UNIVERSITY OF CALIFORNIA, MERCED

Hierarchical Temporal Structure and Deep Learning Methods of Speech and Music

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive & Information Sciences

by

Adolfo G. Ramirez-Aristizabal

Committee in charge:

Professor Chris Kello, Chair
Professor Ramesh Balasubramaniam
Assistant Professor Kristina Backer

2022

The dissertation of Adolfo G. Ramirez-Aristizabal is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Professor Ramesh Balasubramaniam)

(Assistant Professor Kristina Backer)

(Professor Chris Kello, Chair)

University of California, Merced

2022

DEDICATION

Dedicado a mi familia, especialmente a mi mamá. Que juntos seguimos cruzando fronteras and continue on the path to define our dreams.

EPIGRAPH

"...language is a gift as dangerous to humanity as the horse was to the Trojans:
it offers itself to our use free of charge,
but once we accept it,
it colonizes us."
—Slavoj Žižek

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to all my friends, whom have shared drinks, laughs, stories, deep conversations, and tears. I want to also thank 17th street, Little Oven's, and all the usual local businesses that have upheld a space for building community.

0.1 Education

Ph.D., Cognitive Information Science, University of California, Merced, Merced, CA, 2016 - 2022

B.S., Cognitive Science, University of California, Merced, Merced, CA, 2012 - 2016

0.2 Technical Skills

Python – [Keras-TensorFlow, Sci-Kit Learn, Pandas, MNE]
Matlab – [EEG Lab, Linear Mixed Effects Modeling]
R-Studio – [Linear Mixed Effects Modeling]

0.3 Research Projects

0.3.1 Cognitive and Information Sciences

- Complexity matching between auditory stimulus and EEG brain response
- Computer Vision Architectures for Processing EEG Responses to Music Listening
- Linear Mixed Effects Modeling with the Buckeye Corpus, relating linguistic units and acoustic information

0.3.2 Interdisciplinary

- Deep learning researcher in 'automated classification of giant unilamellar vesicle image scans' project

- Collaborator in ICGE research group, product concept development integrating edge computing devices with research models predicting precise air quality in impoverished regions
- Project lead in IAS NRT research group, testing human orienteering performance on hyperbolic and euclidean network space via experimental game setting

0.4 Publications

[1] Schneider, S., **Ramirez-Aristizabal, A. G.**, Gavilan, C., Kello, C. T. (2019). Complexity Matching and Lexical Matching in Monolingual and Bilingual Conversations.*Bilingualism: Language and Cognition*.

[2] Dale, R., Galati, A., Alviar, C., Contreras Kallens, P., **Ramirez-Aristizabal, A. G.**,Tabatabaeian, M., Vinson, D. W. (2018). Interacting Timescales in Perspective Taking. *Frontiers in Psychology*, 9, 1278.

[3] **Ramirez-Aristizabal, A. G.**, Médé, B., Kello, C. T. (2018). Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons Fractals*, 111, 175-179.

[4] **Ramirez-Aristizabal, A. G.**, Ebrahimpour, M., Kello, C. T. (2022). Image-Based EEG Classification of Brain Responses to Song Recordings. *arXiv preprint arXiv:2202.03265*.

[5] **Ramirez-Aristizabal, A. G.**, Kello, C. T. (2022). EEG2Mel: Reconstructing Sound from Brain Responses to Music. *arXiv preprint arXiv:2207.13845*.

0.5 Conference Presentations

[1] **Ramirez-Aristizabal, A.G.** Kello, C.T. (2017, August). Speaking Rate as a Constraint on Hierarchical Temporal Structures. Presentation at the biennial Society for Music Perception and Cognition 2017. University of California San Diego.

[2] **Ramirez-Aristizabal, A.G.**, Comstock, D.C., Kello, C.T. (2019, March). Complexity Matching to EEG Response of Speech and Music. Poster at the annual Cognitive Neuroscience Society 2019. San Francisco.

[3] **Ramirez-Aristizabal, A.G.**, Ebrahimpour, Mohammad., Kello, C.T. (2021, October). Using Deep Learning to Interpret Acoustic Brain Responses. Poster at SACNAS National Diversity in STEM Conference.

[4] **Ramirez-Aristizabal, A.G.**, Ebrahimpour, Mohammad., Kello, C.T. (2022, April). Neural Decoding of Music Stimuli from Auditory Evoked EEG Responses. Talk and Poster at Cognitive Neuroscience Society 2022. San Francisco.

0.6 Honors, Awards, and Training

Innovations in Graduate Education Certificate
-Interdisciplinary Computational Education Program
-National Science Foundation

National Research Training [Participant]
-Intelligent Adaptive Systems
-National Science Foundation

Certificate in Undergraduate Learning Outcomes Assessment
-Pedagogy and Program Planning
-Center for Engaged Teaching and Learning

Miguel Velez Scholarship
-Prestigious scholarship for Latin American graduate students
who exhibit excellence in character and ability

SMPC Student Abstract Submission Award
-Given to the top abstract submissions for the SMPC conference

to cover travel costs

Faculty Mentor Program Fellowship

-Competitive fellowship aiding year long projects and student training

Teaching Fellowship

-Mission is to help new Teaching Assistants navigate the difficulties of teaching during the school's transition back into in-person classes for a large enrollment Intro to Cog Sci course

0.7 Certifications

IBM Data Science Professional Certificate Specialization

- A series of 9 certifications from data science courses teaching machine learning with python for industry
- Courses['What is Data Science','Open Source Tools for Data Science','Data Science Methodology','Python for Data Science AI','Databases SQL','Data Analysis with Python','Data Visualization with Python','Machine Learning with Python','Applied Data Science Capstone']

Other IBM Certifications

- Courses['Fundamentals of Scalable Data Science','Advanced Machine Learning Signal Processing','Applied AI with Deep Learning']

0.8 Teaching Work Experience

Neuromatch Deep Learning Academy

- NMA-DL (Summer 2021): Was a Teaching Assistant on a full time summer course of deep learning in Neuroscience to a group of international students and advised them on student projects

University of California Merced (Teaching Assistant)

- COGS 101 (2016, 2017) Mind, Brain, and Computation, paired with the training received from pedagogy and program planning.
- COGS 01 (2017, 2021) Introduction to Cognitive Science
- COGS 105 (2018) Research Methods
- COGS 103 (2018, 2020) Introduction to Neural Networks
- COGS 05 (2019) Introduction to Linguistics
- COGS 104 (2020) Complex Systems
- POLI 190 (2021) Politics of Economic Development

0.9 Languages

Spanish (Native Fluent)

0.10 Service and Leadership

- Summer 2021-Mentor in a deep learning project for Undergraduate Research Opportunities Center-Summer Undergraduate Research Fellowship (UROC-SURF)
- Fall 2019-Instructor for Data Science with Python Independent Study
- Fall 2018-Student Coordinator for Cognitive Science for the Common Good: **Undocu-Ally Training**
- June-August 2017-Peer Mentor for the Competitive Edge Summer Bridge Program
- April 2017-Latinx Graduate Research Symposium-Panelist and poster presentation
- March 2017 –Workshop Lecturer on 'Intro to Cognitive Science: Thinking About AI' for visiting High School students at UC Merced Splash Conference

Chapter 1

Summary

Complex acoustic signals such as speech and music are central to how people coordinate and communicate temporally. These signals can be described by the way their variability of acoustic energy is built across scales of time. Their structures have been shown to be reflected as people interact with others, themselves, and their environment. The studies in the following chapters show evidence to this end through development of acoustic and linguistic statistical methods, behavioral measurements of speech coordination & production, a behavioral to neural experimental paradigm on measuring temporal structure, and finally new deep learning approaches aiding in answering unresolved questions of the prior experimental paradigm. Theoretical predictions on information transfer have guided expectations across several experimental paradigms relating to this topic, which have brought about a line of successful behavioral studies. But when presented with measurements that face techno-methodological difficulties, such as teasing apart the structure of speech or music in cortical activity, these predictions lead empiricists with a fork in the road. On one path we are faced with the challenge of explicitly looking for the hierarchical temporal structure that defines these signals in human behavior, but at the cost of over parameterizing the experimental controls in data collection. On the other hand, we can simply develop deep learning models that boldly assume the existence of these temporal structures in human behaviors through their training procedures but validates what could easily be an erroneous assumption through the strength of its predictive power. Finally, the chapters here will synthesize results

from both experimental and computational modeling approaches, outlining how the epistemic feedback loop of theoretical predictions, data collection, and the creation of new testable hypothesis from a Complex Systems & Dynamics framework of cognition has shaped our understanding of acoustic temporal structure as it moves through our mouths, bodies, and brains.

1.1 Hierarchical Temporal Structure

When we ask questions about the structure of a system, we seek to understand how a specific system is composed. Furthermore, we can ask whether its composition defines functionality of the system i.e., whether we can attribute the behavior of a system to its compositional specifications. For example, we can think back to a traditional grade school system model such as the atom and ask ourselves these same questions to better understand it. Depending on when you took your first chemistry or physics class and the pedagogical limitations of your academic institutions in primary education, you most likely had a variation of the Niels Bohr planetary model, or the Erwin Schrodinger quantum model assessed on your exams. Both models make an argument for the topography of electrons either following fixed orbits or a probabilistic state of trajectory as they travel around the nucleus. In these cases, the differences in how their structures are composed have shown a difference in their capability to explain the behavior of electrons, with the quantum model having a stronger generalization across different types of newly discovered atoms over the years [Kragh, 2012]. With this, it buys scientists a method to now put testable hypotheses across many different scenarios of atomic variability and it allows for predictions of behaviors to be made when measuring the interaction in more complex procedures.

Moving beyond the common reductionist example of atoms and the passé motivation of understanding the natural world through its most irreducible component structure, we go towards an understanding of systems that is motivated by the measurement of emergent properties across the interactions of system component structures. For example, we can think of the defining features of water such as its

polarity or high heat capacity, and how that property of a water system is not contained in hydrogen or oxygen on its own. Rather, this is a property that emerges from the specific interactions of atoms and molecules. Therefore, studying only hydrogen or oxygen does not tell you much about the systems that it builds at larger scales, and the search for the most irreducible component structure becomes irrelevant to understanding the totality of the natural world. This approach can also be explained as understanding the complexity of a system i.e., the different levels of a system that are responsible for the emergent behaviors from its structural components. Common examples of complex systems and their emergent behaviors consist of ant colonies, the immune system, the world wide web, economies, the brain, and the mind. Ants much like neurons, are simple low-level components of their systems. On their own they demonstrate a limited behavioral capacity, but their coordinated interactions demonstrate behaviors not expected to arise from observing the individual components [Hofstadter, 1979].

When measuring the behaviors of complex systems, we can also measure how their behaviors at different scales of time interact to produce complex behavior. This can be seen anywhere from the biorhythmic variability in heartbeats across time that can be indicative of homeostasis, or perturbations at different scales indicating the shift into a heart attack, much like how we can study small tremors across time being indicative of an upcoming earthquake in geological readings [Kiyono et al., 2005, Lise and Paczuski, 2001]. Empirically this can be measured by observing the energy of the system at different frequencies of measurement. How the structures of these measured signals are clustered across scales of time has been referred to as Hierarchical Temporal Structure (HTS). The term (HTS) was coined in [Falk and Kello, 2017] to measure temporal variations of acoustic energy of infant directed speech and adult directed speech at multiple time scales. Furthermore, here HTS will be a foundational concept to understand how complex signals like speech and music can be understood to interact with other complex systems.

1.2 Information Transfer

Prior paradigms of complexity science have tried to reconcile how information interacts across scales with reductionist epistemology by framing it simply as memory or history of the underlying process [Moss et al., 2004]. This works well in simple models that only account for closed systems and their effect on the outside, as one takes the role of the perturbed and the one who generates those perturbations [Lindenberg and West, 1990]. For example, we can think of a swimmer here as being the model system of interest and the ripples they leave in the water as being simply the history or memory of their actions. What is lost here is the understanding that the water is also affecting how the swimmer interacts. A modern perspective in complexity science would describe those ripples not only as a perturbation, but rather the actualization of how the swimmer can traverse through that space as they push through, and the water pushes back. Here this interaction between the swimmer and the water is what is framed as information transfer [KISH et al., 2001]; the back-and-forth communication of kinetic energy between both systems that defines the emergent act of swimming.

[West et al., 2008] formally synthesize the above-mentioned theoretical framework through the concept of maximal information transfer between complex networks. They review dynamic systems and phase space modeling equations through numerical analyses, while including numerical simulations of artificial networks modeling the production of $1/f$ noise in natural systems such as in neuronal activity. The production of $1/f$ noise is of interest in complexity science as inverse power-law distributions have been experimentally verified as being produced across a variety of systems of interest such as in linguistics [Zipf, 1949], household income [Reed, 2003], scientific citations [Silagadze, 1999], and in connections across the internet [Watts, 1999, Barabási, 2003]. Specifically, self-organized networks have been argued to produce $1/f$ as a signature of ideal long-range correlation within a system that could connect low-level components such as molecules all the way to higher-level systems such as mindfulness [Anderson, 2000]. With this, the conceptual pillars of complexity, emergence, and information transfer are connected for the purpose of understanding how complex systems communicate and the extent

to which the information is preserved across levels of interactions. Further dissemination of the $1/f$ phenomena will be explained throughout the following chapters as the methods of how this can be seen in music and speech is demonstrated.

Here information transfer is studied through the measurement of the HTS in people's produced signals and how their structures are preserved under three scenarios. The first scenario of interest is when people interact with each other. Chapter 2 examines this by looking at empirical studies of coordination between people and specifically focuses on speech and language. The extent to which the nested clustering of linguistic units can be reflected in the HTS of speech is examined, because speech is the physical signal that brains can resonate to. Second, is when people interact with themselves, and this is explained in Chapter 3 from a motor-control aspect. People can modulate their behaviors according to control parameters they impose. This puts into question whether the HTS of a signal normally produced by someone can still preserve its structure across magnitudes of imposed control parameters and whether these adapted changes to produced HTS affect resonance to other complex networks. Lastly, Chapter 4 dives into the study of how HTS is preserved from a perception aspect. People are exposed to different types of stimuli, and they affect our own biorhythms [Redfern et al., 1994]. The data provided specifically looks at the extent to which cortical activity can reflect the structures of listened speech and music stimuli. The theoretical framework of maximal information transfer posed by [West et al., 2008], outlines how discrepancies between the endogenous dynamics of complex systems can limit the bandwidth of information transfer. Here this paradigm in information transfer is extended by looking at the methodological and experimental paradigms that can study this through speech and music. The goal is to understand whether these methods could ultimately validate predictions of maximal information transfer of acoustic stimuli to cortical activity. Technological and methodological limitations to this end have challenged the assumptions made under theoretical frameworks in complexity science. To continue to explore the question of how temporal structures are reflected in movement and neural behavior, a reframing of the hypothesis at hand is performed as seen in Chapters 5 & 6 through the use of Deep Learning

paradigms.

1.3 Speech, Music, and Cortical Activity

As previously mentioned, the term HTS was coined by [Falk and Kello, 2017] to measure temporal variations of acoustic energy of infant directed speech and adult directed speech at multiple time scales. These time scales referred to the embedded hierarchical organization of linguistic units such as phonemes, syllables, phrases, and utterances. Moreover, the importance of the embedded organization in language has been comparatively studied in music as well as in terms of syntactic organization. Hierarchy simply referring to how underlying structures embed measured temporal activity, creating a scaling function of temporal structure as its hierarchy. A shared relation between language and speech organization has led to studies finding meaningful similarities in neural processing [Farbood et al., 2015]. Specifically, syntactic processing in the brain shares activation pathways when parsing the embedded temporal variation of both linguistic and musical units [Patel, 2003]. Such studies have connected the HTS of speech and music signals to human behavior, posing the question of the extent to which one can expect these structures to be preserved. Exactly how that may happen, or if it happens at all, has remained elusive to experimental validation when talking about these signal structures being reflected in collected brain responses to a complex acoustic stimulus. Despite this, plenty of theoretically oriented predictions and modeling work aim to outline these cognitive processes by focusing on the $1/f$ nature of neuronal activity [Buiatti et al., 2007].

Empirical studies have replicated and found that measurements of electroencephalograms (EEG) from cortical activity have, what is termed here as an HTS of $1/f$ distributions, which have been thought of as being an ideal homeostatic state [Allegrini et al., 2009]. Despite how several studies have tried to frame this phenomenon, brains producing $1/f$ noise simply means that the power per frequency interval is inversely proportional to the frequency of the EEG signal. The power in this type of noise is equal at each octave while having power drop off as frequency

scales up. This puts $1/f$ noise in the middle of white noise and brown noise, where the former is activity without changes in power and no correlation while Brownian motion has changes in power but no correlation between them. This is what makes $1/f$ interesting due to its long-range correlation, one which despite its pervasive finding in biological systems has not yet had a formal physical explanation behind its generation [Eliazar and Klafter, 2010]. Furthermore, Brownian motion can simply be obtained by taking the integral of white noise with the reverse being true by taking the derivative of white noise, but such a simple transformation cannot be said for $1/f$. Hence why perspectives on self-organized criticality refer to $1/f$ as a proposed signal of homeostasis in systems due to its minimal stability being configured despite having many degrees of freedom without the need for the modeling of an outside driving force. An example of this phenomena in natural systems is that of the delicate stability of snow fields that can self-organize into beautiful landscapes, but with some perturbation quickly transition into a deadly avalanche. This transition of critical states is also measured in neuronal activity through their own electrical avalanches [Beggs and Plenz, 2003]. Therefore, $1/f$ noise is less so than what one would commonly think of as noise or uncorrelated information but rather a signature of a complex system being sensitive enough to interact with the natural world at multiple scales.

