and human-labeled dataset for audio events." In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261 (cit. on p. 160).

Gillis, Marlies et al. (2021). "Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics." en. In: *Journal of Neuroscience* 41.50, pp. 10316–10329. DOI: 10.1523/JNEUROSCI.0812-21.2021 (cit. on p. 79).

Gilzenrat, Mark S. et al. (2010). "Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function." en. In: *Cognitive, Affective, & Behavioral Neuroscience* 10.2, pp. 252–269. DOI: 10.3758/CABN.10.2.252 (cit. on p. 163).

Giordano, Bruno L. et al. (2023). "Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds." en. In: *Nature Neuroscience* 26.4, pp. 664–672. DOI: 10.1038/s41593-023-01285-9 (cit. on pp. 12, 177).

Giraud, Anne-Lise and David Poeppel (2012). "Cortical oscillations and speech processing: emerging computational principles and operations." en. In: *Nature Neuroscience* 15.4, pp. 511–517. DOI: 10.1038/nn.3063 (cit. on pp. 35, 64).

Godey, B. et al. (2001). "Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients." eng. In: *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 112.10, pp. 1850–1859. DOI: 10.1016/s1388-2457(01)00636-8 (cit. on p. 37).

Goodale, M. A. and A. D. Milner (1992). "Separate visual pathways for perception and action." eng. In: *Trends in Neurosciences* 15.1, pp. 20–25. DOI: 10.1016/0166-2236(92)90344-8 (cit. on p. 11).

Gramann, Klaus et al. (2011). "Cognition in action: imaging brain/body dynamics in mobile humans." In: *Reviews in the Neurosciences* 22.6. DOI: 10.1515/RNS.2011.047 (cit. on p. 57).

Greenwood, Donald D. (1990). "A cochlear frequency-position function for several species—29 years later." In: *The Journal of the Acoustical Society of America* 87.6, pp. 2592–2605. DOI: 10.1121/1.399052 (cit. on p. 6).

Grimm, Giso, Joanna Luberadzka, and Volker Hohmann (2019). "A toolbox for rendering virtual acoustic environments in the context of audiology." In: *Acta acustica united with acustica* 105.3, pp. 566–578 (cit. on p. 129).

Grinfeder, Elie et al. (2022). "What Do We Mean by "Soundscape"? A Functional Description." English. In: *Frontiers in Ecology and Evolution* 10. DOI: 10.3389/fevo.2022.894232 (cit. on pp. 21, 23).

Gross, Joachim et al. (2013). "Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain." en. In: *PLOS Biology* 11.12, e1001752. DOI: 10.1371/journal.pbio.1001752 (cit. on p. 33).

Grothe, Benedikt, Michael Pecka, and David McAlpine (2010). "Mechanisms of Sound Localization in Mammals." In: *Physiological Reviews* 90.3, pp. 983–1012. DOI: 10.1152/physrev.00026.2009 (cit. on p. 7).

Gruijters, Stefan L. K. (2022). "Making inferential leaps: Manipulation checks and the road towards strong inference." In: *Journal of Experimental Social Psychology* 98, p. 104251. DOI: 10.1016/j.jesp.2021.104251 (cit. on p. 161).

Guinan, John J. Jr (2006). "Olivocochlear Efferents: Anatomy, Physiology, Function, and the Measurement of Efferent Effects in Humans." en-US. In: *Ear and Hearing* 27.6, p. 589. DOI: 10.1097/01.aud.0000240507.83072.e7 (cit. on pp. 7, 15).

Gutschalk, Alexander and Andrew R. Dykstra (2014). "Functional imaging of auditory scene analysis." In: *Hearing Research*. Human Auditory NeuroImaging 307, pp. 98–110. DOI: 10.1016/j.heares.2013.08.003 (cit. on p. 93).

Hackett, Troy A. (2011). "Information flow in the auditory cortical network." eng. In: *Hearing Research* 271.1-2, pp. 133–146. DOI: 10.1016/j.heares.2010.01.011 (cit. on p. 9).

Hamilton, Liberty S., Erik Edwards, and Edward F. Chang (2018). "A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus." In: *Current Biology* 28.12, 1860–1871.e4. DOI: 10.1016/j.cub.2018.04.033 (cit. on p. 12).

Hamilton, Liberty S. and Alexander G. Huth (2020). "The revolution will not be controlled: natural stimuli in speech neuroscience." en. In: *Language, Cognition and Neuroscience* 35.5, pp. 573–582. DOI: 10.1080/23273798.2018.1499946 (cit. on pp. 24, 59, 170).

Hamilton, Liberty S. et al. (2021). "Parallel and distributed encoding of speech across human auditory cortex." In: *Cell* 184.18, 4626–4639.e13. DOI: 10.1016/j.cell.2021.07.019 (cit. on pp. 13, 57, 94, 144, 167).

Hari, Riitta (2017). *MEG-EEG Primer*. eng. Cary: Oxford University Press USA - OSO (cit. on p. 31).

Haufe, Stefan et al. (2014). "On the interpretation of weight vectors of linear models in multivariate neuroimaging." In: *NeuroImage* 87, pp. 96–110. DOI: 10.1016/j.neuroimage.2013.10.067 (cit. on pp. 46, 72, 145).

Haumann, Niels T. et al. (2021). "Extracting human cortical responses to sound onsets and acoustic feature changes in real music, and their relation to event rate." In: *Brain Research* 1754, p. 147248. DOI: 10.1016/j.brainres.2020.147248 (cit. on p. 37).

Haupt, Thorge, Marc Rosenkranz, and Martin G. Bleichner (2025a). "Exploring Relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes." en. In: *eNeuro* 12.1. DOI: 10.1523/ENEURO.0287-24.2024 (cit. on pp. 45, 49, 55, 217).

Haupt, Thorge, Marc Rosenkranz, and Martin G. Bleichner (2025b). "Neural response attenuates with decreasing inter-onset intervals between sounds in a natural soundscape." en. In: *eNeuro*. DOI: 10.1523/ENEURO.0210-25.2025 (cit. on pp. 91, 217).

Haupt, Thorge, Marc Rosenkranz, and Martin Georg Bleichner (2024). "Exploring relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes." en-us. In: DOI: 10.31234/osf.io/nuy7e (cit. on pp. 101, 134, 144).

Haupt, Thorge et al. (2025). *Enhancing Mobile Brain and Body Imaging: Open-Source Solutions for Real-World Research Applications*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.5433769 (cit. on p. 160).

Hausfeld, Lars et al. (2021). "Cortical processing of distracting speech in noisy auditory scenes depends on perceptual demand." en. In: *NeuroImage* 228, p. 117670. DOI: 10.1016/j.neuroimage.2020.117670 (cit. on p. 143).

Haykin, Simon and Zhe Chen (2005). "The Cocktail Party Problem." In: *Neural Computation* 17.9, pp. 1875–1902. DOI: 10.1162/0899766054322964 (cit. on p. 123).

Hebart, Martin N. and Chris I. Baker (2018). "Deconstructing multivariate decoding for the study of brain function." In: *NeuroImage*. New advances in encoding and decoding of brain signals 180, pp. 4–18. DOI: 10.1016/j.neuroimage.2017.08.005 (cit. on p. 46).

Hecox, Kurt and Robert Galambos (1974). "Brain Stem Auditory Evoked Responses in Human Infants and Adults." In: *Archives of Otolaryngology* 99.1, pp. 30–33. DOI: 10.1001/archotol.1974.00780030034006 (cit. on p. 27).

Heer, Wendy A. de et al. (2017). "The Hierarchical Cortical Organization of Human Speech Processing." en. In: *The Journal of Neuroscience* 37.27, pp. 6539–6557. DOI: 10.1523/JNEUROSCI.3267-16.2017 (cit. on pp. 57, 62, 70, 159, 175).

Heidlmayr, Karin, Maria Kihlstedt, and Frédéric Isel (2020). "A review on the electroencephalography markers of Stroop executive control processes." en. In: *Brain and Cognition* 146, p. 105637. DOI: 10.1016/j.bandc.2020.105637 (cit. on p. 36).

Heil, Peter (2004). "First-spike latency of auditory neurons revisited." In: *Current Opinion in Neurobiology* 14.4, pp. 461–467. DOI: 10.1016/j.conb.2004.07.002 (cit. on pp. 7, 15).

Helmholtz, H. (1853). "Ueber einige Gesetze der Vertheilung elektrischer Ströme in körperlichen Leitern mit Anwendung auf die thierisch-elektrischen Versuche." In: *Annalen der Physik* 165, pp. 211–233. DOI: 10.1002/andp.18531650603 (cit. on p. 29).

Herrmann, Björn (2024). "Minimal background noise enhances neural speech tracking: Evidence of stochastic resonance." en. In: *eLife* 13. DOI: 10.7554/eLife.100830.1 (cit. on pp. 123, 143).

Herrmann, Björn, Nadine Schlichting, and Jonas Obleser (2014). "Dynamic Range Adaptation to Spectral Stimulus Statistics in Human Auditory Cortex." en. In: *Journal of Neuroscience* 34.1, pp. 327–331. DOI: 10.1523/JNEUROSCI.3974-13.2014 (cit. on pp. 18, 93).

Herrmann, Björn et al. (2013). "Auditory filter width affects response magnitude but not frequency specificity in auditory cortex." In: *Hearing Research* 304, pp. 128–136. DOI: 10.1016/j.heares.2013.07.005 (cit. on pp. 93, 116).

Herrmann, Björn et al. (2016). "Altered temporal dynamics of neural adaptation in the aging human auditory cortex." en. In: *Neurobiology of Aging* 45, pp. 10–22. DOI: 10.1016/j.neurobiolaging.2016.05.006 (cit. on pp. 94, 106, 112–114, 116).

Hershey, Shawn et al. (2017). *CNN Architectures for Large-Scale Audio Classification*. DOI: 10.48550/arXiv.1609.09430 (cit. on p. 160).

Hicks, Jarrod M. and Josh H. McDermott (2024). "Noise schemas aid hearing in noise." In: *Proceedings of the National Academy of Sciences* 121.47, e2408995121. DOI: 10.1073/pnas.2408995121 (cit. on pp. 17, 113, 116).

Hillyard, S. A. et al. (1973). "Electrical signs of selective attention in the human brain." eng. In: *Science (New York, N.Y.)* 182.4108, pp. 177–180. DOI: 10.1126/science.182.4108.177 (cit. on p. 37).

Hjortkjær, Jens et al. (2024). *Real-Time Control of a Hearing Instrument with EEG-based Attention Decoding*. preprint. Neuroscience. DOI: 10.1101/2024.03.01.582668 (cit. on p. 148).

Holdgraf, Christopher R. et al. (2017). "Encoding and Decoding Models in Cognitive Electrophysiology." In: *Frontiers in Systems Neuroscience* 11, p. 61. DOI: 10.3389/fnsys.2017.00061 (cit. on pp. 42, 57, 59, 93, 94, 136, 143–145, 174).

Holleman, Gijs A. et al. (2020). "The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity." English. In: *Frontiers in Psychology* 11. DOI: 10.3389/fpsyg.2020.00721 (cit. on pp. 24, 171).

Holtze, Björn et al. (2021). "Are They Calling My Name? Attention Capture Is Reflected in the Neural Tracking of Attended and Ignored Speech." In: *Frontiers in Neuroscience* 15, p. 643705. DOI: 10.3389/fnins.2021.643705 (cit. on pp. 57, 143).

Holtze, Björn et al. (2022). "Ear-EEG Measures of Auditory Attention to Continuous Speech." In: *Frontiers in Neuroscience* 16 (cit. on pp. 64, 124).

Holtze, Björn et al. (2023). "Eye-Blink Patterns Reflect Attention to Continuous Speech." en. In: *Advances in Cognitive Psychology* 19.2, pp. 177–200. DOI: 10.5709/acp-0387-6 (cit. on pp. 30, 166).

Hotelling, Harold (1936). "Relations Between Two Sets of Variates." In: *Biometrika* 28.3/4, p. 321. DOI: 10.2307/2333955 (cit. on p. 29).

Howard, Mary F. and David Poeppel (2010). "Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension." en. In: *Journal of Neurophysiology* 104.5, pp. 2500–2511. DOI: 10.1152/jn.00251.2010 (cit. on pp. 33, 64, 93).

Huang, Nicholas and Mounya Elhilali (2017). "Auditory salience using natural soundscapes." In: *The Journal of the Acoustical Society of America* 141.3, pp. 2163–2176. DOI: 10.1121/1.4979055 (cit. on pp. 22, 85).

Huang, Nicholas, Malcolm Slaney, and Mounya Elhilali (2018). "Connecting Deep Neural Networks to Physical, Perceptual, and Electrophysiological Auditory Signals." English. In: *Frontiers in Neuroscience* 12. DOI: 10.3389/fnins.2018.00532 (cit. on p. 160).

Hudspeth, A. J. (1985). "The cellular basis of hearing: the biophysics of hair cells." eng. In: *Science (New York, N.Y.)* 230.4727, pp. 745–752. DOI: 10.1126/science.2414845 (cit. on p. 6).

Hyvärinen, A. and E. Oja (2000). "Independent component analysis: algorithms and applications." en. In: *Neural Networks* 13.4-5, pp. 411–430. DOI: 10.1016/S0893-6080(00)00026-5 (cit. on pp. 29, 30).

Hölle, Daniel and Martin G. Bleichner (2023). "Smartphone-based ear-electroencephalography to study sound processing in everyday life." In: *European Journal of Neuroscience* 58.7, pp. 3671–3685. DOI: 10.1111/ejn.16124 (cit. on pp. 39, 57, 62, 85, 95, 118, 124, 162).

Hölle, Daniel, Joost Meekes, and Martin G. Bleichner (2021). "Mobile ear-EEG to study auditory attention in everyday life: Auditory attention in everyday life." en. In: *Behavior Research Methods* 53.5, pp. 2025–2036. DOI: 10.3758/s13428-021-01538-0 (cit. on p. 118).

Hölle, Daniel et al. (2022). "Real-Time Audio Processing of Real-Life Soundscapes for EEG Analysis: ERPs Based on Natural Sound Onsets." In: *Frontiers in Neuroergonomics* 3. DOI: 10.3389/fnrgo.2022.793061 (cit. on pp. 86, 87, 124).

ISO 12913-1:2014(en), Acoustics — Soundscape — Part 1: Definition and conceptual framework (2014) (cit. on p. 21).

Jacobsen, Nadine Svenja Josée et al. (2021). "A Walk in the Park? Characterizing Gait-related Artifacts in Mobile EEG Recordings." In: *European Journal of Neuroscience* 54.12. Ed. by T. Solis-Escalante, pp. 8421–8440. DOI: 10.1111/ejn.14965 (cit. on p. 166).

Jacobsen, Nadine Svenja Josée et al. (2022). "Mobile Electroencephalography Captures Differences of Walking over Even and Uneven Terrain but Not of Single and Dual-Task Gait." In: *Frontiers in Sports and Active Living* 4. DOI: 10.3389/fspor.2022.945341 (cit. on pp. 36, 140).

Jaeger, Manuela et al. (2020). "Decoding the Attended Speaker From EEG Using Adaptive Evaluation Intervals Captures Fluctuations in Attentional Listening." In: *Fron-*

tiers in Neuroscience 14, p. 603. DOI: 10.3389/fnins.2020.00603 (cit. on pp. 31, 33, 37, 123, 145–147).

Jeung, Sein et al. (2023). "Virtual Reality for Spatial Navigation." en. In: *Virtual Reality in Behavioral Neuroscience: New Insights and Methods*. Ed. by Christopher Maymon, Gina Grimshaw, and Ying Choon Wu. Cham: Springer International Publishing, pp. 103–129. DOI: 10.1007/7854_2022_403 (cit. on p. 170).

Johnson, Dominic DP, Pavel Stopka, and Josh Bell (2002). "Individual variation evades the Prisoner's Dilemma." In: *BMC Evolutionary Biology* 2, p. 15. DOI: 10.1186/1471-2148-2-15 (cit. on p. 170).

Joris, P. X., C. E. Schreiner, and A. Rees (2004). "Neural Processing of Amplitude-Modulated Sounds." In: *Physiological Reviews* 84.2, pp. 541–577. DOI: 10.1152/physrev.00029.2003 (cit. on pp. 7, 12).

Kanai, Ryota and Naotsugu Tsuchiya (2012). "Qualia." English. In: *Current Biology* 22.10, R392–R396. DOI: 10.1016/j.cub.2012.03.033 (cit. on p. 161).

Kandel, Eric R. et al., eds. (2021). *Principles of neural science*. en. Sixth edition. New York: McGraw Hill (cit. on pp. 6–8).

Kasten, Florian H., Quentin Busson, and Benedikt Zoefel (2023). *Opposing neural processing modes alternate rhythmically during sustained auditory attention*. en. DOI: 10.1101/2023.10.04.560684 (cit. on p. 162).

Kasten, Florian H., Quentin Busson, and Benedikt Zoefel (2024). "Opposing neural processing modes alternate rhythmically during sustained auditory attention." en. In: *Communications Biology* 7.1, p. 1125. DOI: 10.1038/s42003-024-06834-x (cit. on p. 177).

Kaya, İbrahim (2021). "A Brief Summary of EEG Artifact Handling." en. In: *Brain-Computer Interface*. IntechOpen. DOI: 10.5772/intechopen.99127 (cit. on p. 30).

Kayser, H. et al. (2009). "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses." en. In: *EURASIP Journal on Advances in Signal Processing* 2009.1, p. 298605. DOI: 10.1155/2009/298605 (cit. on pp. 60, 115, 129).

Kayser, Hendrik et al. (2019). "Open Master Hearing Aid (openMHA)—An Integrated Platform for Hearing Aid Research." In: *The Journal of the Acoustical Society of America* 146.4, pp. 2879–2879. DOI: 10.1121/1.5136988 (cit. on p. 130).

Kayser, Stephanie J., Steven W. McNair, and Christoph Kayser (2016). "Prestimulus influences on auditory perception from sensory representations and decision processes." In: *Proceedings of the National Academy of Sciences* 113.17, pp. 4842–4847. DOI: 10.1073/pnas.1524087113 (cit. on p. 162).