Given the predictions made by the maximal information transfer framework, because the brain demonstrates 1/f complexity, then it should also be the most sensitive to signals produced by another 1/f-network [West et al., 2008]. Stimuli such as music have been originally discovered to have 1/f scaling in its amplitude [Voss and Clarke, 1975], and this attribute of self-similarity has been validated by [Hsü and Hsü, 1991] in their comparisons of natural systems to the fractal geometry of music amplitude. Researchers have looked at how music can elicit emotional responses, such as the 'chilling of the spine' effect [Hachinski and Hachinski, 1994]. Complexity science researchers have boldly posed that the complexity in emotions triggered by music could come from the 1/f resonance experienced during music listening [Lewis, 2005]. Such a proposition has not yet been verified through empirical methods due to the difficulty of processing noisy cortical responses. Nevertheless,

such predictions have also been put forth when it comes to language. The well-studied Zipf's law presents a case for this where one can find that written words follow an inverse frequency pattern relative to its rank [Zipf, 1945]. For example, function words such as 'and' will have a high frequency in the document distribution while a highly informational word such as 'supercalifragilisticexpialidocious' might only show up in rare instances. Apart from linguistic units, studies have also found ways to measure the HTS of speech, in the same way that music has [Schneider et al., 2020]. Not so surprising, the amplitude of speech also shows 1/f scaling [Jennings et al., 2004]. Furthermore, conversations between interlocutors show greater coordination when the HTS of their speech signals match, adding further evidence to the proposed resonance from acoustic stimuli to human behavior, which could include cortical activity [Abney et al., 2014].

Lastly, the HTS of speech and music have also shown to match with activity from natural systems. For example, a study adopted methods using acoustic amplitude to measure its scaling across timescales to find commonalities between jazz music having the same HTS as conversations, as well as the HTS from classical music matching the HTS from thunderstorms. Furthermore, signals such as whale calls and earthquakes also present variations of and from $1/f$ HTS [Kello et al., 2017]. If these structures are present in such a variety of systems, then how should we understand how the human brain may resonate with these signals? Here an explanation to this question will start to develop by focusing on a specific purpose for this resonance to occur, and that is for the sake of communication and the extent to which that informational content in communicative signals is preserved between us, within us, and with our environment. The primary informational content discussed will be around the temporal energy of human behaviors and not strictly based on the semantic content of symbolic units. Methodological paradigms have been successful in validating theoretical predictions when it comes to how HTS is preserved between us and within us but have fallen flat when trying to connect it further from environmental stimuli to cortical activity. Cortical activity can also be thought of as a special case, because unlike other interactions, the brain has been framed as a producer of 'ideal' $1/f$. In brief, this notion of

'ideal' simply refers to the fractal exponent in 1/f as being within a range that is characteristic of sub-critical systems [Allegrini et al., 2009]. Why this is ideal, is because it is indicative of the delicate stability which, if perturbed enough, could quickly transition into states that are not characteristic of homeostatic activity [Bak et al., 1987]. This proposed 'ideal' range of brain dynamics is therefore the target level of tuning that the brain has for any stimuli, and any stimuli beyond that range would resonate less and less with the brain. Furthermore, this concept of 'ideal' frames an argument for the connection between maximal information transfer and the ubiquity of $1/f$ dynamics in natural systems. What this ultimately proposes is that information exchange between complex networks is maximized when they both have a long-range sensitivity to perturbations across timescales, which means that they both produce 'ideal' $1/f$ noise. In the following chapters, this concept will be developed to understand to what extent this may be happening and what it would mean for speech and music to either fully resonate in the brain or whether only some of the signal can pass through.

1.4 Acoustic Information Retrieval

Neural entrainment studies have demonstrated a specific case for the extent to which acoustic stimuli can modulate neuronal activity [Power et al., 2012]. The traditional experimental paradigm for these studies takes repeated EEG measurements from participants as they are presented to short simple rhythmic tones [Jones, 2010]. Results from these studies have shown that neural oscillations, which have their own endogenous dynamics, adjust their phase in relation to that of the stimulus [Nozaradan et al., 2011]. These findings have been extended into including complex rhythmic stimuli and finding stronger coupling between the stimulus and neuronal oscillations at specific frequencies [Tierney and Kraus, 2015]. Music and speech processing theoretical frameworks have been developed around these findings following the dichotomy of Bayesian frameworks and that of Complex Systems & Dynamics perspectives [Keller and Mrsic-Flogel, 2018, Dubois, 2003]. Of relevance to the argument being proposed in this dissertation, the Complex Sys-

tems & Dynamics perspectives have explained the phenomenon of pseudo-rhythmic complex acoustic stimuli and entrainment in neural oscillations not simply as one controlling the affect of the other, but rather as the stimulus being used as a referent for coordination between the brain's own endogenous dynamics adjusting its anticipation of incoming information [Rimmele et al., 2018]. Furthermore, empirical validation has been presented to connect hierarchical temporal correlation of speech and neural processing i.e., temporal activity of speech stimulus can be tracked at specific frequencies of neuronal activity [Ding et al., 2016].

Apart from neural entrainment empirical paradigms, some other studies focusing on the localization of neural processing to acoustic stimuli have shown a hierarchical relationship to the temporal dependencies of acoustic stimuli i.e., lower frequency information has longer correlated activation pathways than higher frequency [Farbood et al., 2015]. Some other studies have also argued for methodological techniques finding covariance of complexity measures from acoustic stimuli to EEG signals but further validation of their results and generalizability has been of a limited scope with stimuli used [Teixeira Borges et al., 2019a]. Therefore, a direct approach at finding the HTS of speech and music in cortical activity has been limited, but acoustic information retrieval studies have added another piece of evidence for this endeavor by demonstrating the ability to reconstruct acoustic stimuli using brain responses as input [Coffey et al., 2019]. Why this is of relevance, is because it is expected that the brain's $1/f$ dynamics resonate and reflect stimuli information across multiple scales of time, which means that an information retrieval approach will assume that brain responses contain information to their corresponding stimuli in order to attempt to recover that information.

Early experimental paradigms in acoustic information retrieval have depended on short stimuli and averaged responses from repeated stimuli presentations [Skoe and Kraus, 2010]. Such endeavors have allowed the field at large to see what is possible by allowing to play back about a second of acoustic stimuli reconstructed from brain responses. These early approaches were important proof-of-concepts, but a desire to optimize the methodological procedures arose leading to bigger modeling approaches. Earlier modeling approaches decided to use machine learning

approaches to classify the acoustic stimuli and reconstruct simple features of the stimuli, but depended on extensive feature extraction [Moinnereau et al., 2018]. More recent approaches, as will be discussed in other chapters, have tried to use deep learning to allow for the processing of naturalistic data collection [Ramirez-Aristizabal et al., 2022]. The predictive power of deep learning paradigms has been successful enough to put to question what these networks are actually learning. Why this could be put into question is because neural networks could learn nothing related to a meaningful relationship between the input and target, but its predictive power could generalize to make good enough predictions to produce desired outputs. Despite this, there has been several deep learning studies showing successful acoustic stimuli classification from EEG, and their results across different datasets give confidence that these results could be more than a trivial generalization [Stober et al., 2014, Yu et al., 2018, Sonawane et al., 2021]. Furthermore, some recent deep learning approaches have been able to reconstruct the acoustic stimuli without depending on unnaturalistic data presentations [Ofner and Stober, 2018]. Results from classification models alone are of interest but such models only need to discriminate inputs into target labels, which could mean that the networks learn that input is simply different than the other. Reconstruction studies on the other hand, achieve a more difficult task and in needing to predict acoustic information, give better evidence that EEG inputs have information relating to the original stimulus.

1.5 Conclusion

People produce speech and music signals containing a HTS indicative of $1/f$ dynamics, which have been signatures of complex systems and its long-range temporal correlations demonstrating the capacity of these systems to interact with the natural world at multiple scales of time. These properties point to how complex systems interact with other complex systems, produce communicative signals, and how the information in those signals is preserved within those systems across multiple scales. Theoretical predictions have posed that the structure of these sig-

nals should resonate with corresponding systems, but so far it has been difficult to provide concrete empirical evidence of that happening with speech and music stimuli on cortical activity. Therefore, the question of to what extent the structure of these signals is preserved in cortical signals is of importance because the same methodologies that have worked in empirically validating in other modalities have not worked to the same extent for cortical activity. Whether this is indicative of theoretical predictions not working in the domain of cortical dynamics or simply a techno-methodological limitation is still up for debate. The development of deep learning approaches in acoustic information retrieval has allowed to tackle this question under a different hypothesis testing framework. These models allow for researchers to assume that acoustic information relating to its temporal structure is hidden in brain responses, and can test it through the success of model training procedures. Despite the lack of evidence to frame the problem as EEG being a noisy version of the stimulus, this bold assumption allows for the creation of new testable hypotheses guided by the strength of predictive power of deep neural networks. The integration of deep learning modeling to this end will be argued as a paradigm shift meant to resolve the limitations of current experimental paradigms when exploring how temporal structure of speech and music are reflected in human behaviors.

Chapter 2

Relating Multiscale Linguistic Units with Acoustic Features of Natural Conversations from the Buckeye Corpus

Studying the statistical structure of language in speech has traditionally been dependent on the usage of symbolic linguistic units. A few publications have moved forward and implemented the Allan Factor method to measure the statistical structure of speech data without needing to depend on linguistic transcriptions. The method has previously been used to study the fractal dynamics of neural spike trains as well as to measure speech when people interact with each other. This method takes the normalized variance of acoustic events from the speech signal at various timescales. A scaling function is then produced per signal, which considers the hierarchical structure of its temporal variation. Here the statistics from hand coded linguistic units of speech are analyzed to see how it relates to the multiscale statistics derived from the Allan Factor method. The Buckeye corpus is used here as it provides recorded and transcribed data from interlocutors having spontaneous and natural conversations. Furthermore, this dataset presents a challenging benchmark for understanding how the nested clustering relationship long studied in language

can be measured through speech signals, which is the physical signal interacting with a listener's neuronal activity.

2.1 Emergent Structure of Language & Speech

The tradition of analyzing language structures through their symbolic linguistic units has been facilitated through the availability of written documents and annotated corpora of human interactions and thought [Biber et al., 1998]. Technological advances have also allowed for the development of readily available computational techniques to record and analyze sound patterns such as from human speech. Nevertheless, with advancements in consumer technologies in social media, it has further allowed for the proliferation of available text data to analyze. So much so, that it has become a space for the development of state-of-the-art deep learning natural language processing models with billions of parameters. A notable example is GPT-3, which is a natural language processing generative model used to automatically write movie scripts, news articles, and prose; the caveat with this is that the training of such a model is estimated to have the same carbon footprint as driving a car to the moon and back [Patterson et al., 2021]. At the core of many contemporary natural language processing models in machine and deep learning implementations, is the fundamental mapping of statistical frequencies of linguistic units in documents to specific semantic categories or statistical distributions used for content generation [Wei et al., 2022]. This phenomenon can be traced back to the interdisciplinary approach of cognitive linguistics which would ask about language as representation, pragmatics of discourse, and informational statistics in nature [Rosenberg et al., 1974]. Specifically, the perspective at the intersection of frequency and linguistic structure sought to understand how the production of linguistic information led itself to the emergence of adaptations in complex structure [Hopper and Bybee, 2001]. This perspective understood that there was a correspondence between the content of human social interactions, perceptual mechanisms such as working memory, and the statistical distribution of words [Baddeley, 1998]. Despite the continual high frequency of citations in the

literature, pioneering findings from George K. Zipf remain relevant as the ubiquity of inverse power laws in nature continue to be found and connected to hypotheses about the emergent structure of complex systems [Zipf, 1945].

The study of natural discourse traditionally used linguistic units to understand how language production systems would build hierarchical structures in grammar. Furthermore, a connection between syntax and prosody in grammatical structures would take shape as findings would connect behavioral factors of speech production to shape the statistical frequency of specific units such as words and phrases in underlying produced utterances [Hopper and Bybee, 2001]. For example, when looking at the frequency of co-occurrence between words such as in the sentence 'Do you want to go?', high frequency of use of those elements have presented a case for fusion into new sentences 'wanna go?' [Baddeley, 1998, Boyland, 1996, Bybee and Scheibman, 1999]. This adds a peculiar case for the loss of constituent boundaries in sentences such as the dissipation of a main clause and subordinate clause in sentences using 'wanna', highlighting an adaptation between syntactic elements in construction due to prosodic changes. Word use frequency also demonstrates cases of phonological reductions, such as in the case of binomials written as 'bread and butter' but pronounced with 'and' having a shorter durations and vowel reduction [Krug, 1998, Fenk-Oczlon, 1989]. These examples briefly demonstrate the interplay between behavioral constraints (e.g., high frequency of lexical use) in language production shaping both the prosodic dynamics in natural discourse as well as syntactic construction of specific sentences. Not only does this connect the written symbolic units of language with the actual soundwaves produced in speech, but it further highlights the boundaries between representational levels in language i.e., the sentence 'Do yall' wanna go there?' representing 'Do you all want to go there?' which further represents clearer syntactic boundaries between subject and object [Bybee and Scheibman, 1999]. Lastly, an example moving beyond the scale of word usage and into a more relevant scale of production within a conversation can be seen with lenition. The production of words within one conversation has its own phonological variability, and in some cases, lexical words that are repeated often show a shift into weakened articulation of its consonants. This allows for the

consonants in someone's produced speech to be more sonorant and have shorter durations. With this, the above-mentioned examples outline how structure, both symbolically through linguistic units and in speech amplitudes can emerge through the shifts in variance caused by frequency.

2.2 Measuring Interlocutor Coordination

Methodological and theoretical advancements in Cognitive Linguistics have allowed for thinking of discourse from a systems perspective. This is seen in [Pickering and Garrod, 2004] with their Interactive Alignment Model outlining a theory for explaining the dynamics of convergent behavior between interlocutors. The model explains how a hierarchy of linguistic levels (phonetic, lexical, semantic, and situational) are involved between interlocutors as they share linguistic representations across scales of measurement. Alignment here describes that efficient communication underlies shared linguistic representations which allows for a temporal coordination predicting a coordination that scales within a shared language. Included in methodological advancements has been the shift towards not relying heavily on linguistic units due to the laborious job of annotation. Furthermore, the job of annotation also requires linguistic expertise when studying natural discourse because you would need more levels beyond just words including phonemes, syllables, phrases, and so on. With embodied perspectives in cognition pointing to the importance of studying extra-linguistic information, time series analysis has allowed for both scaling up experiments and presenting comparable analyses to other behavioral measurements. Under this methodological paradigm, [Paxton and Dale, 2013] found that argumentative conversations disrupted behavioral matching while friendly conversations showed evidence for body movements to converge. A follow up study using the same data used the Allan Factor of acoustic onset intervals to show that speech also converges at multiple scales of time when interlocutors are not in argumentative conversations [Abney et al., 2014].

Allan Factor as a method has had comparative success to Detrended Fluctuation Analysis (DFA) in experimental paradigms by allowing a clearer view into the

non-linear relationship across scales of time. The use of AF has not been limited to just speech, but was also used to analyze animal vocalizations, music genres, and thunderstorms [Kello et al., 2017]. Each of these unique acoustic signal categories have their own hierarchical temporal structure (HTS), and while being perceptually distinct between categories some of them are alike in their statistical structure. For example, thunderstorms were found to have the same HTS as classical music while the HTS of conversations were the same as jazz music. Much like the linguistic hierarchy in speech, music also contains its own hierarchy e.g., notes, motifs, phrases and so on [Patel, 2003]. That hierarchy is thought to be reflected through the measured HTS from the AF method as well. This assumption seems to work well when comparing acoustic signals that have a verified symbolic hierarchy of their informational content. That is why it is surprising when thunderstorms produce HTS statistically the same as classical music, because it is not a signal defined by the production of nested informational units like in music or speech language. With that, researchers have asked the question as to what HTS may be relating across signals and some studies have pointed to the HTS as reflecting prosody [Schneider et al., 2020]. Furthermore, predictions from the Interactive Alignment Model are put into question as convergence in HTS is shown when speakers communicate in differing languages (English and Spanish). Although both languages may have similar grammars, they each follow their own syntactic rules and phonological constraints, which means that there may be more to coordination between interlocutors than a shared linguistic representation.

The relation between speech amplitudes and linguistic units in HTS is first investigated in an infant directed speech study. The HTS of speech directed at infants is compared to adult directed speech as well as song. The results find that both song and speech have their own HTS characteristic but that in each category the slope of the HTS is higher when it is infant directed. A higher slope in these functions is attributed to the prosodic exaggeration that mothers produce when speaking to their infants as opposed to an adult [Falk and Kello, 2017]. To further provide evidence for the connection of HTS as prosody and reflecting linguistic hierarchy information, they created models using linguistic transcriptions as pre-

dictors to the linear slope of AF functions. Their approach to test the relationship between linguistic information and the AF functions is adopted here. The model from the infant directed study included the coefficient of variation (CoV) of durations at 3 linguistic levels (syllable, word, and phrasal constituent continuation) as well as speaking rate measured by syllables per second and the standard deviation of phrasal pre-final lengthening. Their data came from 15 speakers averaging 6 min in length and only including utterance final contents. Results from that approach defined a model with $R^2 = 0.936$ and a $p < 0.001$ significant F-Test. Here the approach is extended by using the Buckeye Corpus which contains more participants and longer recordings.

2.3 Methods

2.3.1 Data Acquisition

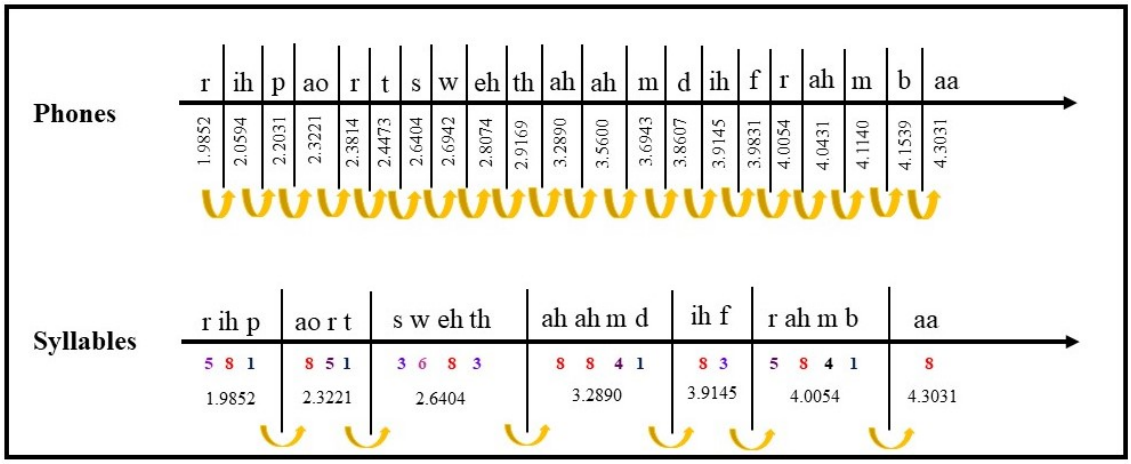

Figure 2.1: Syllables were feature extracted based on a 1-8 sonority scale tagged onto phones. Syllable boundaries as shown above were created when the sonority value increased. Durations were calculated by subtracting the onset time of the current unit from the adjacent identified unit as pointed by the yellow arrows.

The data used comes from the Buckeye Corpus of spontaneous speech, which contains 40 interviews with speakers native to central Ohio. The corpus aimed

to contain a size of 300,000 words available in acoustic recordings, transcriptions, and linguistic unit coding of the transcriptions. The demographic of participants is balanced by age of speaker (under 40, over 40), gender of speaker, and gender of interviewer [Pitt et al., 2005]. Specifically, the corpus provides timestamped phonetic and word level transcriptions, which are the primary data used for linguistic units. In total, we tested four linguistic features (phones, syllables, words, breath groups), two of which (syllables and breath groups) were feature extracted from the provided phonetic and word level transcriptions. Feature extraction had to

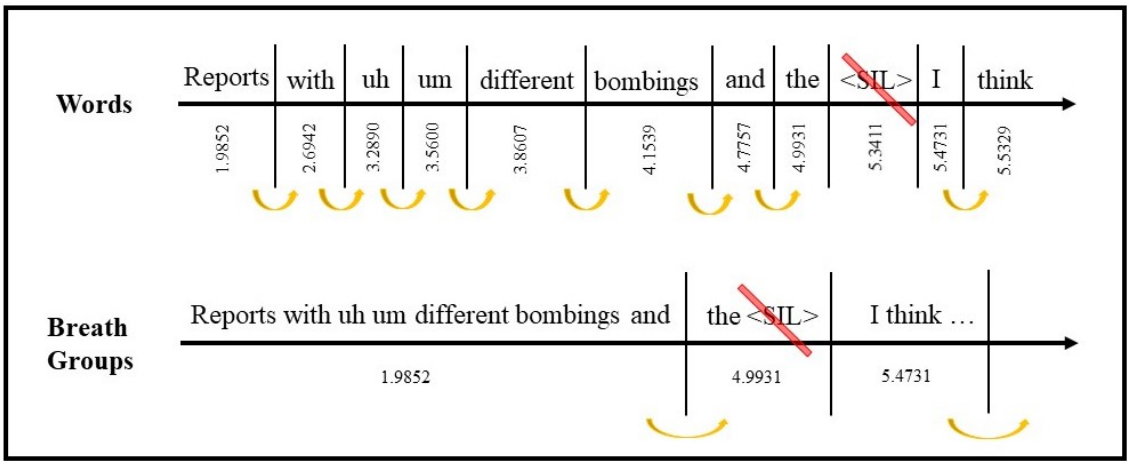

Figure 2.2: Breath groups were feature extracted based on continuous strings of words without interruptions. Any word before a non-speech label was omitted from the breath group and its onset used as a duration stop reference because of transcription errors in timestamping non-speech labels that came with the data set. Yellow arrows reflect that rule by avoiding the crossed-out units as a starting point for duration reference.

consider inconsistencies and errors in transcription timestamps. The features were defined as general proxies for syllables and breath groups as opposed to being exact definitions of those linguistic units as shown in Figure 1 & 2. The strength of this approach is that it allows for simple and consistent reproducibility. Syllable extraction was based on an eight-point sonority ranking of transcribed phonemes (stops, affricates, fricatives, nasals, liquids, glides, syllabified nasals, vowels). The syllable search algorithm simply compares the sonority values of its neighbors and identifies a syllable nucleus if the neighbors of the phonemes have descending sonority values.

Then the syllable boundary is marked by a rising sonority value of a neighboring phoneme (Figure 1). Breath group extraction was based on the word transcription data, and a sequence of uninterrupted words became what is termed here as a breath group. Examples of interruptions in the recordings include laughter, non-speech vocalization, silence, environmental noise, and switch to interviewer for which the data set included its own list of non-speech labels used to identify those events (Figure 2). Both word and phoneme transcriptions were found to have some erroneous and inconsistent timestamp labeling with data adjacent to interviewer switches. To avoid measuring erroneous timestamps, linguistic units occurring before any non-speech label where omitted. Finally, both the audio data and linguistic data were analyzed in four-minute chunks to lineup with the length needed for the acoustic analysis.

2.3.2 Statistics

The audio data was analyzed using the Allan Factor (AF) method used in [Falk and Kello, 2017] in which various categories of acoustic signals were compared based on their Hierarchical Temporal Structure. The specific details of the method can be found there but briefly put: the amplitude envelope of a signal is taken to create peak-amplitude events to measure their variability at windows of different sizes creating a descriptive scaling function. Events are chosen from maximal peaks at a 5 ms sliding window if they are above a set amplitude threshold. The function created consists of variability at 11 different window sizes producing an AF score per window size totaling to an 11-point function. In equation 1 T is

$$1) \qquad A(T) = \frac{\langle (N_i(T) - N_{i+1}(T))^2 \rangle}{2\langle N_i(T) \rangle}$$

timescale and N_i (T) is the event count per window i where variance is the average squared difference of windowed events normalized by two times the average event count per window size. Figure 3 shows an example of what those AF functions look like with the Buckeye Corpus data used. A flattening along the slopes of these functions indicates a loss in variability at those timescales while a rising

slope indicates an increase in variability.

The linguistic data from the Buckeye Corpus needed to be analyzed in a way that is comparable to how the AF method handles acoustic signals. First, to have measures at multiple scales of time we were able to have 4 levels of linguistic units (phonemes, syllables, words, breath groups). With these four features we tested the coefficient of variation (CoV) of their duration for which the AF method is a type of CoV measure. We also measured rate as number of linguistic units over length in seconds along with other information included in the corpus such as demographic data and individual participant variation. These measures were treated as predictors for generalized linear models predicting the linear slope of the AF functions as well as the individual 11 A(T) scores in the function.

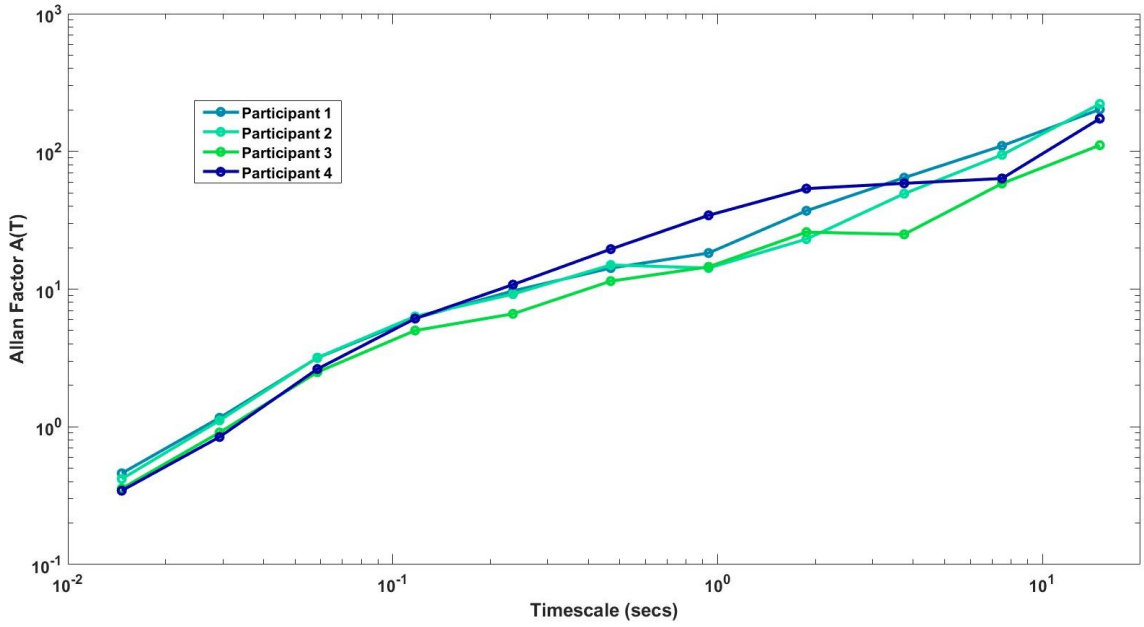

Figure 2.3: This graph contains example AF functions of 4 participants. Across the 11 timescales measured, their AF score scales to have typical slopes for speech categories as tested in [Kello et al., 2017]

2.4 Results

The results are based on models in which the extracted linguistic units become predictors for the AF functions. The audio recordings of the interviews were paired up with their transcriptions and then divided into four-minute sections; meaning that an AF function, rate, and CoV was measured across sections of the participant recording. It is important to note that the linear slope of AF functions is used here as seen in [Falk and Kello, 2017]. Furthermore, terms from quadratic fits of the AF functions were found to not give the best results as response variables. This means that the linear slope allows for models with much higher variance accounted for with the predictors used here and it makes the modeling assumptions a simpler case. Using the averaged per participant dataset we

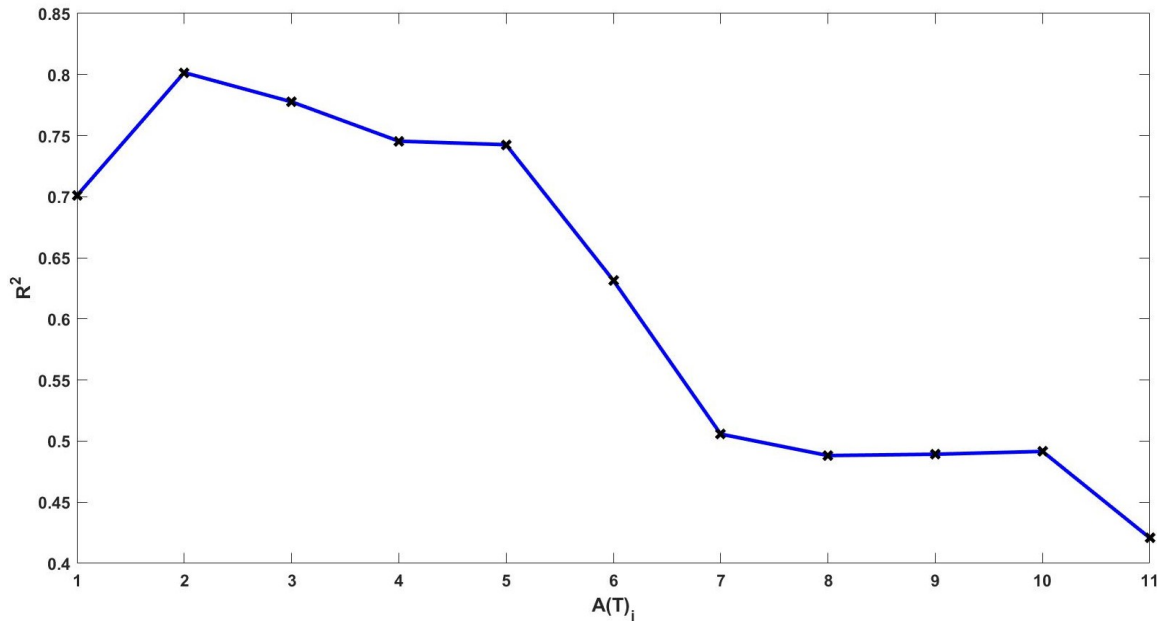

Figure 2.4: Change in R^2 for every $A(T)_i$ used as a response variable, where i is a level in the AF function.

tested ($m \sim \text{phone}_{\text{rate}} + \text{syllable}_{\text{rate}} + \text{word}_{\text{rate}} + \text{breathgroup}_{\text{rate}}$) the rates of the four linguistic features to predict the linear slopes of AF functions. This model had an $R^2 = 0.56$ with a $p < 0.001$ from the model F-Test, providing evidence for the rate of linguistics units corresponding to the AF of their acoustic record-

ings. Using the same data set, models were also trained per linguistic feature on each AF function level ($A(T)_i \sim \text{phone}_{\text{rate}}$, $A(T)_i \sim \text{syllable}_{\text{rate}}$, $A(T)_i \sim \text{word}_{\text{rate}}$, $A(T)_i \sim \text{breathgroup}_{\text{rate}}$). Figure 5 shows how models for each linguistic feature's

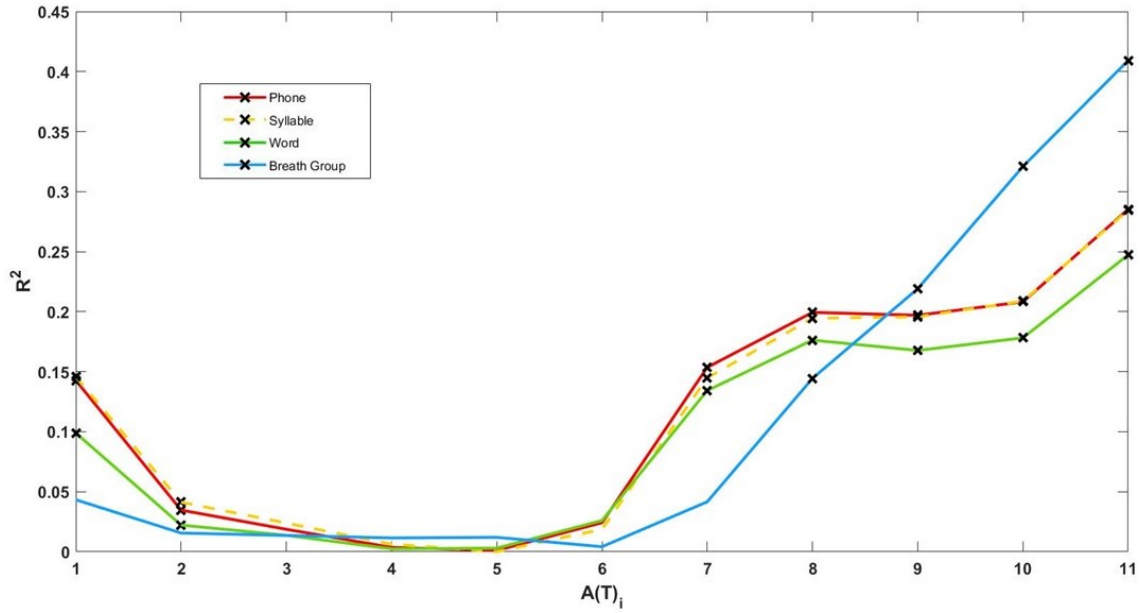

Figure 2.5: Change in R^2 for every $A(T)_i$ used as a response variable using each linguistic feature rate as a predictor. Markers of each line demonstrate $p < 0.05$ F-Test significant models.