Khalighinejad, Bahar et al. (2019). "Adaptation of the human auditory cortex to changing background noise." en. In: *Nature Communications* 10.1, p. 2509. DOI: 10.1038/s41467-019-10611-4 (cit. on p. 17).

King, Andrew J., Sundeep Teki, and Ben D. B. Willmore (2018). *Recent advances in understanding the auditory cortex*. en. DOI: 10.12688/f1000research.15580.1 (cit. on p. 17).

Klimesch, Wolfgang (2012). "Alpha-band oscillations, attention, and controlled access to stored information." In: *Trends in Cognitive Sciences* 16.12, pp. 606 –617. DOI: https://doi.org/10.1016/j.tics.2012.10.007 (cit. on p. 162).

Klimesch, Wolfgang, Paul Sauseng, and Simon Hanslmayr (2007). "EEG alpha oscillations: the inhibition-timing hypothesis." eng. In: *Brain Research Reviews* 53.1, pp. 63–88. DOI: 10.1016/j.brainresrev.2006.06.003 (cit. on p. 162).

Klug, M. et al. (2022). *The BeMoBIL Pipeline for automated analyses of multimodal mobile brain and body imaging data*. en. DOI: 10.1101/2022.09.29.510051 (cit. on p. 166).

Klug, Marius and Klaus Gramann (2021). "Identifying Key Factors for Improving ICA-based Decomposition of EEG Data in Mobile and Stationary Experiments." In: *European Journal of Neuroscience* 54.12, pp. 8406–8420. DOI: 10.1111/ejn.14992 (cit. on p. 166).

Kluger, Daniel S et al. (2021). "Respiration aligns perception with neural excitability." en. In: *eLife* 10, e70907. DOI: 10.7554/eLife.70907 (cit. on p. 164).

Kohl, Carmen, Tiina Parviainen, and Stephanie R. Jones (2022). "Neural Mechanisms Underlying Human Auditory Evoked Responses Revealed By Human Neocortical Neurosolver." en. In: *Brain Topography* 35.1, pp. 19–35. DOI: 10.1007/s10548-021-00838-0 (cit. on p. 37).

Korte, Silvia, Thorge Haupt, and Martin G. Bleichner (2025). "EEG Signatures of Auditory Distraction: Neural Responses to Spectral Novelty in Real-World Soundscapes." en. In: *eNeuro* 12.7. DOI: 10.1523/ENEURO.0154-25.2025 (cit. on pp. 118, 160).

Kosmyna, Nataliya et al. (2019). "AttentivU: Designing EEG and EOG Compatible Glasses for Physiological Sensing and Feedback in the Car." en. In: *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Utrecht Netherlands: ACM, pp. 355–368. DOI: 10.1145/3342197.3344516 (cit. on p. 31).

Kothe, Christian et al. (2025). "The lab streaming layer for synchronized multimodal recording." en. In: *Imaging Neuroscience* 3, IMAG.a.136. DOI: 10.1162/IMAG.a.136 (cit. on p. 130).

Krakauer, John W. et al. (2017). "Neuroscience Needs Behavior: Correcting a Reductionist Bias." In: *Neuron* 93.3, pp. 480–490. DOI: 10.1016/j.neuron.2016.12.041 (cit. on pp. 170, 174).

Kriegeskorte, Nikolaus and Pamela K Douglas (2019). "Interpreting encoding and decoding models." In: *Current Opinion in Neurobiology*. Machine Learning, Big Data, and Neuroscience 55, pp. 167–179. DOI: 10.1016/j.conb.2019.04.002 (cit. on pp. 46, 47, 72, 94, 174, 175).

Kulasingham, Joshua P. et al. (2024). "Level-Dependent Subcortical Electroencephalography Responses to Continuous Speech." en. In: *eNeuro* 11.8. DOI: 10.1523/ENEURO.0135-24.2024 (cit. on p. 27).

Ladouce, Simon, Magda Mustile, and Frédéric Dehais (2021). "Capturing Cognitive Events Embedded in the Real-World Using Mobile EEG and Eye-Tracking." In: *bioRxiv*. DOI: 10.1101/2021.11.30.470560 (cit. on p. 93).

Lalor, Edmund C. and John J. Foxe (2010). "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution." eng. In: *The European Journal of Neuroscience* 31.1, pp. 189–193. DOI: 10.1111/j.1460-9568.2009.07055.x (cit. on p. 132).

Lalor, Edmund C. et al. (2009). "Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli." In: *Journal of Neurophysiology* 102.1, pp. 349–359. DOI: 10.1152/jn.90896.2008 (cit. on pp. 85, 93, 123).

Lanting, Cornelis P. et al. (2013). "Mechanisms of adaptation in human auditory cortex." In: *Journal of Neurophysiology* 110.4, pp. 973–983. DOI: 10.1152/jn.00547.2012 (cit. on pp. 17, 36, 37, 112, 114, 116).

Larson, Eric and Adrian K.C. Lee (2013). "The cortical dynamics underlying effective switching of auditory spatial attention." en. In: *NeuroImage* 64, pp. 365–370. DOI: 10.1016/j.neuroimage.2012.09.006 (cit. on pp. 24, 123).

Lartillot, Olivier, Petri Toiviainen, and Tuomas Eerola (2008). "A Matlab Toolbox for Music Information Retrieval." In: *Data Analysis, Machine Learning and Applications*. Ed. by Christine Preisach et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 261–268. DOI: 10.1007/978-3-540-78246-9_31 (cit. on p. 65).

Laszlo, Sarah et al. (2014). "A direct comparison of active and passive amplification electrodes in the same amplifier system." en. In: *Journal of Neuroscience Methods* 235, pp. 298–307. DOI: 10.1016/j.jneumeth.2014.05.012 (cit. on p. 31).

Lee, Adrian K. C. et al. (2014). "Using neuroimaging to understand the cortical mechanisms of auditory selective attention." In: *Hearing Research*. Human Auditory NeuroImaging 307, pp. 111–120. DOI: 10.1016/j.heares.2013.06.010 (cit. on p. 93).

Lei, Xu and Keren Liao (2017). "Understanding the Influences of EEG Reference: A Large-Scale Brain Network Perspective." English. In: *Frontiers in Neuroscience* 11. DOI: 10.3389/fnins.2017.00205 (cit. on p. 31).

Lindboom, Elsa et al. (2023). "Incorporating models of subcortical processing improves the ability to predict EEG responses to natural speech." In: *Hearing Research* 433, p. 108767. DOI: 10.1016/j.heares.2023.108767 (cit. on p. 159).

Lohse, Michael et al. (2020). "Neural circuits underlying auditory contrast gain control and their perceptual implications." en. In: *Nature Communications* 11.1, p. 324. DOI: 10.1038/s41467-019-14163-5 (cit. on p. 16).

Lorenzi, Christian et al. (2023). "Human Auditory Ecology: Extending Hearing Research to the Perception of Natural Soundscapes by Humans in Rapidly Changing Environments." EN. In: *Trends in Hearing* 27, p. 23312165231212032. DOI: 10.1177/23312165231212032 (cit. on p. 170).

Lu, Hao and W. Owen Brimijoin (2022). "Sound Source Selection Based on Head Movements in Natural Group Conversation." In: *Trends in Hearing* 26, p. 233121652210977. DOI: 10.1177/23312165221097789 (cit. on p. 12).

Luck, Steven J. and Emily S. Kappenman (2011). *ERP Components and Selective Attention*. Oxford University Press. DOI: 10.1093/oxfordhb/9780195374148.013.0144 (cit. on pp. 36, 37).

Luo, Huan and David Poeppel (2007). "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex." In: *Neuron* 54.6, pp. 1001–1010. DOI: 10.1016/j.neuron.2007.06.004 (cit. on pp. 33, 64).

Lutzenberger, Werner, Friedemann Pulvermüller, and Niels Birbaumer (1994). "Words and pseudowords elicit distinct patterns of 30-Hz EEG responses in humans." In: *Neuroscience Letters* 176.1, pp. 115–118. DOI: 10.1016/0304-3940(94)90884-2 (cit. on p. 93).

López-Caballero, F (2025). "N1 facilitation at short Inter-Stimulus-Interval (ISI) occurs under 400 ms and is dependent on ISI from previous sounds: Evidence using an unpredictable auditory stimulation sequence." en. In: (cit. on pp. 17, 113).

López-Caballero, Fran et al. (2023). "Intensity and inter-stimulus-interval effects on human middle- and long-latency auditory evoked potentials in an unpredictable auditory context." en. In: *Psychophysiology* 60.4, e14217. DOI: 10.1111/psyp.14217 (cit. on pp. 37, 93, 104, 105, 113, 115).

MacLean, Jessica et al. (2024). "Short- and long-term neuroplasticity interact during the perceptual learning of concurrent speech." In: *Cerebral Cortex* 34.2, bhad543. DOI: 10.1093/cercor/bhad543 (cit. on p. 38).

Mahajan, Yatin, Varghese Peter, and Mridula Sharma (2017). "Effect of EEG Referencing Methods on Auditory Mismatch Negativity." English. In: *Frontiers in Neuroscience* 11. DOI: 10.3389/fnins.2017.00560 (cit. on p. 31).