rate respond differently across fast to slower timescales. The middle timescales for all features do not show up as significant. Most of the significant models appear at the longest timescales. It is also evident that the phone and syllable features are almost identical across all timescales. The largest linguistic feature demonstrated higher R^2 values in the longer timescales compared to the shorter linguistic features. Meanwhile, phone, syllable, and word had a single significant R^2 model value at the shortest timescale for which phone and syllable were higher than word. Using the same type of analysis, Figure 6 demonstrates how models for each linguistic feature's CoV respond across each timescale. In this case, only one model had a significant F-Test effect, and it came from the breath group level. Most models in Figure 6 have R^2 values close to zero and the only F-test significant model had an $R^2 = 0.1451$. At that same timescale in Figure 5, the R^2 was lower

than in Figure 6. Finally, models trained using the CoV as predictors showed R^2

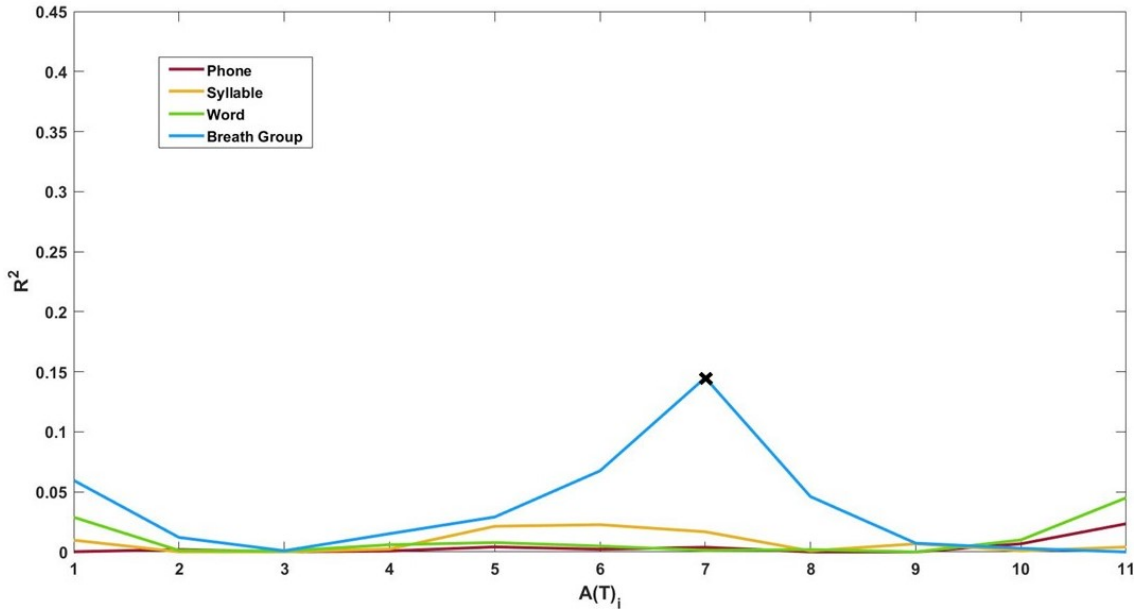

Figure 2.6: Change in R^2 for every $A(T)_i$ used as a response variable using each linguistic feature CoV as a predictor. Markers of each line demonstrate $p < 0.05$ F-Test significant models.

values close to zero. This is true when using both the per four-minute and averaged per participant dataset. When breaking it up to a model per timescale, as seen in Figure 6, it can be seen that there might be something going on at least with the breath group level. Prior studies using the AF method have found it useful to break up the linear slope of AF functions into fast and slow segments. Such a breakup can highlight whether there are relationships more correspondent to either fast or slow timescales. Here the linear slope of AF scores 1-6 was used to define the fast timescales and scores 6-11 for the slow timescales. Using the averaged per participant dataset, models were run comparing CoV and rate predictors of phone, word, and breath-group units e.g., ($m \sim \text{phone}_{\text{CoV}} + \text{word}_{\text{CoV}} + \text{breathgroup}_{\text{CoV}}$). To avoid issues of collinearity, we decided to exclude the syllable unit because of its almost identical effect in comparison to phones as seen in Figure 5. The results, as shown in Figure 7, demonstrate that although CoV cannot have a significant model for the slow timescales, that it does show up as significant with

an R^2 of 0.1981 for the fast timescales. On the other hand, rate models show a slight increase going from fast R^2 of 0.4094 and a slow R^2 of 0.47.

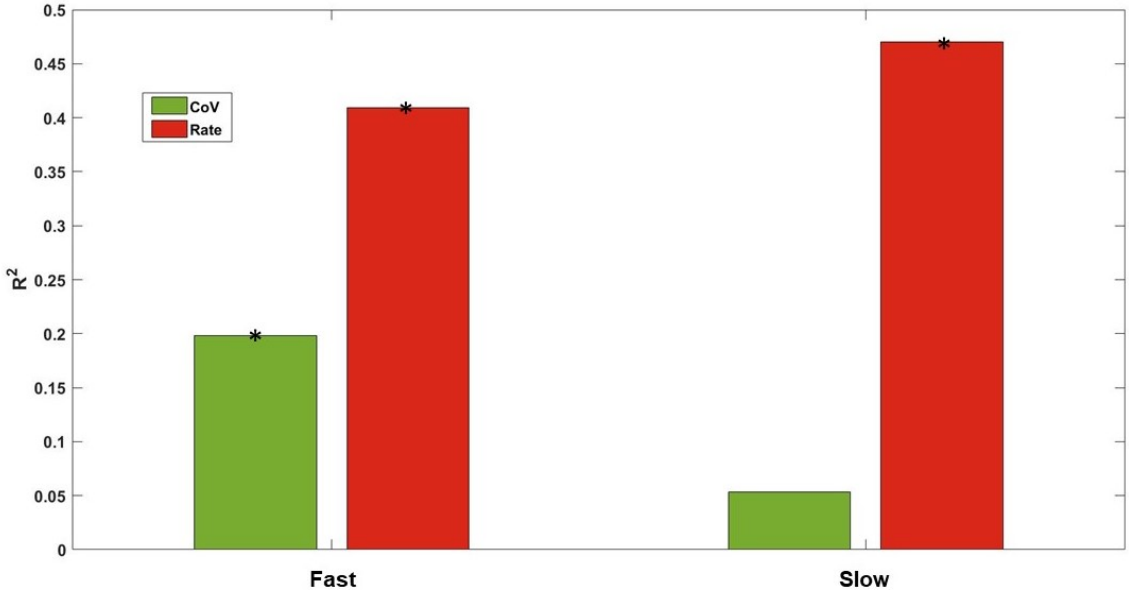

Figure 2.7: Model R^2 values of coefficient of variance (CoV) and rate predictors of fast and slow timescale slopes. Asterisks indicate $p < 0.05$ F-Test significant models.

2.5 Discussion

Using the Buckeye Corpus, this study was able to provide evidence for the HTS of speech recordings corresponding to the hierarchical structure of its own linguistic units. This extends results previously reported by [Falk and Kello, 2017] in which a limited sample was used. The corpus allowed us to be able to test one more linguistic feature than the previous study. The accessibility of phonemes helped to show that the relationship of phonemes and syllables to the AF functions were statistically the same. Therefore, we stick to highlighting the features with the least number of redundancies (phonemes, words, breath groups).

Across all linguistic levels, speaking rate gave the strongest relationship with the AF functions. In Figure 4 we can see that breath group has higher R^2 val-

ues in the three longest timescales. This makes sense because it demonstrates that longer linguistic features have more influence in the slower timescales when it comes to describing HTS. Similarly, both phonemes and syllables had a higher R^2 value than word and breath group in the fastest timescale. A previous study had looked at the effects of speaking rate and HTS as measured by the AF method. The study had compared how the HTS of speeches were different when the participant was manipulated to slow or speed up their speaking rate in comparison to when those changes are done algorithmically. Results from that study showed that participants were able to manipulate their HTS in fast and slow speaking rate conditions so that they differ in the middle to longer timescales. On the other hand, algorithmic manipulations allowed for a change in HTS at all levels for both conditions. Through the Buckeye Corpus we can see that for all the linguistic features, most of the significant relationship in the models are in the longer timescales [Schneider et al., 2020]. Figure 5 outlines how the middle to smaller timescales tend to have insignificant results. Despite their differences, the nested relationships of linguistic units in the corpus and the temporal variation of acoustic energy in the recordings share a similar outcome for pointing to speaking rate relating to the longer timescales in speech.

Relationships between individual timescales are not limited to speaking rate. In a prior study, the AF method was used to measure the extent of speech convergence between bilingual interlocutors. That study showed that the HTS of speech could be compared despite differences in languages or lexicon within a language. The between participants' results encouraged them to investigate variability within a participant and they correlated a person's own HTS across trials. It was found that a person is highly correlated with themselves, despite their HTS being flexible enough to converge with another speaker's, even when going from speaking English to Spanish. This demonstrated that people have their own unique HTS, which in speech has been shown to be akin to someone's own prosodic style. The convergence that occurs has also been shown to occur primarily in the longest timescales [Schneider et al., 2020]. Here it is argued that such a phenomenon is reflected by the data as seen in Figure 4 5. Participant identifier as a predictor for

AF functions was shown to have a significant relationship. Furthermore, when it was broken down and made into models for each level, we can see that the strongest models happen in the faster timescales and the strength drops off as the timescales become larger. This demonstrates that a person's unique prosodic style in their speech is primarily reflected by the fast temporal variations in their speech. Both when a participant tries to manipulate their own speech to be faster and slower or when their speech is changed through interlocutor interaction, the changes happen in the longer timescales and what remains constant is in the faster timescales. That is to say that it might be easier to control the prosodic variability of one's sentences as opposed to the variability of finer details such as the articulation of speech that occurs in about 30 ms.

Finally, a corpus wide model using CoV of linguistic units as predictors failed to show a significant relationship with the AF functions. The results from this may be a consequence in differences of approach and data used from [Schneider et al., 2020]. First, the corpus used here contained long recordings of natural spontaneous speech in comparison to shorter recordings of scripted stories. The corpus contained many natural pauses, interruptions, laughter, stuttering, and unverified words/vocalizations that act as noise in these natural settings. Furthermore, here the data was not limited to only utterance final and used an approach that would capture as much of the recorded speech while avoiding timestamp inconsistencies. As discussed previously, a dichotomy of effects between fast and slow timescales tends to be commonplace when using the AF method. Results in Figure 6 also gave some insight in detailing that CoV of linguistic unit durations as predictors may not be completely null. Figure 7 demonstrates that CoV of linguistic units show a significant model with a smaller amount of variance explained when predicting the linear slope of only the fast timescales. The model for the slow timescales shows a smaller amount of variance explained and no significant effect. On the other hand, Figure 7 shows that rate predictors in fast and slow models have a much larger amount of variance explained, with a slight increase in the slow timescales. Furthermore, correlations between rate and CoV for each linguistic unit also show that they are not related. CoV is also limited by the need to omit breath groups

that occurred adjacent to interviewer switches, as the annotation contained some erroneous timestamps. This led to omitting several shorter breath groups, which already is indicative of limits in the variance measured at that level.

The AF method produces scaling functions which not only reflect the HTS of acoustic energy but also of the nested relationship of linguistic units. The exact correspondence of AF timescales and linguistic unit levels is not completely clear, but the analyses here were able to replicate and extend results from previous studies. Because of that, we can at least know that shorter linguistic units (phone, syllable, and word) have more influence in the shortest timescales and that the largest linguistic unit (breath-group) has more influence in the three longest timescales. It may also be possible that speaking rate as used in [Schneider et al., 2020] and for every linguistic unit in our study is a more salient predictor for the variability in natural speech. From the prior speaking rate study, we know that speaking rate puts limitations to the variability in speaking prosody. To further study its effect, we would need to have data at the scale of the Buckeye Corpus that would also control for speaking rate. Participant demographic information (speaker gender, speaker age, interviewer gender) did not show a relationship with AF functions and points to the method as capturing universally produced patterns that are solely unique from person to person.

2.6 Conclusion

Lastly, it was evident that annotation of linguistic units complicated analyses focusing on encompassing multiscale relationships. Feature extraction was needed and at every level of human transcription and annotation that was needed, room for human error opened. This study argues for the utility of methodology in measuring the scaling complexity of speech amplitudes as it regularizes the analyses to reduce opportunities of human error or re-interpretation of specific linguistic units and what they may represent in the discourse. The traditional linguistic unit annotation vs time-series decomposition methods have their own trade-offs and when asking questions about natural discourse that are also comparable to

other behavioral measures, the time-series decomposition type of analyses seem to have an advantage while also showing a connection the original representational units in the language. The studies using the Allan Factor method therefore have an advantage as they can not only easily scale the measurement of naturalistic speech but also of any other correlated behaviors, which allows for a complexity science perspective on cognition to be extended and embodied. The following chapter will highlight the success of this methodology at capturing the HTS of modulated productions. Given that so far, we have seen the strength in utility for capturing both linguistic and extra-linguistic relationships when information between people is occurring in naturalistic settings, the next challenge will be to see how perceptual changes may offer insight to understanding how HTS is preserved in behavioral measures such as in cortical activity.

Chapter 3

Effects of Speaking Rate and Naturalness on Hierarchical Temporal Structure of Speech

Recordings of speech exhibit nested clustering of peak amplitude events that reflects the hierarchical temporal structure of language. Previous studies have found variations in nested clustering to correspond with variations of speech production seen in prosody and social interaction. In the present study, two specific dimensions of variation are tested in produced speech hypothesized to have differing effects on hierarchical temporal structure: Speaking rate and naturalness. Rate was manipulated both algorithmically and experimentally, and naturalness was manipulated using synthesized speech, with sinewave speech as a comparison. Allan Factor analysis was used to quantify nested clustering of peak amplitude events in speech recordings as a function of timescale. For fast speech, nested clustering was found to shift into shorter timescales, whereas for synthesized speech, nested clustering was found to decrease in the longer timescales. Results lead to a discussion on how the hierarchical temporal structure of speech signals are preserved when modulations in the control parameters of their production vary and its implications to theoretical predictions of how neural and perceptual processes might respond to such changes.

3.1 Producing Hierarchical Temporal Structure

On an everyday basis, people are responsible for producing a variety of informational signals including speech and music. Researchers have measured such signals and found that they commonly follow power laws [Kello et al., 2017]. It has been argued that these power laws underlying neural, behavioral, and social processes could be usefully theorized in terms of complex networks [West et al., 2008], because power laws are a natural consequence of their non-stationary, non-ergodic statistics. Ergodicity refers to stochastic processes where any random sample of measurement represents the average of the overall dynamics and stationarity refers to an unwavering shift of mean and variance of a system's dynamics over time. A fundamental question about complex networks, as well as cognitive and social systems, is how they respond to inputs from their environments. For example, the dynamics of complex perceptual networks are responsive to their sensory inputs, and language networks are responsive to inputs from verbal interactions. The former is an example of unidirectional influence because sensory systems do not directly affect the sensory world, only indirectly via the perception-action loop [Haykin, 2012]. The latter is an example of bidirectional influence because participants in language interactions directly affect each other as the variability of the amplitude in their behaviors adjust to coordinate with the other person they are interacting with.

This view of cognitive and social systems as complex networks leads to predictions based on theories of how complex networks respond to external inputs. Specifically, [West et al., 2008] formulated the principle of complexity matching, which states that complex networks are most responsive to perturbations that match their own temporal complexity. This is in reference to information transfer between networks, where the process is maximized if the information being produced by either system have corresponding bandwidths. Diverging bandwidths of information transfer here would mean that there is a smaller degree of overlap between information channels, meaning less information passing through. Complexity is measured in terms of exponents that define power laws in network activity which accounts for activity at multiple levels of the system at hand. In terms of

the measurement of specific power laws the matching corresponds to similarity in the exponents characterizing the networks in question, and their environmental inputs. The original work defined network activity in terms of $1/f$ noise and fractal time series of events, the latter being analyzed in terms of waiting times (inter-event-intervals) τ, where $P(\tau) \sim 1/\tau\mu$, and $1 < \mu < 2$ [West et al., 2008]. In the previous chapter, we saw that behavioral scientists have tested for complexity matching in human coordination and speech. Their experimental paradigms were in part motivated by the premise that human complex networks are highly adaptive [Baronchelli et al., 2013]. One testable hypothesis is that human complex networks may adapt by "bending" the statistics of their dynamics towards those of their inputs, to better match the environment and other complex networks. Matching is hypothesized to increase the response sensitivity of complex brain and behavioral networks. When inputs are power law distributed, matching manifests as a convergence in power law exponents of brain and behavioral networks towards the exponents of their inputs. Such flexibility in power law exponents would not be expected for less adaptive complex systems. With this we can begin to think of what differentiates power laws across natural systems, such as in seismic amplitudes, and whether adaptability can be considered in the same way for the varying natural systems [Marković and Gros, 2014].

Among the first experiments to explicitly test for complexity matching in human behavior, researchers examined the dynamics of finger tapping, and pendula being swung together [Stephen et al., 2008, Marmelat and Delignières, 2012]. The tapping experiment used a fractal metronome that participants tried to follow as closely as possible. Fluctuations in inter-tap intervals exhibited $1/f$ noise, and power law exponents matched those of their fractal metronomes, i.e. unidirectional influence of the metronome on tapping. By contrast, the pendula experiment showed that power law $1/f$ exponents of angular fluctuations converged with each other, instead of a fixed stimulus like a metronome. The swinging of one pendulum by one participant was affected by the swinging of the other pendulum by the other participant, and vice versa, via perceptual and physical coupling, i.e. bidirectional influence. Together, these two studies provide evidence that human

complexity matching can occur in response to stimuli in the environment, and also in response to human interactions. Both systems need to be coupled at their levels of interaction for a corresponding adaptive process to occur.

3.2 Coupled Dynamics

One of the most natural kinds of human interaction is speech, which has also been found to exhibit complexity matching effects [Abney et al., 2014]. Furthermore, coupling between systems is extended to the concept of shared common ground which moves beyond the previously mentioned physical coupling and perceptual coupling to a social framework of agreement. The authors in this example recorded pairs of individuals having conversations about friendly topics that were to not spur controversy, versus polarizing topics with conversational partners on opposite ends. They converted the speech waveform for each speaker into a series of acoustic onset events, and found inter-onset-intervals (IOIs) to be power law distributed like critical events of complex networks. Complexity matching was found not in IOI exponents, but in the power law clustering of events that reflects the hierarchical temporal structure of language. Specifically, Allan Factor (AF) functions for event series were closer together for conversational partners compared with baseline, but only for friendly topics for which speakers shared common ground. Polarizing conversations showed no detectable complexity matching, suggesting that the coupling of human complex networks is multifaceted across physical, psychological, and social factors.

[Abney et al., 2014] used the AF function to measure hierarchical temporal structure in speech waveforms recorded from conversations, over time scales of 30ms-30s. Variations in this range of time scales are perceptible to the human auditory system, and complexity matching suggests that auditory brain networks adapt the statistics of their dynamics to those of their acoustic inputs [Ding et al., 2016]. Given the relationship between complexity matching and psychological processes reported by Abney and colleagues, we hypothesize that hierarchical temporal structure in speech, as measured by AF functions, should be reflected in auditory ex-

perience by way of complexity matching in auditory networks. In support of this hypothesis, Kello and colleagues [Kello et al., 2017] found that the shapes of AF functions reflect at least three perceivable variations in complex acoustic signals: social interaction, prosodic variation, and musical composition. Greater nested clustering in peak amplitude events (as opposed to acoustic onset events) can be perceived as acoustic interactions among people, prosodic emphasis in speech, or metrical structure in music. These results are consistent with the working hypothesis at hand, but they are quite general and do not inform how specific variations in AF functions relate to specific variations in perceivable features of speech, music, and other complex acoustic signals. This is in reference to modulations of produced signals by people due to control parameters of their motor systems. Directionality of information transfer, as discussed previously, reframes whether complexity matching can be thought of as the passive act of resonance (bi-directional) and active adjustment (unidirectional), and here it is further discussed in terms of passive reflection (unidirectional) i.e., the hierarchical temporal structure of a stimulus signal being preserved and reflected in a perceptual system. Focusing on stimuli signals, it is of importance here to understand the extent to which someone can modulate their hierarchical temporal structure and if that can change predictions on complexity matching effects on perceptual systems.

3.3 Analyses of Speaking Rate and Naturalness

In the present study, we test two types of perceptual variations in speech that we predict to have differing effects on hierarchical temporal structure: Speech rate and naturalness. Previous studies have demonstrated consistent effects of speech rate on prosodic variation, the latter being shown to affect hierarchical temporal structure. For instance, [Jun, 2003] found that more syllables are packed into fewer accentual phrases at faster versus slower speaking rates, thereby reducing variability by reducing the number of accentual phrases. [Dellwo et al., 2003] varied speech rates in English, French, and German, and found reduced variability in consonant durations for faster versus slower speaking rates. A modeling study in

Mandarin indicated that the effect of speaking rate affects variability across several hierarchical levels of prosodic organization [Chen et al., 2014], consistent with a study of speaking rate in Mandarin [Tseng and Lee, 2004]. In summary, previous studies indicate that faster speech should reduce prosodic variability across hierarchical levels, and thereby reduce hierarchical temporal structure across a wide range of timescales. Speech naturalness is also predicted to affect hierarchical temporal structure, but in a different way compared with speaking rate. In particular, human generated speech is predicted to have more hierarchical temporal structure compared with text-to-speech synthesis, particularly in the longer timescales. Variability in prosodic intonation and timing is difficult for traditional text-to-speech synthesizers because they do not model the meanings of sentences or discourse contexts [Ze et al., 2013]. As a result, synthesized speech is often perceived as having flat affect compared with human-generated speech. Relatively flat affect should correspond with reduced hierarchical temporal structure in time scales on the order of a second and longer, as previously shown by [Falk and Kello, 2017]. They measured AF functions in recordings of German-speaking mothers reading a story or singing a song, either to their infants or to other adults. The exaggerated prosody of infant-directed speech resulted in generally steeper AF functions, but the authors did not report a more fine-grained analysis. With respect to naturalness, [Kello et al., 2017] showed that AF functions for synthesized speech were flatter than those for natural speech, but again, the authors did not quantify the effect, nor did they compare it with speaking rate.

Here AF analyses of fast versus slow speech is reported, as well as natural versus synthesized speech. The analyses are designed to measure more stringent hypotheses about perceivably different effects of these manipulations on hierarchical temporal structure. Specifically, faster speech is predicted to result in less variability across all perceptible timescales, which should correspond with shallower, flatter AF functions. By contrast, synthesized speech is predicted to result in less variability in the longer timescales only, which should lead to shallower but more curved AF functions due to selective effects on longer timescales. The effect of speech rate is tested using both algorithmic and experimental manipulations,

whereas the effect of naturalness is tested using two different algorithmic manipulations. For the latter, results are compared with synthesized versus sinewave speech [Remez et al., 1981]. With sinewave speech being a synthetic control that retains most of the hierarchical temporal structure in the original signal.

3.4 Methods

3.4.1 Data Acquisition

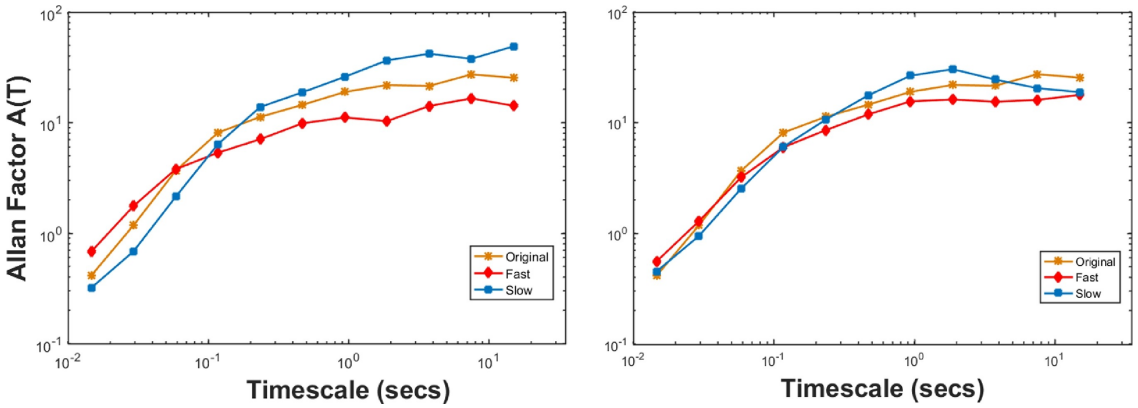

Figure 3.1: Left: AF functions of the original Obama speech, and fast and slow versions. Right: AF functions of the fast and slow teleprompter conditions.

Analyses of speaking rate were based on Barack Obama's address at George Mason University on the 21st Century Economy (1/08/09, 17:08 mins). The élastique algorithm (https://products.zplane.de/) was used to manipulate speaking rate without affecting the vocal pitch. The "fast" condition was 2x faster than the original recording, and the "slow" condition was 2x slower. In addition to these algorithmic manipulations, an experiment was conducted in which ten University of California Merced students read two excerpts from the speech off a teleprompter. Half of the participants read the first excerpt at a slow pace and the second at a fast pace, and vice versa for the other half. On average, the fast paced and slow paced excerpts took 4.5 and 10.1 min to complete, respectively. Participants were instructed to read the speech from the teleprompter as smoothly as possible, and their readings were recorded for subsequent acoustic analyses.

Analyses of naturalness were based on ten recordings of TED talks (mean length = 6.41 min, SD = 1.14 min) reported by [Kello et al., 2017]. The TED intro and outro theme was trimmed from the recordings, along with any applause at the beginnings or ends of the talks. A synthesized version of each talk was created by submitting the transcript to Google speech synthesis, and recording the output. The synthesized versions (mean length = 6.62 min, SD = 1.16 min) were recorded using Garage Band version 10.1.0. Garage Band was also used to set the lengths of the synthesized recordings roughly equal to the original recordings (within ± 30s). Lastly, sinewave speech recordings (mean length = 6.46 min, SD 1.16 min) were created from the ten trimmed TED talks by using the Matlab sinewave speech code provided by [Ellis, 2004], with default parameters provided by Haskins Laboratories. The software tracks speech formants and assigns a single sinewave to each one. The sinewave amplitudes and frequencies are modulated to track the formants over time. The result is a combination of whistling sounds that preserve most of temporal structure in speech. Sinewave speech is typically perceived as speech-like, but the words spoken are difficult to discern unless the listener is given information about what is being said.

3.5 Results

Audio recordings were analyzed using the same method as reported in [Kello et al., 2017]. Details can be found there, but briefly: Each recording was divided into four-minute segments, and analyses were averaged across segments to yield a single AF function per recording. The Hilbert envelope was calculated for each segment and peaks above threshold were analyzed as time series of acoustic events. An AF function was computed for each segment where T is the timescale,

$$1) \qquad A(T) = \frac{\langle (N_i(T) - N_{i+1}(T))^2 \rangle}{2\langle N_i(T) \rangle}$$

$N_i(T)$ is the event count in each window i, and $A(T)$ is AF variance. AF variance captures the degree of event clustering at a given timescale, and for timeseries with nested clustering, $A(T)$ increases with T. Self similar clustering across timescales

yields a power law, $A(T) \sim T^{\alpha}$, where $0 < \alpha < 2$. The AF function was computed for 11 values of T in between 15 ms and 15s, logarithmically spaced to compute the orthonormal basis. AF functions for speaking rate analyses are shown in Figure

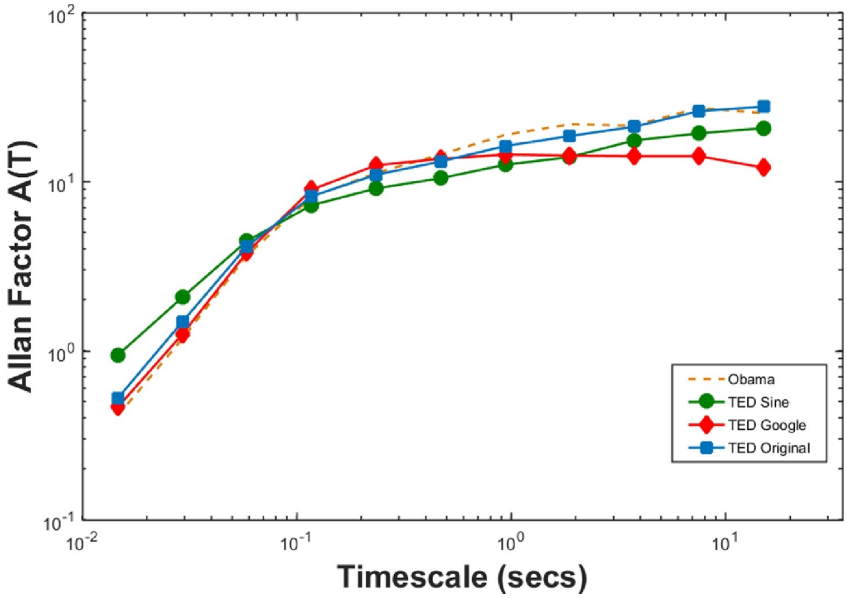

Figure 3.2: Mean AF functions for TED talks and their two different synthesized versions, Google text-to-speech and sine wave speech. The AF function for Obama's speech is shown for comparison.

3.1. The left panel shows the effect of algorithmic speaking rate manipulations on the original Obama recording, and the right panel shows mean AF functions for the slow and fast teleprompter conditions, with the original Obama recording as a reference. AF variance for the Obama recording steadily increased as a function of timescale, consistent with analyses of TED talk recordings reported by [Kello et al., 2017]. [Falk and Kello, 2017] found evidence to suggest that this AF shape is common to speech because it reflects the nesting of linguistic units like syllables in words, words in phrases, and phrases in sentences. Figure 3.1 shows that an algorithmic increase in speaking rate causes clustering to generally shift left into the shorter timescales, whereas an algorithmic decrease causes a rightward shift into the longer timescales. Figure 3.1 also shows that the teleprompter had a similar effect, except that there was a drop in AF variance at the longest timescales for slow speaking rates. We hypothesize that this drop comes from the artificially

even pace of speaking caused by the slow, even pace of the teleprompter. This evenness creates isochrony and isochrony reduces clustering and hence AF variance. We leave it for future research to test this hypothesis explicitly.

AF functions for naturalness analyses are shown in Figure 3.2. The mean AF function for the original TED talk recordings has the same basic shape as that for the original Obama recording. This similarity is consistent with [Kello et al., 2017] who found that monologues have common, distinctive AF functions compared with dialogues and singing TED talks and the Obama speech are both types of monologues. AF functions for synthesized versions of TED talks were very similar to the original recordings in the shorter timescales, but they diverged in the longer timescales. Specifically, synthesized AF functions were flat compared with original recordings, which indicates a lack of nested clustering in timescales corresponding with prosody and intonation. By contrast, AF functions for sinewave speech had the same overall shape as the TED talk recordings from which they were created, with a slight leftward shift of clustering as if the sinewave speech rate was faster than the original recording.

The perceptual distinction between natural and synthesized speech is very clear, as is the distinction between slow versus fast speaking rates. Moreover, these two dimensions of variation are perceptually distinct from each other. The effects of speaking rate and naturalness were also different from each other, as verbally described above, but it is necessary to quantify this difference to better understand it and relate it to complexity matching. To do so, we fitted a second-order polynomial to each individual AF function, which allowed us to capture their convex shapes in terms of linear and quadratic coefficients.

Coefficients are plotted in Figure 3.3 for fast and slow speaking rates, as well as natural and synthesized speech. The graph shows that speaking rate had the same effect on linear coefficients but opposite effects on quadratic coefficients. Fast speech was comparable to synthesized speech in that linear coefficients were closer to zero compared with slow speech and natural speech, respectively. This similar result was due to the overall flattening effect of these conditions. However, fast speech was less convex than slow speech, whereas synthesized speech was more

convex than natural speech. This difference was due to the selective effect of synthesis on longer timescales, versus the overall effect of speaking rate across all measured timescales. Finally, sinewave synthesis had a small effect on coefficients akin to the effect of fast speech. It would be interesting to test whether sinewave is perceived as being faster than normal speech, even though the same signal variations unfold over the same time periods.

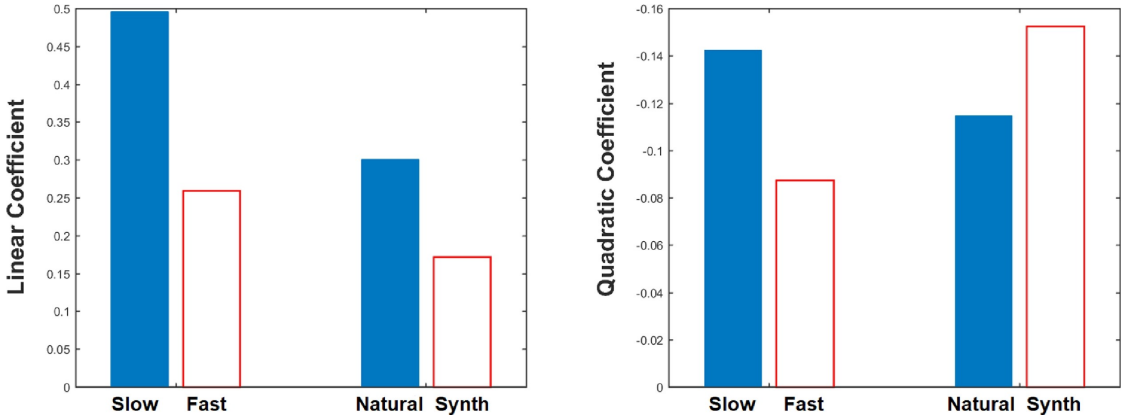

Figure 3.3: Linear and quadratic coefficients for fast versus slow speech, and natural versus synthesized speech. The two different manipulations had the same effect on linear coefficients, but opposite effects on quadratic coefficients.

3.6 Discussion

In the present study, we investigated the effect of manipulating speaking rate and naturalness on hierarchical temporal structure in speech. Using AF analysis, we showed that nested clustering in peak amplitude events is affected differently by these two manipulations. Changes in speaking rate shifts the entire measured hierarchy into shorter or longer timescales, whereas changes in naturalness flatten or steepen the longer timescales of the hierarchy, i.e. on the order of seconds and longer. Other studies have shown that acoustic events in speech appear to be crucial events [Abney et al., 2014], including a recent study by [Pease et al., 2018] in the special issue edited by [Grigolini, 2017]. Taken together, these studies suggest that neural and perceptual processes may be highly responsive to speech inputs

by means of complexity matching. Specifically, power laws in neural and perceptual dynamics may take the general shape of power laws in speech dynamics by means of complexity matching, while having distinct trajectories because of myriad differences in neural versus acoustic "substrate", so to speak. The present results are consistent with this application of complexity matching, in that the different perceptual experiences associated with speaking rate and naturalness have corresponding differences in hierarchical temporal structure. These perceptual differences may have their roots in complexity matching of auditory networks within coming speech signals.