Makeig, Scott et al. (1995). "Independent component analysis of electroencephalographic data." In: *Advances in neural information processing systems* 8 (cit. on pp. 29, 30).

Makin, Joseph G., David A. Moses, and Edward F. Chang (2020). "Machine translation of cortical activity to text with an encoder–decoder framework." en. In: *Nature Neuroscience* 23.4, pp. 575–582. DOI: 10.1038/s41593-020-0608-8 (cit. on p. 167).

Malmierca, Manuel S. (2003). "The structure and physiology of the rat auditory system: an overview." eng. In: *International Review of Neurobiology* 56, pp. 147–211. DOI: 10.1016/s0074-7742(03)56005-6 (cit. on p. 9).

Marmarelis, Vasilis Z. (2004). *Nonlinear Dynamic Modeling of Physiological Systems*. en. 1st ed. Wiley. DOI: 10.1002/9780471679370 (cit. on p. 42).

Mathewson, Kyle E. et al. (2009). "To See or Not to See: Prestimulus Phase Predicts Visual Awareness." en. In: *Journal of Neuroscience* 29.9, pp. 2725–2732. DOI: 10.1523/JNEUROSCI.3963-08.2009 (cit. on p. 162).

May, Patrick J. C. and Hannu Tiitinen (2010). "Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained." en. In: *Psychophysiology* 47.1, pp. 66–122. DOI: 10.1111/j.1469-8986.2009.00856.x (cit. on pp. 18, 36, 37, 93, 112, 113).

McDermott, Josh H. (2009). "The cocktail party problem." English. In: *Current Biology* 19.22, R1024–R1027. DOI: 10.1016/j.cub.2009.09.005 (cit. on p. 123).

Mcgurk, Harry and John Macdonald (1976). "Hearing lips and seeing voices." en. In: *Nature* 264.5588, pp. 746–748. DOI: 10.1038/264746a0 (cit. on p. 23).

Meiser, Arnd, Anna Lena Knoll, and Martin G. Bleichner (2024). "High-density ear-EEG for understanding ear-centered EEG." en. In: *Journal of Neural Engineering* 21.1, p. 016001. DOI: 10.1088/1741-2552/ad1783 (cit. on p. 31).

Meiser, Arnd et al. (2020). "The Sensitivity of Ear-EEG: Evaluating the Source-Sensor Relationship Using Forward Modeling." In: *Brain Topography* 33.6, pp. 665–676. DOI: 10.1007/s10548-020-00793-2 (cit. on p. 146).

Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen (2016). "Metrics for Polyphonic Sound Event Detection." en. In: *Applied Sciences* 6.6, p. 162. DOI: 10.3390/app6060162 (cit. on p. 160).

Mesgarani, Nima et al. (2009). "Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex." In: *Journal of Neurophysiology* 102.6, pp. 3329–3339. DOI: 10.1152/jn.91128.2008 (cit. on pp. 45, 67, 99).

Mesik, Juraj and Magdalena Wojtczak (2023). "The effects of data quantity on performance of temporal response function analyses of natural speech processing." English. In: *Frontiers in Neuroscience* 16. DOI: 10.3389/fnins.2022.963629 (cit. on pp. 62, 79, 87, 117).

Metzger, Sean L. et al. (2023). "A high-performance neuroprosthesis for speech decoding and avatar control." en. In: *Nature* 620.7976, pp. 1037–1046. DOI: 10.1038/s41586-023-06443-4 (cit. on p. 167).

Michel, Christoph M. and Denis Brunet (2019). "EEG Source Imaging: A Practical Review of the Analysis Steps." English. In: *Frontiers in Neurology* 10. DOI: 10.3389/fneur.2019.00325 (cit. on p. 37).

Mikkelsen, Kaare B. et al. (2021). "EEGs Vary Less Between Lab and Home Locations Than They Do Between People." In: *Frontiers in Computational Neuroscience* 15, p. 565244. DOI: 10.3389/fncom.2021.565244 (cit. on p. 31).

Mirkovic, Bojana et al. (2015). "Decoding the Attended Speech Stream with Multi-Channel EEG: Implications for Online, Daily-Life Applications." In: *Journal of Neural Engineering* 12.4, p. 046007. DOI: 10.1088/1741-2560/12/4/046007 (cit. on pp. 31, 118, 123, 124, 139, 143, 147).

Mirkovic, Bojana et al. (2016). "Target Speaker Detection with Concealed EEG Around the Ear." In: *Frontiers in Neuroscience* 10. DOI: 10.3389/fnins.2016.00349 (cit. on pp. 64, 124, 148).

Mirkovic, Bojana et al. (2019). "Effects of Directional Sound Processing and Listener's Motivation on EEG Responses to Continuous Noisy Speech: Do Normal-Hearing and Aided Hearing-Impaired Listeners Differ?" In: *Hearing Research* 377, pp. 260–270. DOI: 10.1016/j.heares.2019.04.005 (cit. on pp. 69, 123, 148).

Moerel, Michelle, Federico De Martino, and Elia Formisano (2014). "An anatomical and functional topography of human auditory cortical areas." English. In: *Frontiers in Neuroscience* 8. DOI: 10.3389/fnins.2014.00225 (cit. on p. 11).

Muncke, Jan, Ivine Kuruvila, and Ulrich Hoppe (2022). "Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise." In: *Frontiers in Neuroscience* 16 (cit. on p. 143).

Murphy, Sandra, Charles Spence, and Polly Dalton (2017). "Auditory perceptual load: A review." en. In: *Hearing Research* 352, pp. 40–48. DOI: 10.1016/j.heares.2017.02.005 (cit. on p. 143).

Mustafa, Mishal and Avinash Krishnamurthy (2025). "A comprehensive review on real-world challenges faced by hearing aid users and innovative solutions for background noise—from lab to life." In: *The Egyptian Journal of Otolaryngology* 41.1, p. 62. DOI: 10.1186/s43163-025-00814-6 (cit. on p. 123).

Müller, Meinard (2021). *Fundamentals of Music Processing: Using Python and Jupyter Notebooks*. en. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-69808-9 (cit. on pp. 63, 100).

Młynarski, Wiktor and Josh H. McDermott (2019). "Ecological origins of perceptual grouping principles in the auditory system." In: *Proceedings of the National Academy of Sciences* 116.50, pp. 25355–25364. DOI: 10.1073/pnas.1903887116 (cit. on p. 22).

Nastase, Samuel A., Ariel Goldstein, and Uri Hasson (2020). "Keep it real: rethinking the primacy of experimental control in cognitive neuroscience." In: *NeuroImage* 222, p. 117254. DOI: 10.1016/j.neuroimage.2020.117254 (cit. on p. 161).

Niedermeyer, Ernst and F. H. Lopes da Silva (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. en. Lippincott Williams & Wilkins (cit. on p. 33).

Norman-Haignere, Sam, Nancy G. Kanwisher, and Josh H. McDermott (2015). "Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition." In: *Neuron* 88.6, pp. 1281–1296. DOI: 10.1016/j.neuron.2015.11.035 (cit. on p. 12).

Norman-Haignere, Sam V. et al. (2022). "Multiscale temporal integration organizes hierarchical computation in human auditory cortex." en. In: *Nature Human Behaviour* 6.3, pp. 455–469. DOI: 10.1038/s41562-021-01261-y (cit. on p. 12).

Nourifard, Mahan (2025). *A Survey on Single-Channel Blind Source Separation in Communications*. DOI: 10.36227/techrxiv.174494865.59517175/v1 (cit. on pp. 23, 160).

Nunez, Paul L. and Ramesh Srinivasan (2006). *Electric Fields of the Brain*. Oxford University Press. DOI: 10.1093/acprof:oso/9780195050387.001.0001 (cit. on p. 37).

Näätänen, R. and T. Picton (1987). "The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure." eng. In: *Psychophysiology* 24.4, pp. 375–425. DOI: 10.1111/j.1469-8986.1987.tb00311.x (cit. on pp. 93, 113).

Näätänen, Risto (1990). "The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function." en. In: *Behavioral and Brain Sciences* 13.2, pp. 201–233. DOI: 10.1017/S0140525X00078407 (cit. on p. 173).

Näätänen, Risto (1992). *Attention and brain function*. Attention and brain function. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc (cit. on p. 161).

Näätänen, Risto (2001). "The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm)."

en. In: *Psychophysiology* 38.1, pp. 1–21. DOI: 10.1111/1469-8986.3810001 (cit. on pp. 18, 37, 93).

Obleser, Jonas and Christoph Kayser (2019). "Neural Entrainment and Attentional Selection in the Listening Brain." English. In: *Trends in Cognitive Sciences* 23.11, pp. 913–926. DOI: 10.1016/j.tics.2019.08.004 (cit. on pp. 23, 34, 36, 57, 143, 173).

Oertel, Donata and Eric D. Young (2004). "What's a cerebellar circuit doing in the auditory system?" In: *Trends in Neurosciences* 27.2, pp. 104–110. DOI: 10.1016/j.tins.2003.12.001 (cit. on p. 8).

Oganian, Yulia and Edward F. Chang (2019). "A speech envelope landmark for syllable encoding in human superior temporal gyrus." In: *Science Advances* 5.11, eaay6279. DOI: 10.1126/sciadv.aay6279 (cit. on pp. 64, 167).

Oganian, Yulia et al. (2023). "Phase Alignment of Low-Frequency Neural Activity to the Amplitude Envelope of Speech Reflects Evoked Responses to Acoustic Edges, Not Oscillatory Entrainment." en. In: *Journal of Neuroscience* 43.21, pp. 3909–3921. DOI: 10.1523/JNEUROSCI.1663-22.2023 (cit. on pp. 22, 35, 36, 85, 159, 173).

Ogg, Mattson and L. Robert Slevc (2019). "Acoustic Correlates of Auditory Object and Event Perception: Speakers, Musical Timbres, and Environmental Sounds." English. In: *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.01594 (cit. on p. 22).

Okamoto, H. et al. (2004). "N1m recovery from decline after exposure to noise with strong spectral contrasts." In: *Hearing Research*. 40th Inner Ear Biology (IEB) Workshop 196.1, pp. 77–86. DOI: 10.1016/j.heares.2004.04.017 (cit. on p. 112).

Oostenveld, R. and P. Praamstra (2001). "The five percent electrode system for high-resolution EEG and ERP measurements." eng. In: *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 112.4, pp. 713–719. DOI: 10.1016/s1388-2457(00)00527-7 (cit. on p. 31).

Oostenveld, Robert et al. (2011). "FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data." In: *Computational Intelligence and Neuroscience* 2011, p. 156869. DOI: 10.1155/2011/156869 (cit. on p. 69).

Orne, Martin T. (1962). "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." In: *American Psychologist* 17.11, pp. 776–783. DOI: 10.1037/h0043424 (cit. on p. 170).

O'Sullivan, James A. et al. (2015). "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG." In: *Cerebral Cortex* 25.7, pp. 1697–1706. DOI: 10.1093/cercor/bht355 (cit. on pp. 123, 147).

Pantev, C. et al. (1988). "Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields." eng. In: *Electroencephalography and Clinical Neurophysiology* 69.2, pp. 160–170. DOI: 10.1016/0013-4694(88)90211-8 (cit. on p. 38).

Pantev, C. et al. (1989). "Neuromagnetic evidence of an amplitopic organization of the human auditory cortex." In: *Electroencephalography and Clinical Neurophysiology* 72.3, pp. 225–231. DOI: 10.1016/0013-4694(89)90247-2 (cit. on p. 38).

Pantev, C et al. (1995). "Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings." In: *Electroencephalography and Clinical Neurophysiology* 94.1, pp. 26–40. DOI: 10.1016/0013-4694(94)00209-4 (cit. on p. 37).

Panzeri, Stefano et al. (2015). "Neural population coding: combining insights from microscopic and mass signals." eng. In: *Trends in Cognitive Sciences* 19.3, pp. 162–172. DOI: 10.1016/j.tics.2015.01.002 (cit. on p. 174).

Parras, Gloria G. et al. (2017). "Neurons along the auditory pathway exhibit a hierarchical organization of prediction error." en. In: *Nature Communications* 8.1, p. 2148. DOI: 10.1038/s41467-017-02038-6 (cit. on pp. 9, 16).

Pavlovic, Caslav et al. (2018). "Open portable platform for hearing aid research." en. In: *The Journal of the Acoustical Society of America* 143.3_Supplement, pp. 1738–1738. DOI: 10.1121/1.5035670 (cit. on p. 124).

Pearce, David and Hans-Günter Hirsch (2000). "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions." In: *6th International Conference on Spoken Language Processing (ICSLP 2000)*. ISCA, vol. 4, 29–32–0. DOI: 10.21437/ICSLP.2000-743 (cit. on p. 29).

Peelle, Jonathan E. (2018). "Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior." en-US. In: *Ear and Hearing* 39.2, p. 204. DOI: 10.1097/AUD.0000000000000494 (cit. on p. 163).

Peters, T. M. (1998). "Introduction to the Fourier Transform." en. In: *The Fourier Transform in Biomedical Engineering*. Ed. by Terry M. Peters and Jackie Williams. Boston, MA: Birkhäuser, pp. 1–24. DOI: 10.1007/978-1-4612-0637-8_1 (cit. on p. 33).

Petersen, Eline Borch et al. (2017). "Neural tracking of attended versus ignored speech is differentially affected by hearing loss." In: *Journal of Neurophysiology* 117.1, pp. 18–27. DOI: 10.1152/jn.00527.2016 (cit. on p. 64).

Pfurtscheller, G. (1992). "Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest." eng. In: *Electroencephalography and Clinical Neurophysiology* 83.1, pp. 62–69. DOI: 10.1016/0013-4694(92)90133-3 (cit. on p. 36).

Pion-Tonachini, Luca, Ken Kreutz-Delgado, and Scott Makeig (2019). "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website." en. In: *NeuroImage* 198, pp. 181–197. DOI: 10.1016/j.neuroimage.2019.05.026 (cit. on p. 30).

Plack, Christopher J. (2023). *The Sense of Hearing*. 4th ed. London: Routledge. DOI: 10.4324/9781003303329 (cit. on p. 6).

Poeppel, David and Federico Adolfi (2020). "Against the Epistemological Primacy of the Hardware: The Brain from Inside Out, Turned Upside Down." en. In: *eNeuro* 7.4. DOI: 10.1523/ENEURO.0215-20.2020 (cit. on p. 42).

Popov, Vencislav, Markus Ostarek, and Caitlin Tenison (2018). "Practices and pitfalls in inferring neural representations." In: *NeuroImage* 174, pp. 340–351. DOI: 10.1016/j.neuroimage.2018.03.041 (cit. on p. 72).

Popper, Karl R. and Karl Raimund Poper (1991). *The open universe: an argument for indeterminism*. The Postscript to The logic of scientific discovery. London: Routledge (cit. on p. 91).

Pratchett, Terry (1996). *Diggers*. eng. London: Corgi Books (cit. on p. 121).

Pratchett, Terry (2005). *A hat full of sky: a story of Discworld*. eng. London: Corgi Books (cit. on p. 177).

Pratchett, Terry (2008). *Interesting times: a novel of Discworld*. eng. New York, N.Y.: Harper (cit. on p. 41).

Pratchett, Terry (2009). *Equal Rites*. eng. Place of publication not identified: HarperCollins (cit. on p. 153).

Pratchett, Terry (2010). "The discworld series. 13: Small gods / Terry Pratchett." eng. In: Corgi ed., [Nachdr.] A Corgi book. London: Corgi Books (cit. on p. 25).

Proverbio, Alice Mado, Sacha Santoni, and Roberta Adorni (2020). "ERP Markers of Valence Coding in Emotional Speech Processing." In: *iScience* 23.3, p. 100933. DOI: 10.1016/j.isci.2020.100933 (cit. on p. 36).

Puschmann, Sebastian et al. (2019). "Hearing-Impaired Listeners Show Increased Audiovisual Benefit When Listening to Speech in Noise." In: *NeuroImage* 196, pp. 261–268. DOI: 10.1016/j.neuroimage.2019.04.017 (cit. on p. 23).

Puvvada, Krishna C. and Jonathan Z. Simon (2017). "Cortical Representations of Speech in a Multitalker Auditory Scene." en. In: *Journal of Neuroscience* 37.38, pp. 9189–9196. DOI: 10.1523/JNEUROSCI.0938-17.2017 (cit. on p. 145).

Rahman, Monzilur et al. (2020). "Simple transformations capture auditory input to cortex." en. In: *Proceedings of the National Academy of Sciences* 117.45, pp. 28442–28451. DOI: 10.1073/pnas.1922033117 (cit. on pp. 48, 84, 93, 160).

Rauschecker, J. P. and B. Tian (2000). "Mechanisms and streams for processing of "what" and "where" in auditory cortex." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.22, pp. 11800–11806. DOI: 10.1073/pnas.97.22.11800 (cit. on p. 11).

Regev, Tamar I et al. (2021). "Context Sensitivity across Multiple Time scales with a Flexible Frequency Bandwidth." en. In: *Cerebral Cortex* 32.1, pp. 158–175. DOI: 10.1093/cercor/bhab200 (cit. on p. 114).

Reiss, Lina A.J and Michelle R. Molis (2021). "An Alternative Explanation for Difficulties with Speech in Background Talkers: Abnormal Fusion of Vowels Across Fundamental Frequency and Ears." en. In: *Journal of the Association for Research in Otolaryngology* 22.4, pp. 443–461. DOI: 10.1007/s10162-021-00790-7 (cit. on p. 123).

Robbins, Kay et al. (2021). "Capturing the nature of events and event context using hierarchical event descriptors (HED)." en. In: *NeuroImage* 245, p. 118766. DOI: 10.1016/j.neuroimage.2021.118766 (cit. on pp. 57, 59).

Robinson, Benjamin L., Nicol S. Harper, and David McAlpine (2016). "Meta-adaptation in the auditory midbrain under cortical influence." en. In: *Nature Communications* 7.1, p. 13442. DOI: 10.1038/ncomms13442 (cit. on p. 15).

Robles, Luis and Mario A. Ruggero (2001). "Mechanics of the Mammalian Cochlea." In: *Physiological reviews* 81.3, pp. 1305–1352. DOI: 10.1152/physrev.2001.81.3.1305 (cit. on p. 7).