The application of complexity matching to speech perception leads to questions about how power laws in auditory networks are affected when temporal structures in speech signals do not follow a single power law. [Kello et al., 2017] showed that many categories of speech and music deviate from power law AF functions. In fact, the only categories that closely followed a power law in nested event clustering were classical music and thunderstorms. Monologues like those analyzed herein were consistently found to have a distinct flattening in the longer timescales, and the shape of this deviation varies as a function of speaking rate and naturalness. What do such deviations imply for complexity matching?

One possibility is that neural and perceptual dynamics become less responsive to speech dynamics when they deviate from a power law, because brains are attuned to power laws in sensory inputs. As previously mentioned, the principal of maximal information transfer underlies complexity matching effects, and the proposed hypothesis would mean that deviations in production also create a divergence in the informational bandwidth between interacting systems. Another possibility is that neural dynamics bend along with the dynamics of speech being listened to. The latter would correspond to a neural correlate of perceiving and following the sounds of speech. In this case, a unidirectional relationship is outlined like in the finger tapping experiments where the metronome would be the speech signals and cortical activity would make an effort to adjust its endogenous deviations despite it being off center to perfect $1/f$ dynamics. The same question can also be asked of music, with the same possible hypotheses [Ding et al., 2017].

Indeed, the effect of prosody on temporal hierarchies in speech has been argued to have an analog in music [Hausen et al., 2013, Palmer and Hutchins, 2006]. This analog leads to the idea that music perception, as hypothesized for speech perception, may be partly supported by a form of complexity matching that enables temporal hierarchies in neural dynamics to conform to those of speech and music. Lastly, here another possibility is put forth when considering the brain's need to maintain homeostasis through the delicate balance of being at a sub-critical state [Bak et al., 1987]. With this, the hierarchical temporal structure of cortical activity as a whole would be limited in how much it could bend towards a perceived signal. A more sensible expectation would be that specific neuronal pathways and perceptual systems have varying degrees of corresponding change in affect. Therefore, this phenomenon could be better framed as an act of reflection of hierarchical temporal structure because neuronal activity will not be able to completely bend towards environmental factors but rather reflect in part some facets of the originally perceived signal.

3.7 Conclusion

Through the experiments presented in this study, examples of how hierarchical temporal structure changes via modulations in their production were outlined. Changes varied from a nonlinear shift of nested clustering in AF functions to a simple flattening or swelling on longer timescales. Despite some of these changes being likened to that structure of other natural systems, such modulations did not completely erase what makes these AF functions in speech unique. Given that the stimuli were able to maintain some of its identity in AF functions, such a persistence can also be expected from the brain's own endogenous dynamics. Complexity matching effects are seen under different coupling relationships, but here results from modulation to hierarchical temporal structure allows for fine tuning predictions from how we might see the structure of stimuli in measured brain responses. The following chapter will dive into this through the AF methodology and what limits such a methodology may have. A practical hypothesis is presented

here so far, which frames complexity matching effects in empirical studies as an act of reflection of hierarchical temporal structure in cortical activity. The intersection of the complexities between endogenous neuronal dynamics and epistemological limits of the current empirical methodology further cements the endeavor as the attempt to find reflections of hierarchical temporal structure in the vast waves of cortical activity.

Chapter 4

The Search for Auditory Stimuli Structure in EEG Responses

Coordination of cognitive systems have been seen to follow complexity matching principles in their information exchange. Experimental approaches have sought methods to quantify the statistical complexity of signals in these systems to infer the extent of matching across scales of time that may occur during interaction. One type of coordination demonstrates a system that plays the role of the perturber and another one as the perturbed. In the present study, auditory stimuli are framed as perturbations to auditory processing in the brain, and whether these perturbations can be captured in EEG signals through complexity matching is tested. The Allan Factor method is adapted into a novel experimental paradigm to try to capture hierarchical temporal structure of EEG recordings from natural auditory stimuli that allow for longer, natural stimuli lengths without the limits of repeated stimuli exposure. Current results of this methodology affirm the potential for capturing hierarchical temporal structure reflected in EEG responses, but complexity matching stays elusive through limitations of localization, feature extraction, and noise in the signal. This chapter ends in a discussion on methodological tradeoffs and outlining which approaches can be most conducive to advancing and confirming theoretical predictions.

4.1 Processing Temporal Structure

Music and language are both constructed from smaller to larger units that hierarchically organize content e.g., determiner & noun to build a noun phrase in language and varying note pitches that build a larger harmonic phrase. This hierarchical construct has been predicted to be parsed by similar neural pathways whether the incoming information is categorized as music, or language [Patel, 2003]. The shared processing pathways in the cortex are indicative of the importance of hierarchical temporal structure of stimuli and perceptual systems, as evidence from these types of experiments seem to suggest a domain general approach [Tao et al., 2021]. The previous chapters have given various data and examples of how hierarchical temporal structure is shared between natural systems and framed it as a structural feature important to understanding how people can interact and be coupled to the natural world. A domain general account of perceptual systems and the intersection of complexity in coupled interactions defines a generalized adaptiveness. In the previous chapters, this type of adaptiveness has been useful in explaining how information is maintained despite modulations to its production and how coordination between systems can emerge. First, let us look at auditory perception research starting from simple stimuli responses to complex multilevel responses, to build up to how hierarchical response processing may inform methodological procedures and experimental paradigms trying to verify complexity matching predictions in neuronal activity.

Neural entrainment research has demonstrated that neural oscillations are sensitive to temporal features, as seen when participants synchronize to perceived metrical onsets of an isochronous auditory stimuli [Tal et al., 2017]. Experiments controlling for stimuli frequency have shown that faster metrical rhythms show a corresponding array of entrained neural responses that are also oscillating at faster frequencies as well as slower responses for slower rhythms [Musacchia et al., 2014]. In addition, pitch variance in auditory stimuli can be used to induce a stronger meter percept. Stronger identification has been shown for stimuli that have shorter semitone distances, which further tests the multidimensional aspect of complex auditory stimuli [Jones et al., 2006]. These results are simply a small sample of

published experimental results demonstrating that you can model auditory perception as the coupling of self-sustained oscillators. Such a modeling is argued here to have been advantageous not just in extending core physical theoretical predictions of coupled oscillators but also in allowing for a practical simplification of experimental design. In fact, many of collected auditory processing data have been shown to be compatible with the neural entrainment framework [Rimmele et al., 2018]. The limits to this approach can be described as a type of spot checking of multidimensional perceptual features in auditory processing because of its reliance on analyzing one level of interaction at a time.

To mend such limitations, researchers have designed experiments via a cortical tracking framework. In this case, a theoretical assumption of neural processing being anticipatory is put forth to release restrictions of one-to-one synchrony at specific scales of measurement. This also shows importance to the complexities of the overarching endogenous dynamics of the cortex being not simply noise to be filtered out but rather information to be better captured. With respect to complex stimuli analyzed hierarchically, the use of natural speech stimuli presented at isochronous intervals has been used to demonstrate hierarchical cortical tracking of linguistic units [Ding et al., 2016]. Neural responses show tracking occurring at the sentence level (1 Hz), phrase level (2 Hz), and syllable level (4 Hz). Similarly, evidence for a hierarchical system in auditory processing is posited by demonstrating tracking at lower frequencies of phonemic units in natural speech 155 s [Di Liberto et al., 2015]. Longer stimuli are also used in the attempt to analyze hierarchical processing of music and speech with lengths closer to their natural lengths ∼4:15 mins [Farbood et al., 2015]. They present conditions with scrambled structure at three different timescales of the music separately to trained musicians. Using fMRI, larger musical timescales were found to be processed longer when approaching higher order brain topography in the auditory cortex. It is evident that such an approach was helpful in opening up experimental controls to allow for more naturalistic stimuli, but still, many of these experiments found it difficult to move away from stimuli presentations at regular intervals.

Natural discourse and everyday events in auditory perception are not so cut

and dry. Anything from talking to a friend whose spontaneous laughter, coughs, or limited attention can disrupt a story they are telling you, to making inferences from heterochronus streams of events at a social gathering aka., 'cocktail party effect', deviate from the isochrony of stimuli in traditional laboratory experiments [Golumbic et al., 2013, Rimmele et al., 2015]. Even in music, with its often-characterized rhythmicity as a stimulus, presents cases for aperiodic meters still building up temporal expectations such as in dancing [Polak, 2020]. This is all to highlight that the temporal processing of auditory events goes beyond the entrainment to periodic stimuli and that experimental methods interested in capturing that interaction should be primarily focused on its naturalness, and the hierarchical nature of auditory processing.

4.2 Allan Factor Paradigm

The main reason for tracking experiments to present speech in isochronous rates, is because unlike music, speech stimuli is defined as being pseudo-rhythmic [Nolan and Jeon, 2014, Ten Oever and Martin, 2021]. Moreover, results and discussion from chapter 2 have already demonstrated how much messier data from natural discourse can be. In the recordings of the Buckeye corpus, annotations captured many non-speech productions such as coughs, laughter, stuttering, and miscellaneous noises. This is all on top of the natural pseudo-rhythmic nature of speech. Nevertheless, the Allan Factor analysis of peak-amplitude events from acoustic recordings was robust enough to still capture its hierarchical temporal structure, replicating the identifiable scaling nature of conversations from previous studies [Falk and Kello, 2017, Kello et al., 2017]. Modeling work of linguistic units from the annotated Buckeye corpus also replicated a previous study showing a relationship between the temporal hierarchy in language and speech [Falk and Kello, 2017]. This highlights the relevance of such a method to capture meaningful scaling of variance in a complex and noisy signal such as speech. This is the case because the method in part includes some feature extraction steps, since the actual variance comes from events captured. In studies involving movement

and speech, researchers have found it useful to extract events from vocalization onsets to better compare the lower frequency of movement events captured through a frame differencing method [Abney et al., 2021]. In studies only analyzing speech [Schneider et al., 2020, Ramirez-Aristizabal et al., 2018], it was useful to look at peak-amplitude events which are selected by a max value and threshold parameters. In brief, the threshold parameter was used to set irrelevant events to values of zero and all peaks above it to 1. The threshold parameter defined an amplitude relative the sample rate and length of the recording so that recordings would be comparable, and variance not affected by trivial factors such as a longer recording or a recording with higher sampling rate. Then maximal peaks were identified using a $\pm$ 5 ms sliding window. Both parameters allowed for the regularization of features across various recordings while also filtering out irrelevant noise or activity with less salience in the signal.

If the Allan Factor method can handle noise in natural speech, along with environmental noise in animal vocalizations and thunderstorms [Kello et al., 2017], then it gives confidence to try it with the noise associated from brain responses recorded from electroencephalogram (EEG) systems. Therefore, the goal of this study is to test the Allan Factor method to capture hierarchical temporal structure (HTS) in EEG responses and investigate the extent that complexity matching principles can be applied. Stimuli are taken from the [Kello et al., 2017] study to test an array of distinct types of sounds and unique HTS categories. Each type of stimuli is only presented once at full length of recording. Time series distributions are parsed from peak amplitude events in the EEG recording. Independent component analysis is used to localize data in topographic points of interest within frequency ranges of 1 - 50 Hz. Components located in the shared regions of activation (auditory cortex) will be the primary focus for testing HTS preservation in brain responses.

4.3 Methods

4.3.1 Data Collection

The approach taken here sought to collect cortical responses through EEG to an array of 6 auditory stimuli e.g., electronic music, bird song, ted talk, sine transform of ted talk, classical music, and the same classical song repeated. The lengths of the original stimuli ranged from 4:20 – 4:42 mins and the EEG recording lengths where up to 4:20 mins. The audio was down sampled to the sampling rate of the EEG system (32 channel ANT Waveguard electrode cap) at 2056 Hz. There was a total of 11 participants and the presentation of the stimuli was randomized between subjects. Headphones were used and the volume was adjusted to comfortable levels for the participants. During the presentation of stimuli, participants were told to stare at a black screen and to minimize unnecessary movements including eye blinks. Collected data was further processed to remove muscle artifacts and unwanted noise. The removal of 60 Hz sinusoids was implemented to clean up noise from electrical appliances. A 0.1 Hz high pass filter was used to remove drift from data. Bad channels were rejected using a probability function predicting the probability a channel recorded meaningful data.

4.3.2 ICA

Adaptive Mixture Independent Component Analysis (AMICA) was used to localize EEG components that topographically cluster in the auditory cortex and other points of interest. Independent Component Analysis (ICA) weights were retrieved from modified recordings which were down sampled from 2048 Hz to 1024 Hz and band passed at 1-50 Hz. Then those weights were transferred to the original EEG recordings which were then down sampled to 1024 Hz. The EEGLab plugin AMICA was used with 3 models using posterior probabilities for each model to keep the best components among the different models. K-means clustering was used taking into account spectral, and topographic information of the components across participants as seen in Figure 4.1. An n of 9 was used, which placed 9 distinct means and classified components with similar means into a cluster. The

n parameter was also a practical choice, because it allowed for the creation of a cluster near the auditory cortex with at least one component per participant and condition. The Allan Factor of the auditory cortex clustered ICA components were taken to test HTS, as well as any other components of relevance.

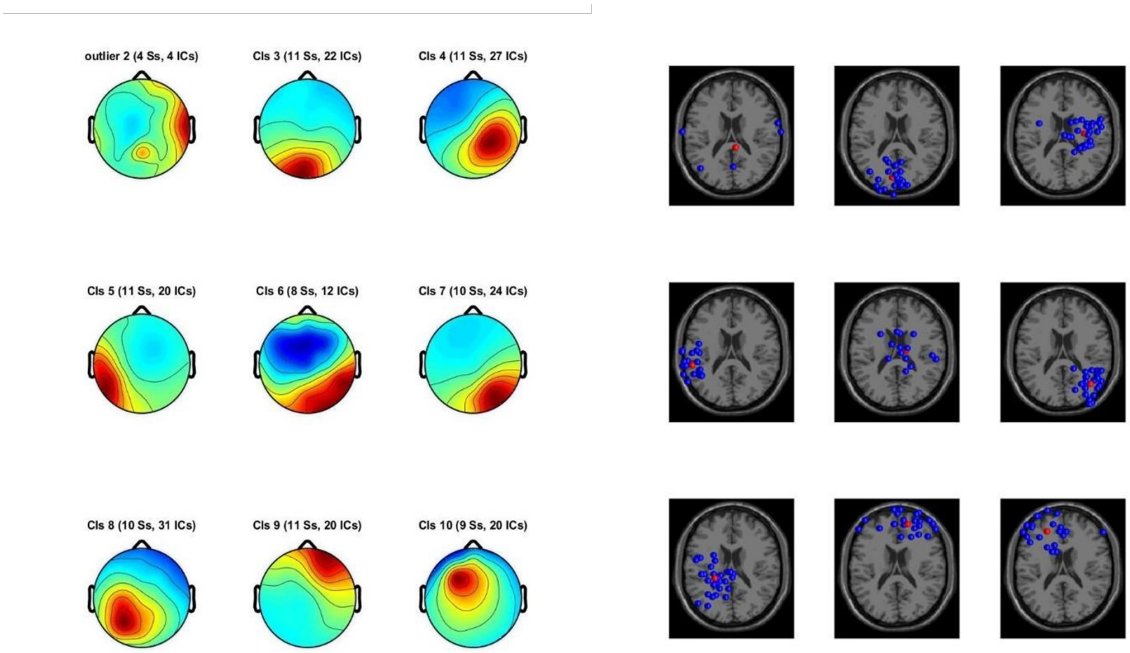

Figure 4.1: AMICA components of individual participants were put through K-means clustering and broken down into 9 clusters containing components across participants. First panel (left) shows a heatmap of spectral clustering and the second panel (right) breaks down components by dipole fitting locations. The first cluster at the top left of both panels contains components that were located outside the cortex and treated as artifacts, as well as components that were over 15% variance in the group.

4.3.3 Allan Factor

Auditory stimuli and EEG components were analyzed using Allan Factor. First, signals were divided into four-minute segments, and the Hilbert envelope was calculated for each segment. Analyses were averaged across segments to yield a single AF function per signal. Peak amplitude events were then selected using two pa-

rameters. The first parameter identified maximal peaks within a ± 5 ms window, then all peaks were preserved if their amplitudes were over the H parameter, a breakdown of this process is seen in Figure 4.2. Selected events created a time series in which the AF statistic was implemented on as follows: where T is the

$$1) \qquad A(T) = \frac{\langle (N_i(T) - N_{i+1}(T))^2 \rangle}{2\langle N_i(T) \rangle}$$

timescale, $N_i(T)$ is the event count in each window i, and $A(T)$ is AF variance. AF variance captures the degree of event clustering at a given timescale, and for a time series with nested clustering, $A(T)$ increases with T. Self-similar clustering across timescales yields a power law, $A(T) \sim T^{\alpha}$, where $0 < \alpha < 2$. The AF function was computed for 11 values of T in between 15ms and 15s, logarithmically spaced to compute the orthonormal basis.

4.3.4 Classification

Machine learning through Matlab's Classification Learning application was used to test a parametric training approach to finding differences in variances across scales of time. A matrix of the input data was formatted using the AF scores of each ICA component from every participant and from the four clusters. The target response for classification was the six stimuli categories and the predictors were the eleven AF values which account for the eleven timescales in the analysis. Different combinations of predictors were tested as well by adding the cluster labels and another case which tested the response variable of the four clusters for classification. Training for classification used all the Support Vector Machine (SVM) options which include linear, quadratic, cubic, fine gaussian, medium gaussian, and coarse gaussian SVMs. The data was split and tested using a 5-fold cross validation method. This approach for using AF values for label classification is akin to the approach by [Kello et al., 2017].

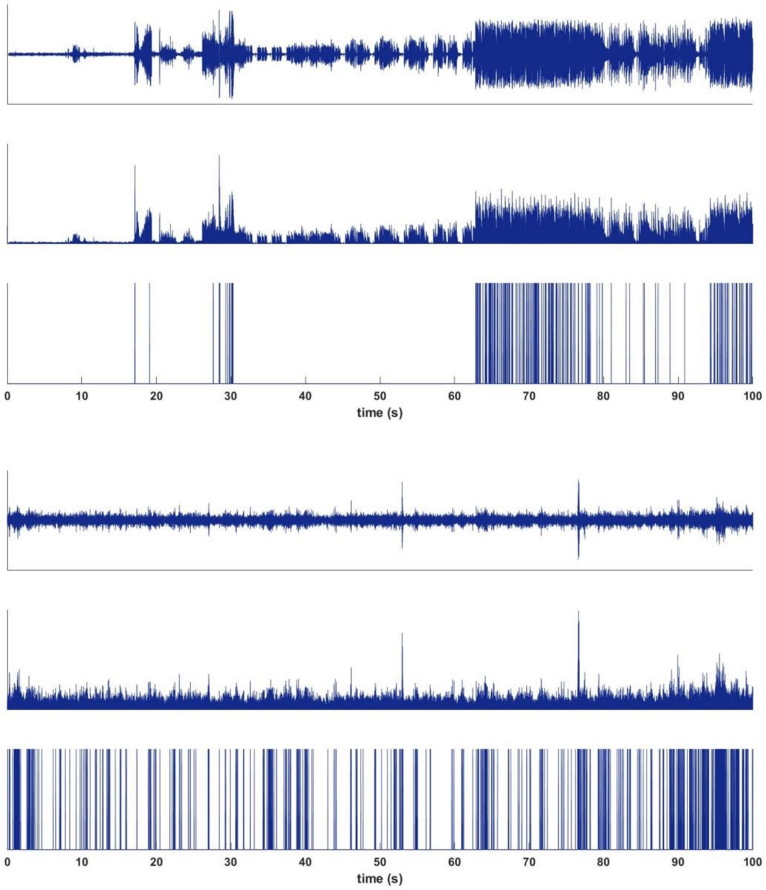

Figure 4.2: An example of peak amplitude event selection into a time series is shown for the electronic dance music stimulus (top) and a corresponding EEG response (bottom) of a 100 second length. For both top and bottom panels, the first section shows the signal waveform, followed by its Hilbert envelope, and lastly the time series in which the AF statistic is applied to.

4.4 Results

The first set of results presented demonstrate clear distinctions in AF function shapes for the stimuli (Figure 4.3). An exception is observed in the similarity of event clustering at the longer timescales for the Ted talk and its sine wave transform. The characteristics of the stimuli AF functions demonstrate scaling expected from their categories as seen in the [Kello et al., 2017] study.

The next set of analyses took ICA decompositions of individuals and organized them into 9 clusters through K-means to locate regions of interests within the cor-

tex. The initial focus was on doing AF on ICA components that fell into a cluster close to the auditory cortex, but the shapes of AF functions looked similar to $1/f$ scaling. Then this was tested to see if the pattern persisted across clusters. Clusters that had at least one component for each participant per condition were taken into account for this analysis. Four clusters met that criterion and they belonged to the frontal cortex, occipital, sensory-motor (right hemisphere), and auditory cortex (left hemisphere) regions (Figure 4.1). All clusters were shown to follow the same $1/f$ pattern despite their differences on topography. The clusters were averaged together and plotted by condition (Figure 4.4). Despite AF functions having $1/f$ scaling throughout, the longest timescales showed a notable difference. Linear fit slopes of the AF functions of all components were taken to test for differences. Figure 4.4 shows the mean linear slopes of ICA components when organized by condition and by order of condition presented. A two-way repeated measures Analysis of Variance (ANOVA) of linear slopes shows a p-value of 0.0633 and the one for condition order has a p-value of 0.0969. A Tukey's Honest Significant Difference Test is performed on both conditions' and condition orders' linear slopes with the only significant difference occurring in between the Symphony 1 and Symphony 2 (the second time participants heard the symphony stimuli) condition slopes with a p-value of 0.0471.

Classification learning of AF values demonstrated results that support the above presented statistics. The first case had only AF values as predictors and the stimuli categories as response variables. In all SVMs the classification did not go above the chance classification rate of 16.67%. When the cluster labels and participant labels were added as predictors it would lower classification performance. The shorter timescales and the longer timescales were also tested by breaking up the classification testing in two. The shorter timescales used only the first three AF values as the predictors and yielded lower performance than chance, with the best performing SVM yielding 15.7%. The longer timescales used the last three AF values and had performance higher than chance at 22.3% with a fine gaussian SVM. The second classification case involved the cluster labels as target variables and the AF values as predictors. The best SVM classification came from

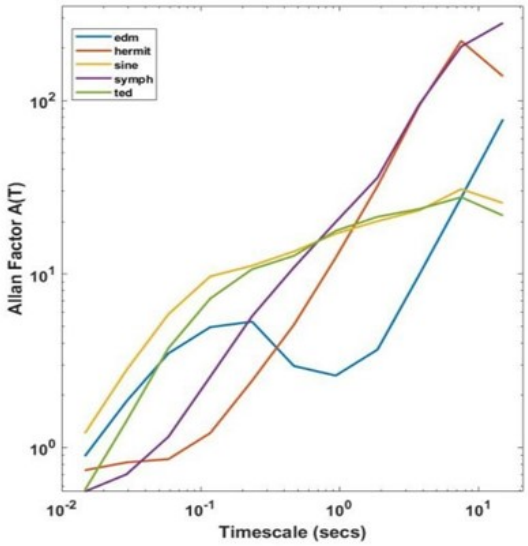

Figure 4.3: AF functions for the down-sampled auditory stimuli are presented above with 'hermit' referring to the Hermit thrush bird song and 'symph' referring to classical music symphonies.

the quadratic SVM with a performance of 28.4% above chance (chance 25%) from the four cluster categories. The frontal lobe cluster showed to be the most unique in its classification performance with the highest rate of true positives at 68%.

	EDM	**Hermit**	**Sine**	**Symph1**	**Symph2**	**Ted**
Song Means	1.0457	0.9587	0.9858	1.1355	0.9237	0.9799
Song SEs	0.0504	0.0565	0.0543	0.0575	0.0462	0.0495
	1	**2**	**3**	**4**	**5**	**6**
Order Means	0.9664	0.999	1.0815	1.0815	1.0154	0.8928

Table 4.1: The average AF slopes from ICA components organized conditions and condition order and their standard errors are presented. Slopes come from a linear fit on the longest timescales where the biggest differences exist.

4.5 Complexity Matching

A direct complexity matching effect such as in speech studies [Abney et al., 2014, Schneider et al., 2020], was not seen using the current methodology. The AF func-

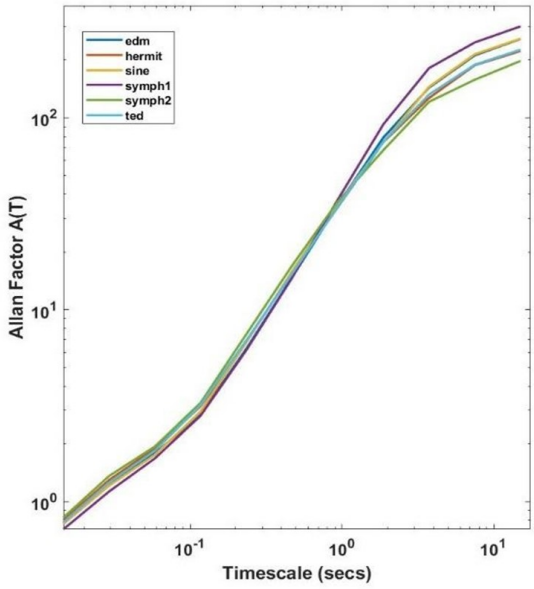

Figure 4.4: Aggregated Allan Factor functions of ICA components are averaged by condition across clusters.

tions of brain responses collected in this experiment at the individual EEG channel level and ICA compositions from different topographic cluster locations all yielded AF functions akin to $1/f$ event clustering. This finding replicates studies of spontaneous brain activity of healthy brains which is explained as a sign of self-organized criticality [Freeman and Holmes, 2005]. Complexity matching principles were tested by [Allegrini et al., 2010] as $1/f$ brain dynamics are shown to have optimal $\mu = 2$ scaling. They propose that the brain should be most sensitive to $1/f$ signals as seen in classical music and other phenomena in nature. Our results put this into question, because HTS of brain responses were found to be $1/f$ despite the original stimuli not necessarily being $1/f$ and with the only statistically significant difference between AF functions coming from repetitions of classical music presentation. This result held for all clusters including the ICA components near the auditory cortex. Furthermore, a few limitations of the present experiment should be noted that qualify the conclusions and warrant further investigation. First, although stimulus order did not come out significant ($p = 0.0969$), orders were not balanced across participants. Next, a larger sample size would be helpful

in clarifying the two-way ANOVA statistic between AF slopes and types of stimuli used.

So, is it fair to say that complexity matching effects should not be expected from measuring brain responses using EEG? Perhaps not, as two current studies have adjusted their own methodologies to make new arguments towards the efficacy of complexity matching effects of auditory stimuli and recorded brain responses using their own unique EEG experimental paradigms. The first study collects brain responses corresponding to a combination of both music stimuli presentations and behavioral tasks guiding attention to the music stimuli [Carpentier et al., 2020]. This breaks the tradition of having tasks that focus solely on passive cortical activity while listening to some acoustic stimuli. The researchers argue that it not only allows for them to have more salient brain responses but that it also allows to test for emotional and perceptual features during the presentation of the stimuli. In their study they also deviate from common complexity measures previously used in complexity matching studies and choose to use Multiscale Entropy (MSE); the details of this algorithmic implementation will be discussed in the next section along with other relevant timeseries processing methods. Participants in their study were presented with an array of 40 classical song snippets varying from around 40 seconds to a minute while they were tasked to move a computer mouse into quadrants from either emotional ('Stimulating, Relaxing', 'Pleasant, Unpleasant') or perceptual ('Fast Tempo, Slow Tempo', 'High Pitch, Low Pitch') valances to match how a participant categorizes the stimuli at any given point during the presentation. Furthermore, the processing of brain responses relied on spectral power feature selection like with the ICA components used here but their localization depended on source localization of predefined regions of interest. Results from their proposed experimental paradigm showed evidence for complexity matching effects being salient when participants were attending to perceptual features of pitch and tempo on components located near regions associated to music cognition [Zatorre et al., 2002].

Another study presents strong evidence for complexity matching effects by focusing on adapting feature selection and EEG processing methods while stay-

ing with relevant methodology to complexity matching studies instead of adding behavioral tasks [Teixeira Borges et al., 2019b]. The researchers collected brain responses from 28 participants, which half of them had formal musical training, passively listening to piano performances of 12 classical songs for about 2 minutes per song. The chosen complexity measure is the Detrended Fluctuation Analysis (DFA), which has been used widely in complexity matching studies with behavioral data [Coey et al., 2016, Almurad et al., 2018]. Unlike in the present study, they did not depend on extracting components from the data using ICA, partly because of the higher resolution of their recording system and its 128 recording channels. Instead, data was taken from each channel and split into 7 frequency components using Empirical Mode Decomposition (EMD), which instead of using preset frequency parameters defining the range of each frequency bands, the process becomes a signal-dependent time-variant style of filtering. Simply put, EMD only needs one hyperparameter n and divides the signal into that many frequency components that act as power spectrum modes. This way of defining components from brain responses leveraged the larger number of recording channels and adaptively multiplied the amount of data by relevant feature references. This was useful in showing that specific channels and specific frequency components akin to α, β, and γ ranges had strong correlations between the slope of the DFA functions from the brain responses to the ones from the corresponding stimulus. In line with the approach presented in the current study, they performed DFA on the amplitude of the stimuli as well as pitch and note onsets with pitch DFA functions giving the strongest correlations. Therefore, such results safely show complexity matching effects under their specific experimental parameters.

These two studies stand as the sole empirical examples validating complexity matching effect predictions of complex acoustic stimuli and brain responses, originally proposed by [West et al., 2008]. Each of these studies find unique methodological implementations to achieve these results, but it is argued here that the generalizability of those results faces some limitations worth discussing. This is especially true when going back to the main goal of understanding whether the Hierarchical Temporal Structure can be reflected in brain responses, which through

complexity matching effects can at least be assumed that it happens to some extent but does not fully answer the question. Furthermore, it is argued here that methodological limitations are primarily obscuring complexity matching results in EEG experimental paradigms. The following sections flush out the details of this claim to better understand how data processing and collecting methods influence validation of theoretical predictions as well as their generalizability.

4.6 Limitations of Timeseries Approaches

The understanding of a signal's complexity via the processing of timeseries data has been the precursor for relating the HTS of coupled behavior i.e., measuring complexity matching effects. The complexity matching literature has stuck by several tried and true methods for processing time-series of behavioral and neuronal data. All methods typically include different ways of extracting salient activity while averaging out uninteresting regular activity. As with any feature extraction analyses, there exists trade-offs in what information is filtered out or not. Here I will discuss the details and limitations of Detrended Fluctuation Analysis (DFA), Multiscale Entropy (MSE), and Allan Factor (AF) methods which have been used in behavioral as well as neuronal data which is of focus here.

4.6.1 Detrended Fluctuation Analysis

In relevance to the Cognitive Science literature, the detrended fluctuation analysis (DFA) was first used by [Stephen et al., 2008] to propose a method for capturing long-range correlations of a time series. The purpose of using this method was to make a case for strong anticipation, which involved the activity of a signal at multiple time scales. Simply put, arguments for either weak or strong anticipation were a part of theoretical frameworks in Cognitive Science that contrasted approaches depending on 'prediction' and 'simulation' of cognitive processes happening at multiple levels of analysis such as in what is known as Bayesian and Predictive Processing frameworks, which will be discussed at greater length in Chapter 7. This is also in contrast to looking at correlations that happen at specific

frequency rates and creates a scaling function instead. The use of DFA was first established in a DNA sequencing paper which tried to consolidate a method for understanding the meaningfulness of long-range correlations in nucleotide sequences [Peng et al., 1994]. Such a method was an extension of the classic fluctuation analysis, in which the purpose of DFA was to handle non-stationary activity through a detrending process. Successful mitigation of non-stationary trends in a time series is meant to demonstrate the scale invariant relationship of the signal. The process can be broken down into two steps. First a given timeseries x_i is shifted by its mean (see Equation 2). Then, the integrated series X_t is segmented by windows of different time scales n in which the integrated values are fit linearly per window Y_t and the mean squared residuals $F(n)$ are calculated as the fluctuations of the signal (see Equation 3).

$$(2) \quad \mathrm{X_t} = \sum_{i=1}^{t}(x_i - \langle x \rangle)$$

$$(3) \quad F(n) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(X_t - Y_t)^2}$$

The fluctuations $F(n)$ are a measure of variability at specific resolutions n, that come from the averaged dispersion of residuals extracted from the local linear fits of the integrated values X_t. Power law scaling is demonstrated through log-log plotting of $F(n)$ over n in a fluctuation plot and its trend linear as delineated by the α scaling exponent which approximates the Hurst exponent. This is in reference to the original work demonstrating a heuristic measure of long-range dependence in signals dominated by stochastic properties, termed the Hurst exponent [Hurst, 1951]. Newer measures such as DFA are applied to varying natural data in the attempt to estimate Hurst scaling. The DFA method faced criticism for its ability to create artefactual curvature in the fluctuation plot. Despite its original intent to handle non-stationary noise, the robustness of the method was shown to be limited to signals with either purely stationary or weak-nonlinear trends [Bryce and Sprague, 2012]. Results from the testing of artefactual curva-

ture in DFA put many studies into question. In the case of complexity matching or the related strong anticipation studies, the strength of correlations between α scaling exponents can be problematic for unknown variability coming from artefactual curvature. Newer methods have attempted to go beyond and use a multifractal version of the DFA [Almurad et al., 2018]. Such a method was useful in providing more data of subsystem scaling during behavioral experiments but so far has not been adapted or proposed as useful for EEG data. This could be because it would only give more $1/f$ scaling functions, which would simply multiply the data into subcomponents of the same trend. The original intent for using a multi-fractal adaptation of DFA was to give more resolution at faster timescales [Delignières et al., 2016], but it is hard to tell whether that alone would give any interesting results.

4.6.2 Multiscale Entropy

Measuring the complexity of neuronal activity from EEG recordings has used MSE to process their data, generally for the purposes of having signatures for states of neuronal activity correlated with neurodevelopment syndromes such as Autism spectrum condition [Mandy and Lai, 2016]. This method can be simplified as happening in two steps; first a down sampling of the processed signal via coarse-graining and sample entropy deployed for every coarse-grained scale. The first step is simple as it creates a list of time-series delineated by the number of scale factor where at each scale factor τ the samples are averaged by a non-overlapping window of length t. Then each element of a coarse-grained time series $\gamma_j(\tau)$ is calculated as follows: where the length of each coarse-grained timeseries is N/τ. Also, the

$$(4) \quad \gamma_j^{(\tau)} = {}^{1}\!/_{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i$$

first scale is not necessarily coarse-grained as it is simply the original time-series. The total number of scale factors is often a hyper parameter chosen by researchers that is constrained both by the resolution of interest and overall sampling rate of the signal at hand. [Carpentier et al., 2020] set that hyper parameter to 100 given

that the sampling rate of their EEG data was 512 Hz for recording lengths on average around 1 minute.