Rosburg, Timm and Ralph Mager (2021). "The reduced auditory evoked potential component N1 after repeated stimulation: Refractoriness hypothesis vs. habituation account." en. In: *Hearing Research* 400, p. 108140. DOI: 10.1016/j.heares.2020.108140 (cit. on p. 113).

Rosburg, Timm, Michael Weigl, and Ralph Mager (2022). "No evidence for auditory N1 dishabituation in healthy adults after presentation of rare novel distractors." In: *International Journal of Psychophysiology* 174, pp. 1–8. DOI: 10.1016/j.ijpsycho.2022.01.013 (cit. on pp. 113, 161).

Rosenkranz, Marc and Martin Georg Bleichner (2022). "Auditory Perception during 3D Tetris." en-us. In: DOI: 10.17605/OSF.IO/SGVK6 (cit. on p. 31).

Rosenkranz, Marc et al. (2023). "Investigating the attentional focus to workplace-related soundscapes in a complex audio-visual-motor task using EEG." English. In: *Frontiers in Neuroergonomics* 3. DOI: 10.3389/fnrgo.2022.1062227 (cit. on pp. 57, 59, 60, 66, 73–75, 93, 95, 97).

Rosenkranz, Marc et al. (2024). "Using mobile EEG to study auditory work strain during simulated surgical procedures." en. In: *Scientific Reports* 14.1, p. 24026. DOI: 10.1038/s41598-024-74946-9 (cit. on pp. 93, 104, 118).

Rossion, Bruno and Corentin Jacques (2012). "The N170: Understanding the time course of face perception in the human brain." In: *The Oxford handbook of event-related potential components*. Oxford library of psychology. New York, NY, US: Oxford University Press, pp. 115–141 (cit. on p. 36).

Rudner, Mary, Jerker Rönnberg, and Thomas Lunner (2011). "Working Memory Supports Listening in Noise for Persons with Hearing Impairment." In: *Journal of the American Academy of Audiology* 22.3, pp. 156–167. DOI: 10.3766/jaaa.22.3.4 (cit. on p. 143).

Ruusuvirta, Timo (2021). "The release from refractoriness hypothesis of N1 of event-related potentials needs reassessment." In: *Hearing Research*. Stimulus-specific adaptation, MMN and predicting coding 399, p. 107923. DOI: 10.1016/j.heares.2020.107923 (cit. on p. 113).

Ryck, Iris Van de et al. (2025). *EEG-based Decoding of Auditory Attention to Conversations with Turn-taking Speakers*. en. DOI: 10.1101/2025.06.20.660726 (cit. on p. 148).

Sams, M et al. (1985). "Auditory frequency discrimination and event-related potentials." In: *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 62.6, pp. 437–448. DOI: 10.1016/0168-5597(85)90054-1 (cit. on p. 18).

Sanderson, Brandon (2010). *The way of kings*. 1st ed. New York: Tor (cit. on p. 51).

Sanderson, Brandon (2017). *Oathbringer*. eng. First edition. New York: Tom Doherty Associates (cit. on p. 55).

Sanes, Dan H. and Shaowen Bao (2009). "Tuning up the Developing Auditory CNS." In: *Current opinion in neurobiology* 19.2, pp. 188–199. DOI: 10.1016/j.conb.2009.05.014 (cit. on p. 13).

Sanmiguel, Iria, Juanita Todd, and Erich Schröger (2013). "Sensory suppression effects to self-initiated sounds reflect the attenuation of the unspecific N1 component of the auditory ERP." en. In: *Psychophysiology* 50.4, pp. 334–343. DOI: 10.1111/psyp.12024 (cit. on p. 161).

Santoro, Roberta et al. (2014). "Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex." en. In: *PLOS Computational Biology* 10.1, e1003412. DOI: 10.1371/journal.pcbi.1003412 (cit. on p. 177).

Scanlon, Joanna E. M. et al. (2017). "Your brain on bikes: P3, MMN/N2b, and baseline noise while pedaling a stationary bike." en. In: *Psychophysiology* 54.6, pp. 927–937. DOI: 10.1111/psyp.12850 (cit. on p. 95).

Scanlon, Joanna E. M. et al. (2021). "Does the electrode amplification style matter? A comparison of active and passive EEG system configurations during standing and walking." en. In: *European Journal of Neuroscience* 54.12, pp. 8381–8395. DOI: 10.1111/ejn.15037 (cit. on pp. 31, 146).

Scanlon, Joanna E.M. et al. (2019). "Taking off the training wheels: Measuring auditory P3 during outdoor cycling using an active wet EEG system." en. In: *Brain Research* 1716, pp. 50–61. DOI: 10.1016/j.brainres.2017.12.010 (cit. on p. 172).

Schmuckler, Mark A. (2001). "What Is Ecological Validity? A Dimensional Analysis." en. In: *Infancy* 2.4, pp. 419–436. DOI: 10.1207/S15327078IN0204_02 (cit. on p. 171).

Schofield, Brett R. (2005). "Superior Olivary Complex and Lateral Lemniscal Connections of the Auditory Midbrain." en. In: *The Inferior Colliculus*. Ed. by Jeffery A. Winer and Christoph E. Schreiner. New York, NY: Springer, pp. 132–154. DOI: 10.1007/0-387-27083-3_4 (cit. on p. 9).

Schreiner, Leonhard et al. (2024). "Mapping of the central sulcus using non-invasive ultra-high-density brain recordings." en. In: *Scientific Reports* 14.1, p. 6527. DOI: 10.1038/s41598-024-57167-y (cit. on p. 31).

Schutz, Michael and Jessica Gillard (2020). "On the generalization of tones: A detailed exploration of non-speech auditory perception stimuli." en. In: *Scientific Reports* 10.1, p. 9520. DOI: 10.1038/s41598-020-63132-2 (cit. on pp. 93, 94).

Schöne, Benjamin et al. (2023). "The reality of virtual reality." English. In: *Frontiers in Psychology* 14. DOI: 10.3389/fpsyg.2023.1093014 (cit. on p. 171).

Scott, Brian H. et al. (2017). "Thalamic connections of the core auditory cortex and rostral supratemporal plane in the macaque monkey." en. In: *Journal of Comparative Neurology* 525.16, pp. 3488–3513. DOI: 10.1002/cne.24283 (cit. on p. 11).

Shahin, Antoine et al. (2005). "Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds." en-US. In: *NeuroReport* 16.16, p. 1781. DOI: 10.1097/01.wnr.0000185017.29316.63 (cit. on p. 38).

Shamay-Tsoory, Simone G. and Avi Mendelsohn (2019). "Real-Life Neuroscience: An Ecological Approach to Brain and Behavior Research." EN. In: *Perspectives on Psychological Science* 14.5, pp. 841–859. DOI: 10.1177/1745691619856350 (cit. on pp. 24, 171).

Shan, Tong, Madeline S. Cappelloni, and Ross K. Maddox (2024). "Subcortical responses to music and speech are alike while cortical responses diverge." en. In: *Scientific Reports* 14.1, p. 789. DOI: 10.1038/s41598-023-50438-0 (cit. on pp. 88, 159).

Shannon, C. E. (1948). "A mathematical theory of communication." In: *The Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 173).

Shannon, Samantha (2020). *The priory of the orange tree*. eng. London: Bloomsbury (cit. on p. 165).

Shine, James M. et al. (2023). "The impact of the human thalamus on brain-wide information processing." en. In: *Nature Reviews Neuroscience* 24.7, pp. 416–430. DOI: 10.1038/s41583-023-00701-0 (cit. on pp. 10, 114).

Silva, Daniel M. R., Danilo B. Melges, and Rui Rothe-Neves (2017). "N1 response attenuation and the mismatch negativity (MMN) to within- and across-category phonetic contrasts." en. In: *Psychophysiology* 54.4, pp. 591–600. DOI: 10.1111/psyp.12824 (cit. on pp. 113, 117).

Snyder, Joel S., Claude Alain, and Terence W. Picton (2006). "Effects of Attention on Neuroelectric Correlates of Auditory Stream Segregation." In: *Journal of Cognitive Neuroscience* 18.1, pp. 1–13. DOI: 10.1162/089892906775250021 (cit. on p. 37).

Snyder, Joel S. et al. (2012). "Attention, Awareness, and the Perception of Auditory Scenes." In: *Frontiers in Psychology* 3. DOI: 10.3389/fpsyg.2012.00015 (cit. on pp. 37, 38).

Somervail, R et al. (2021). "Waves of Change: Brain Sensitivity to Differential, not Absolute, Stimulus Intensity is Conserved Across Humans and Rats." en. In: *Cerebral Cortex* 31.2, pp. 949–960. DOI: 10.1093/cercor/bhaa267 (cit. on p. 114).

Somervail, Richard et al. (2025). *A Two-system Theory of Sensory-evoked Brain Responses*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.5199548 (cit. on pp. 10, 114).

Southworth, Michael (1969). "The sonic environment of cities." In: *Environment and Behavior* 1.1, pp. 49–70. DOI: 10.1177/001391656900100104 (cit. on p. 21).

Stam, C. J. (2005). "Nonlinear dynamical analysis of EEG and MEG: review of an emerging field." eng. In: *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 116.10, pp. 2266–2301. DOI: 10.1016/j.clinph.2005.06.011 (cit. on p. 94).

Stangl, Matthias, Sabrina L. Maoz, and Nanthia Suthana (2023). "Mobile cognition: imaging the human brain in the 'real world'." en. In: *Nature Reviews Neuroscience* 24.6, pp. 347–362. DOI: 10.1038/s41583-023-00692-y (cit. on pp. 24, 171).

Steinmetzger, Kurt and André Rupp (2023). *The auditory P2 evoked by speech sounds consists of two separate subcomponents*. en. DOI: 10.1101/2023.06.30.547226 (cit. on p. 38).

Steinmetzger, Kurt and André Rupp (2024). "The auditory P2 is influenced by pitch changes but not pitch strength and consists of two separate subcomponents." In: *Imaging Neuroscience* 2, imag–2–00160. DOI: 10.1162/imag_a_00160 (cit. on p. 38).

Stekelenburg, Jeroen J. and Jean Vroomen (2007). "Neural correlates of multisensory integration of ecologically valid audiovisual events." eng. In: *Journal of Cognitive Neuroscience* 19.12, pp. 1964–1973. DOI: 10.1162/jocn.2007.19.12.1964 (cit. on p. 23).

Stevens, S. S., J. Volkmann, and E. B. Newman (1937). "A Scale for the Measurement of the Psychological Magnitude Pitch." en. In: *The Journal of the Acoustical Society of America* 8.3, pp. 185–190. DOI: 10.1121/1.1915893 (cit. on p. 65).

Stowell, Dan et al. (2019). "Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge." en. In: *Methods in Ecology and Evolution* 10.3. Ed. by David Orme, pp. 368–380. DOI: 10.1111/2041-210X.13103 (cit. on p. 160).

Straetmans, L. et al. (2022). "Neural Tracking to Go: Auditory Attention Decoding and Saliency Detection with Mobile EEG." In: *Journal of Neural Engineering* 18.6, p. 066054. DOI: 10.1088/1741-2552/ac42b5 (cit. on pp. 124, 146, 148).

Straetmans, Lisa, Kamil Adiloglu, and Stefan Debener (2024). "Neural speech tracking and auditory attention decoding in everyday life." English. In: *Frontiers in Human Neuroscience* 18. DOI: 10.3389/fnhum.2024.1483024 (cit. on pp. 123–125, 143, 147).

Studnicki, Amanda, Ryan J. Downey, and Daniel P. Ferris (2022). "Characterizing and Removing Artifacts Using Dual-Layer EEG during Table Tennis." en. In: *Sensors* 22.15, p. 5867. DOI: 10.3390/s22155867 (cit. on pp. 57, 166).

Taberner, Annette M. and M. Charles Liberman (2005). "Response properties of single auditory nerve fibers in the mouse." eng. In: *Journal of Neurophysiology* 93.1, pp. 557–569. DOI: 10.1152/jn.00574.2004 (cit. on p. 15).

Tallus, Jussi et al. (2015). "Effects of Auditory Attention Training with the Dichotic Listening Task: Behavioural and Neurophysiological Evidence." en. In: *PLOS ONE* 10.10, e0139318. DOI: 10.1371/journal.pone.0139318 (cit. on p. 123).

Thornton, A. Roger D., Matthew Harmer, and Brigitte A. Lavoie (2007). "Selective attention increases the temporal precision of the auditory N100 event-related potential." In: *Hearing Research* 230.1, pp. 73–79. DOI: 10.1016/j.heares.2007.04.004 (cit. on p. 145).

Thornton, Mike, Danilo Mandic, and Tobias Reichenbach (2024). *Comparison of linear and nonlinear methods for decoding selective attention to speech from ear-EEG recordings*. DOI: 10.48550/arXiv.2401.05187 (cit. on pp. 124, 148).

Tibshirani, Robert (1996). "Regression Shrinkage and Selection Via the Lasso." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x (cit. on p. 47).

Tikhonov, A.N. and V.B. Glasko (1965). "Use of the regularization method in nonlinear problems." en. In: *USSR Computational Mathematics and Mathematical Physics* 5.3, pp. 93–107. DOI: 10.1016/0041-5553(65)90150-3 (cit. on p. 47).

Timm, Jana et al. (2013). "The N1-suppression effect for self-initiated sounds is independent of attention." en. In: *BMC Neuroscience* 14.1, p. 2. DOI: 10.1186/1471-2202-14-2 (cit. on p. 161).

Tuckute, Greta et al. (2023). "Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions." en. In: *PLOS Biology* 21.12, e3002366. DOI: 10.1371/journal.pbio.3002366 (cit. on p. 160).

Ulanovsky, Nachum, Liora Las, and Israel Nelken (2003). "Processing of low-probability sounds by cortical neurons." en. In: *Nature Neuroscience* 6.4, pp. 391–398. DOI: 10.1038/nn1032 (cit. on pp. 17, 113).

Vallet, William and Virginie Van Wassenhove (2023). "Can cognitive neuroscience solve the lab-dilemma by going wild?" en. In: *Neuroscience & Biobehavioral Reviews* 155, p. 105463. DOI: 10.1016/j.neubiorev.2023.105463 (cit. on pp. 24, 94, 170, 171).

Vanthornhout, Jonas, Lien Decruy, and Tom Francart (2019). "Effect of Task and Attention on Neural Tracking of Speech." In: *Frontiers in Neuroscience* 13 (cit. on p. 143).

Wang, An Li et al. (2008a). "The Enhancement of the N1 Wave Elicited by Sensory Stimuli Presented at Very Short Inter-Stimulus Intervals Is a General Feature across Sensory Systems." en. In: *PLOS ONE* 3.12, e3929. DOI: 10.1371/journal.pone.0003929 (cit. on pp. 93, 94, 105, 112–114).

Wang, X. et al. (2008b). "Neural coding of temporal information in auditory thalamus and cortex." In: *Neuroscience*. From Cochlea to Cortex: Recent Advances in Auditory Neuroscience 154.1, pp. 294–303. DOI: 10.1016/j.neuroscience.2008.03.065 (cit. on p. 12).

Wang, Yanmei et al. (2022). "Auditory and cross-modal attentional bias toward positive natural sounds: Behavioral and ERP evidence." English. In: *Frontiers in Human Neuroscience* 16. DOI: 10.3389/fnhum.2022.949655 (cit. on p. 112).

Wang, Yijun, Shangkai Gao, and Xiaorong Gao (2005). "Common Spatial Pattern Method for Channel Selelction in Motor Imagery Based Brain-computer Interface." In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 5392–5395. DOI: 10.1109/IEMBS.2005.1615701 (cit. on p. 12).

Wascher, Edmund et al. (2022). "Visual Demands of Walking Are Reflected in Eye-Blink-Evoked EEG-Activity." en. In: *Applied Sciences* 12.13, p. 6614. DOI: 10.3390/app12136614 (cit. on pp. 30, 166).

Widmann, Andreas and Erich Schröger (2012). "Filter Effects and Filter Artifacts in the Analysis of Electrophysiological Data." In: *Frontiers in Psychology* 3 (cit. on p. 35).

Widmann, Andreas, Erich Schröger, and Burkhard Maess (2015). "Digital filter design for electrophysiological data – a practical approach." en. In: *Journal of Neuroscience Methods*. Cutting-edge EEG Methods 250, pp. 34–46. DOI: 10.1016/j.jneumeth.2014.08.002 (cit. on p. 35).

Willmore, Ben D. B. and Andrew J. King (2023). "Adaptation in auditory processing." In: *Physiological Reviews* 103.2, pp. 1025–1058. DOI: 10.1152/physrev.00011.2022 (cit. on pp. 15, 114, 176).

Winkler, Irene et al. (2015). "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP." In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4101–4105. DOI: 10.1109/EMBC.2015.7319296 (cit. on pp. 61, 98).

Winkler, István, Susan L. Denham, and Carles Escera (2013). "Auditory event-related potentials." hu. In: ed. by Dieter Jaeger and Ranu Jung. SpringerReference (cit. on pp. 22, 24, 37, 123, 134, 139, 145).

Winslow, R. L. and M. B. Sachs (1987). "Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise." In: *Journal of Neurophysiology* 57.4, pp. 1002–1021. DOI: 10.1152/jn.1987.57.4.1002 (cit. on p. 15).

Wit, Lee de et al. (2016). "Is neuroimaging measuring information in the brain?" en. In: *Psychonomic Bulletin & Review* 23.5, pp. 1415–1428. DOI: 10.3758/s13423-016-1002-0 (cit. on pp. 172, 173).

Wong, Daniel D. E. et al. (2018). "A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding." In: *Frontiers in Neuroscience* 12 (cit. on pp. 47, 145).

Woodman, Geoffrey F. (2010). "A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention." In: *Attention, perception & psychophysics* 72.8, 10.3758/APP.72.8.2031. DOI: 10.3758/APP.72.8.2031 (cit. on p. 36).

Xu, Jiawei et al. (2017). "Active Electrodes for Wearable EEG Acquisition: Review and Electronics Design Methodology." In: *IEEE Reviews in Biomedical Engineering* 10, pp. 187–198. DOI: 10.1109/RBME.2017.2656388 (cit. on p. 31).

Yarden, Tohar S. and Israel Nelken (2017). "Stimulus-specific adaptation in a recurrent network model of primary auditory cortex." en. In: *PLOS Computational Biology* 13.3. Ed. by Abigail Morrison, e1005437. DOI: 10.1371/journal.pcbi.1005437 (cit. on p. 17).

Zacharias, Norman, Reinhard König, and Peter Heil (2012). "Stimulation-history effects on the M100 revealed by its differential dependence on the stimulus onset interval." en. In: *Psychophysiology* 49.7, pp. 909–919. DOI: 10.1111/j.1469-8986.2012.01370.x (cit. on pp. 93, 102, 105, 106, 112, 114, 116).

Zekveld, Adriana A., Thomas Koelewijn, and Sophia E. Kramer (2018). "The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge." EN. In: *Trends in Hearing* 22, p. 2331216518777174. DOI: 10.1177/2331216518777174 (cit. on p. 163).

Zion Golumbic, Elana M. et al. (2013). "Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party"." In: *Neuron* 77.5, pp. 980–991. DOI: 10.1016/j.neuron.2012.12.037 (cit. on p. 33).

Zoefel, Benedikt, Sanne ten Oever, and Alexander T. Sack (2018). "The Involvement of Endogenous Neural Oscillations in the Processing of Rhythmic Input: More Than a Regular Repetition of Evoked Neural Responses." English. In: *Frontiers in Neuroscience* 12. DOI: 10.3389/fnins.2018.00095 (cit. on p. 34).

Zuk, Nathaniel J., Emily S. Teoh, and Edmund C. Lalor (2020). "EEG-based Classification of Natural Sounds Reveals Specialized Responses to Speech and Music." In: *NeuroImage* 210, p. 116558. DOI: 10.1016/j.neuroimage.2020.116558 (cit. on pp. 36, 57, 85, 88).

SCIENTIFIC CONTRIBUTIONS

PUBLICATIONS

- **Thorge Haupt**, Marc Rosenkranz, and Martin G. Bleichner (2025a). *"Exploring Relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes."* en. In: eNeuro 12.1. Publisher: Society for Neuroscience Section: Research Article: New Research. doi: 10.1523/ENEURO.0287-24.2024

- **Thorge Haupt**, Marc Rosenkranz, and Martin G. Bleichner (2025b). *"Neural response attenuates with decreasing inter-onset intervals between sounds in a natural soundscape."* en. In: eNeuro. Publisher: Society for Neuroscience Section: Research Article: New Research. doi: 10.1523/ENEURO.0210-25.2025

- **Thorge Haupt**, Lisa Straetmans, Kami Adiloglu, Martin G. Bleichner, and Stefan Debener (2025d). *"Auditory Attention Decoding to go with mobile and portable hardware"*. (Submitted to Hearing Research)

- **Thorge Haupt**, Paul Maanen, Mareike Daeglau, Miguel Contreras Altamirano, Anouk Sophie Strizke, Franziska Kiene, Julius Welzel, Mandy Roheger, and Stefan Debener (2025c). *"Enhancing Mobile Brain and Body Imaging: Open Source Solutions for Real-World Research Applications."* en. SSRN Scholarly Paper. Rochester, NY. doi: 10.2139/ssrn.5433769 (Submitted to iScience, shared first authorship)

- Marc Rosenkranz, **Thorge Haupt**, Manuela Jaeger, Verena N. Uslar, and Martin G. Bleichner (2024). *"Using mobile EEG to study auditory work strain during simulated surgical procedures."* en. In: Scientific Reports 14.1. Publisher: Nature Publishing Group, p. 24026. doi: 10.1038/s41598-024-74946-9

- Silvia Korte, **Thorge Haupt**, and Martin G. Bleichner (2025). *"EEG Signatures of Auditory Distraction: Neural Responses to Spectral Novelty in Real World Soundscapes."* en. In: eNeuro 12.7. Publisher: Society for Neuroscience Section: Research Article: New Research. doi: 10.1523/ENEURO.0154-25.2025

CONFERENCE ABSTRACTS

- **Thorge Haupt**, Marc RosenKranz, and Martin G. Bleichner (2023) *"Neural Tracking of Acoustic Onsets; Towards understanding the brain beyond the lab"*. 10th Mind Brain Body Symposium, Berlin, Germany
- **Thorge Haupt**, Marc RosenKranz, and Martin G. Bleichner (2023) *"Neural Tracking of Acoustic Onsets; Towards understanding the brain beyond the lab"*. 10th BCI Meeting, Brussels, Belgium
- **Thorge Haupt**, Marc RosenKranz, and Martin G. Bleichner (2023) *"Neural tracking of acoustic features: a step towards understanding the neural correlates of auditory processing beyond the lab"*. Methods in Mobile EEG 2.0, Belgrade, Serbia
- **Thorge Haupt** and Martin G Bleichner (2024) *"Investigating environmental characteristics driving mobile EEG measures of everyday life auditory perception."*. Neuroscience of Everyday World Conference, Boston, USA
- **Thorge Haupt**, Marc Rosenkranz, and Martin G. Bleichner (2025) *"The Importance of Environmental Information to Bridge the Gap between Laboratory and Beyond the Lab Measures of Auditory Perception"*. Cognitive Hearing Science Communication, Lincoping, Sweden

CONFERENCE TALKS

- **Thorge Haupt** and Martin G. Bleichner (2024) *"The Importance of Environmental Information to Bridge the Gap between Laboratory and Beyond the Lab Measures of EEG-based Auditory Perception"*. 5th International Neuroergonomics Conference, Bordeaux, France
- **Thorge Haupt**, Marc Rosenkranz, and Martin G. Bleichner (2024) *"The Importance of Environmental Information to Bridge the Gap between Laboratory and Beyond the Lab Measures of Auditory Perception"*. Cognitive Hearing Science Communication, Lincoping, Sweden
- **Thorge Haupt**, Lisa Straetmans, Martin G. Bleichner, and Sefan Debener (2025) *"Mobile Speech Tracking in Everyday Life with Ear-EEG"*. Methods in Mobile EEG 3.0, Belgrade, Serbia

WORKSHOPS

- Mareike Daeglau, **Thorge Haupt**, and Jakab Pilaszanovich (2024). *Advanced EEG Workshop*. Leipzig, Germany
- *"Exploring Temporal Response Functions (TRFs) in Real-World Speech Processing"* (2024). Delmenhorst, Germany

AWARDS

- Travel Award for the Neuroscience of Everyday World Conference, Boston, USA

DECLARATIONS

AUTHOR CONTRIBUTIONS

I hereby confirm that Thorge Haupt contributed to the aforementioned studies as stated below:

Study I

Thorge Haupt, Marc Rosenkranz, and Martin G. Bleichner (Jan. 2025a). "Exploring Relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes." en. In: *eNeuro* 12.1. Publisher: Society for Neuroscience Section: Research Article: New Research. DOI: 10.1523/ENEURO.0287-24.2024

Author Contribution

Thorge Haupt: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization.
Marc Rosenkranz: Investigation, Data Curation, Software, Writing - Review & Editing
Martin G. Bleichner: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Study II

Thorge Haupt, Marc Rosenkranz, and Martin G. Bleichner (Sept. 2025b). "Neural response attenuates with decreasing inter-onset intervals between sounds in a natural soundscape." en. In: *eNeuro*. Publisher: Society for Neuroscience Section: Research Article: New Research. DOI: 10.1523/ENEURO.0210-25.2025

Author Contribution

Thorge Haupt: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization.
Marc Rosenkranz: Investigation, Data Curation, Software, Writing - Review & Editing

Martin G. Bleichner: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Study III

Thorge Haupt, Lisa Straetmans, Kamil Adiloglu, Martin G. Bleichner, and Stefan Debener (submitted). "Auditory Attention Decoding to go with mobile and portable hardware"

Author Contribution

Thorge Haupt: Conceptualization; Formal Analysis; Investigation; Writing – Original Draft Preparation; Writing – Review & Editing; Visualization.
Lisa Straetmans: Conceptualization; Methodology; Data Curation; Writing – Review & Editing; Supervision.
Kamil Agidolu: CConceptualization; Methodology; Software; Data Curation; Writing – Review & Editing
Martin G. Bleichner: Conceptualization; Writing – Review & Editing; Supervision.
Stefan Debener: Conceptualization; Resources; Writing – Review & Editing; Supervision; Project Administration; Funding Acquisition.

Prof. Dr. Martin G. Bleichner

Thorge Haupt

DECLARATION OF ORIGINALITY

I have completed the work independently and used only the indicated facilities. This dissertation is my own work. All the sources of information have been acknowledged by means of references.

This dissertation has neither as a whole nor in part been published or submitted to assessment in a doctoral procedure at another university.

This is to confirm that I am aware of the guidelines of good scientific practice of the Carl von Ossietzky University of Oldenburg and that I observed them.

This is to confirm that I have not availed myself of any commercial placement or consulting services in connection with my promotion procedure.

Oldenburg, March 2026

Thorge Haupt

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of March 30, 2026 (`classicthesis` version 4.2).