Then in the second step, sample entropy is calculated for each coarse-grained timeseries $\gamma_j(^{\tau})$ as follows. Given that N is the length of the timeseries, sample

$$(5) \quad S_E(m, r, N) = ln \frac{\sum_{i=1}^{N-m} n_i' m}{\sum_{i=1}^{N-m} n_i' m+1}$$

entropy calculates the predictability within a timeseries by finding the conditional probability that any two sequences of m consecutive data points are like each other given a similarity criterion r that could remain similar at the next point $m + 1$ [Richman and Moorman, 2000]. The [Carpentier et al., 2020] study set the pattern length parameter to $m = 2$, and the similarity criterion parameter to r=0.5.

4.6.3 Allan Factor and Limitations

In the previous chapters including this one, the details regarding the AF analysis have already been discussed. To avoid redundancies, this section will not repeat a tutorial but instead will jump right into its own limitations along with the other timeseries processing approaches. So far, we know that AF has been used both with onset intervals and peak amplitude events. Unlike the DFA and MSE, the AF method performs its measurement on an extracted event series rather than on extracted timeseries. The main difference being that feature extraction for both MSE and DFA are dependent on averaging processes. In behavioral data, this does not seem to be much of a problem for DFA, because movements are salient activity compared to any recording noise. Regarding EEG data, this could be difficult as it really depends on the quality of pre-processing and specific data cleaning pipelines implemented by the researchers. Specifically, it begs the question of whether specific time window sizes have more or less noise included in the samples it averages. As previously mentioned, there were already several cases in which DFA included artifactual curves in the longer timescales, which could be attributed to the detrending steps not being sufficient for noise at those timescales. Of course, this

does not mean that the AF method is without fault here simply because it does adhere to an event series that reduces free parameters and focuses on only capturing relevant amplitude. In fact, it is because of this capability that it limits the minimum size of signals measured. With music recordings, the rule of thumb has been that the stimuli must be around 4-5 minutes long minimum to have enough samples to process at 11 timescales [Kello et al., 2017]. On top of this limitation, 'relevant' amplitudes are also tricky to assume with EEG data because any mistakes in the preprocessing of the data could leave in artifactual peaks being picked up without any balancing that an averaging process includes.

Across all three methods, the complexity comes from some sort of variability measure whether it would be a Coefficient of Variance such as in AF, residual fluctuations in DFA, or predictability based on differences in neighboring values such as in MSE. I argue here that the limitations of these three methods does not come from their variability measures but rather in their feature extraction steps. The fact that any of these methods can give even limited results is surprising when it comes to EEG data, given the noisy nature of the signal. What does this leave us with? Perhaps with the impetus to model the noise itself. In summary of the limitations, all of the above-mentioned methods open room for mistakes during their feature extraction steps with either overly correcting a trend such as in DFA, having too many free parameters which can add too much emphasis at specific scales of time such as in MSE, and the potential for incorrectly identifying noise as a relevant event such in AF.

4.7 Repeated Measures

Apart from data processing methods, experimental paradigms in data collection also play a big part in not just the quality of the data but also what information the data contains and the extent to which researchers can generalize their findings. One of the peculiar standard procedures of many Cognitive Neuroscience experiments is the practice of measuring repeated responses to the same stimulus, a practice fundamental to event related potential (ERP) measurement [Horváth, 2015].

Such a procedure depends on the averaging of multiple presentations to the same stimulus, but such an approach does not work well with measuring complexity matching effects where the stimulus is several minutes long and complex. Despite this limitation, it is still of interest to understand whether a repetition of a complex acoustic stimuli can evoke a more salient response that could potentially make the signal easier to process during feature extraction. In the present study, this is investigated by making one of the song presentations the same classical symphony song with its order randomly put across participants. In fact, we saw that the only statistically significant result using complexity matching analyses came from that difference in the first presentation to the second presentation of the classical symphony stimulus. What does this mean? Well, this is complicated because a simple correlation would be difficult to interpret because of the similarity of $1/f$ from both the classical music stimulus and spontaneous cortical activity. But what it does demonstrate, is that at the very least a repetition of the same stimulus does change the slope of neuronal scaling at the longest timescales. An easy but perhaps wrong interpretation can be that participants attune better to $1/f$ stimuli and that a change in affect through repeated presentation is evidence given the significant change in slope. That would easily tie back to theoretical predictions from maximal information transfer theory. Of course, what is missing are baselines and controls for this to be the case in the present study, since only the classical song was repeated. Ideally such an interpretation could be given evidence if every song was also repeated and only the classical music was had a significant difference in slopes. But this result does help in interpret results from other studies, if not at least put forth worthwhile testable hypothesis.

If we think back to the results from both studies using the MSE and DFA to measure the complexity of brain responses, one peculiarity arises in the quality of the stimuli used. The first obvious one is that they both present to participants an array of only classical music. Further investigation towards their stimuli used reveals that both studies songs repeating the same composer although the actual songs are different pieces. Furthermore, songs from [Teixeira Borges et al., 2019a] include songs not just from the same composer but they also include consecutive

movements from the same symphony. Movements within the same symphony normally share many musical similarities, such as the same orchestral arrangements, musical motifs, and articulatory phrasing meant to connect with the consecutive movements within one symphony. Although not strictly repeated measures, these similarities are worth pointing out for future studies to test whether actual repeated measures impact the salience of brain responses to the stimuli and whether complexity matching effects can be untangled from any potential entrainment, familiarity, or attention factors that comes with the repeated presentation of stimuli. Lastly, during the steps of extracting components from brain responses we also see a multiplication of the data as there can be more components than the original number of recording channels. In the present study this was not the case as only four of the identified clusters were closely looked at which gave less data than the total number of recorded channels. In the DFA study the was augmented by having [128 Channels $\times$ 7 frequencies] for a total of 896 components while the MSE study used source localization of 68 regions of interest, increasing their data slightly by four. EEG methodology gives an interesting way for thinking about repeated measures because each recording channel on its own can be thought of as an observation of the same event. This is further augmented through methods that multiply the number of observations by components. Ultimately suggesting that repeated measures of this nature boost statistical power of frequentist statistics.

4.8 Localization

Given that the brain is involved in many different processes beyond the processing of acoustic information, it is fair to say that the $1/f$ scaling measured by EEG responses are a summation of subprocesses that include activity irrelevant to the stimuli. To maintain homeostasis the brain has its own default endogenous dynamics, and an approach focusing on signal localization could help in extracting brain responses with a higher correspondence to the stimulus presented. Although making assumptions about the locality of where correspondence to a stimulus can raise difficult arguments about the theoretical framing of what is or is not relevant

activity in cortical dynamics, we can at least treat localization as just another step aiding with noise reduction. We know that brain dynamics are not simply modular, and that correlated activity varies spatially throughout complicated topographies of activity that do not always follow simple parallel heuristics. The ICA decomposition and clustering analyses performed in the current study make a very simplistic effort at trying to define regions of interest without trying to depend on predefined assumptions of locality. One of the clusters conveniently located components in what could roughly be characterized as a region where auditory processing occurs. Despite that, it seemed that the results were more or less the same when compared to components in other brain regions. The [Carpentier et al., 2020] study similarly leveraged localization but instead of using ICA components they extracted source signals of 68 predetermined regions of interest while using a 64-channel recording system. Such an approach was good at producing a standardized multiplicity of components to analyze but it depends on topographic assumptions of salient activity. In contrast, the approach presented in the current study used a more hands off approach by letting the data organize saliency itself based on orthogonal spectral relationships and a data driven approach to defining regions of interest, which through our clustering analysis gave out only 9 (of which we focused on 4) compared to 68. On the other hand, the [Teixeira Borges et al., 2019b] study analyzed data at the electrode channel level without extracting components based on topographic location, but this could be because their system had 128 recording channels. Furthermore, each channel was broken into 7 frequency mode components which can loosely adhere to the same spatio-temporal categorization performed as with ICA and source localization procedures. In all these cases, the extraction of brain responses into some spatio-temporal dimension aided in trying to capture relevant correspondence to the stimulus, thus putting forth a standard of needing to have higher resolution because there exists cortical activity that could potentially reflect more of the stimulus. A simple framing of this problem is like finding a needle in a haystack, where there might be one component that perfectly reflects the HTS of the auditory stimulus. But perhaps it makes more sense that the problem is more difficult, and it is more like trying to find a needle in a needlestack.

All three studies found stronger results in specific locations that could be tied back one way or another to auditory, music, or attention cognition literature. MSE results found that certain regions of interest were correlated with behavioral ratings of emotional and perceptual features of the stimuli that were backed up by music cognition studies. Results were stronger in brain responses adhering to the perceptual behavioral task that participants were tasked during music listening. Here, I put forth criticism worth consideration because the MSEs collected, despite them going through standard muscle and ocular artifact cleaning, are representative not just of the passive brain response to the auditory stimuli but it also includes the motor activity of the behavioral task itself. It is evident that such a task helps provide stronger responses but their argument of such behavioral tasks providing stronger responses because the participant is attending more closely to the stimulus seems to not be clear, given that stronger responses could be an artefact of the motor dynamics produced by the task rather than guiding in stronger perceptual adherence to the stimulus. Complexity matching effects of MSEs were also stronger in parietal and temporal lobe regions whereas in the studies using DFA and AF they both showed stronger results in the frontal lobe region. In fact, the present study was only able to show some evidence of complexity matching effects occurring using a machine learning approach. The AF scores of ICA components were used as predictors for the stimuli category labels in a classification task. Results were above chance performance in models trained only using the longest timescales, while showing that data coming from the frontal lobe had the highest rate of true positives.

4.9 Conclusion

Complexity matching effects between complex acoustic stimuli and brain responses were shown in studies using MSE and DFA as their methods to measure complexity. Experimental and data processing parameters of those approaches raise questions as to how generalizable such results can be. This is evident when considering the that their stimuli were played for shorter durations than in the

present study while only including similar classical music examples. Complexity matching effects in prior behavioral studies are able to show the alignment of HTS between coupled signals across scales of time. When measuring cortical activity, the extent to which one can expect to find this reflection of HTS of one signal on another is limited due to the limits of spatial resolution, as many responses become summations of the default spontaneous $1/f$ scaling that normally exists. This is why results of the three discussed studies find stronger results when looking at specific time scales, or at the averaged variability from the compared scaling functions. It does not make sense for an acoustic stimulus to shape the entirety of cortical scaling and it is difficult to find a highly corresponding signal that could reflect the subtleties of the HTS from the stimulus. The present study does shed some light on extending this methodology to using machine learning, as it provides results above chance performance despite the limits of the experiment at hand. A general machine learning approach would deviate from the specific framing in data collection and processing as it no longer asks the question of whether we can find that needle in the needlestack. It then asks whether the brain response has enough information correlated to the stimulus that it can then tie it to acoustic features. This chapter ends by highlighting the importance of this switch in hypothesis testing from complexity matching to information retrieval, as such methods could free up experimental and data processing parameters to include naturalistic data.

Chapter 5

Image-Based EEG Classification of Brain Responses to Song Recordings

Classifying EEG responses to naturalistic acoustic stimuli is of theoretical and practical importance, but standard approaches are limited by processing individual channels separately on very short sound segments (a few seconds or less). Recent developments have shown classification for music stimuli (∼ 2 mins) by extracting spectral components from EEG and using convolutional neural networks (CNNs). This chapter proposes an efficient method to map raw EEG signals to individual songs listened for end-to-end classification. EEG channels are treated as a dimension of a [Channel × Sample] image tile, and images are classified using CNNs. The experimental results here (88.7%) compete with state-of-the-art methods (85.0%), yet our classification task is more challenging by processing longer stimuli that were similar to each other in perceptual quality, and were unfamiliar to participants. A transfer learning scheme using a pre-trained ResNet-50 is adopted, confirming the effectiveness of transfer learning despite image domains being unrelated from each other. The strength of efficient image-based modeling of EEG responses to music puts forth an example for the strength of information retrieval. This is considered as an alternative method to complexity matching paradigms as it boasts high classification performance of mapping to features of the stimuli, adding

evidence to what type of information is contained in brain responses.

5.1 Introduction

The prior chapters demonstrated evidence for understanding Hierarchical Temporal Structure (HTS) via behavioral coordination and brain response studies using speech and music as the signals of interest in the analysis. Complexity Matching effects allowed for understanding how coupled systems affect the HTS of signals in behavioral experiments but limitations in methodology have found it harder to show the same strength in results with brain responses. Core theoretical predictions in Maximal Information Transfer have been keen on connecting how the structure of systems in the surrounding environment can be tied to neuronal activity during perceptual tasks. Like many other natural systems, the brain also demonstrates a characteristic $1/f$ scaling in its temporal structure. Early computational simulations and experimental measurements discuss how the brain's $1/f$ nature should mean that it is most sensitive to attune or resonate with stimulus also showing $1/f$ dynamics. The $1/f$ nature of the brain in these early perspectives came from coarse measurements capturing a summation of subprocesses that during that time had limited techno-methodological capabilities, unable to have fine grain localization of stimuli correspondence in cortical activity. The advancements in technological capabilities have influenced experimental parameterization and the data processing allowing for more naturalistic acoustic stimuli. Complexity Matching effects were still limited as results came from only from classical music stimuli, while also not being able to dissect how the overall HTS across all scales are being affected. Different types of acoustic stimuli show deviations from $1/f$ HTS and whether that information can be related to a brain response has not yet been verified. The previous chapter presented a case for the strengths in a machine learning approach, as it can pick up on statistical subtleties from the input being mapped to a desired target feature through an iterative learning process. Although a parametric modeling approach via machine/deep learning changes how the question is asked about finding a correspondence from speech and music HTS, the goal

is to see whether strength in performance and generalizability can continue the discourse of this topic further.

Information retrieval is the wide tent term referring to the modeling task of mapping a noisy input signal to a desired feature you are fishing for. Brain responses here are then treated as noisy versions of the corresponding acoustic stimulus, and the modeling task simply tries to retrieve information contained in that brain response. Different sub-disciplines overlap in these approaches to different ends but with relevant experimental paradigms which are reviewed here to develop an efficient methodology. For example, Brain Computer Interface (BCI) research seeks to interpret information retained in brain responses that relates to perceived stimuli for the purpose to extend and coordinate cognition via wearable technologies. Traditional BCI approaches have leveraged correlated behaviors measured through brain responses, such as modeling the relationship between ocular directionality or mouse tracking and cortical activity [Petrushin et al., 2018, Stawicki et al., 2017]. When feeding these data to deep learning networks, classification of specific actions across varying contextual scenes becomes strong e.g., leveraging pupil dilation with cortical activity to classify click actions when navigating the internet [Slanzi et al., 2017].

Consequentially, this leads to interest in researching methods that could interpret the passive responses to complex, naturalistic stimuli which open the possibility of including populations with limited motor capabilities. Users of these BCI technologies will not always be in a position to depend on correlated motor movements so passive cortical activity is of interest. An example of this in the image-stimuli domain has shown the ability to both classify and reconstruct image categories that participants were passively exhibited. [Spampinato et al., 2017] used 10 image classes from ImageNet [Krizhevsky et al., 2012] and randomly presented examples of 0.5 second presentation length. The electroencephalogram (EEG) responses to these stimuli were then fed to a recurrent neural network (RNN). In the case of classification, the trained RNN models were able to take EEG data as input to classify whether a participant was seeing a category of images such as ‘dogs’, ‘cars’, ‘fruits’ and so on.

Methods focused on only processing cortical activity have fallen under the neural decoding umbrella, with a similar goal of relating external behaviors to the internal cortical activity. One of the primary endeavors in neural decoding research is to model temporal correlations of stimuli to their corresponding brain response via regression or classification-based decoders [Glaser et al., 2020]. Common practices with neural spike trains and EEG recordings have included using RNN frameworks to capture temporal transitions over time. EEG recordings show high temporal resolution but a low spatial resolution tradeoff [Livezey and Glaser, 2021]. Here we focus on the comparison of EEG studies related to processing temporal relationships using deep learning methods.

Some approaches use Convolutional Neural Networks (CNNs) along with traditional feature extraction techniques. This usually involves taking a specified EEG channel and passing it a wavelet transform to turn it into a 2D spectral input representation [Golshan et al., 2020, Wang et al., 2019]. Then the 2D spectral input is sent to the CNN for further processing. For instance, [Supratak et al., 2017] utilized the single-channel recordings along with 1D CNN layers as an input for a Long short-term memory (LSTM) model. [Qin et al., 2018] proposed EEGNet, which leverages from the depthwise and separable convolution technique. This allows for the model to use the EEG channel ordering to approximate filter-bank common spatial pattern (FBCSP) and bilinear discriminate component analysis. This is successful because EEG channel ordering follow topographic relationships of how they are recorded on the scalp, thus giving a reference point for how to properly organize EEG channels to create a relevant portrait [Lawhern et al., 2018, Livezey and Glaser, 2021].

Recent studies explored neural decoding concepts and applied them to acoustic EEG responses of complex stimuli. Stober et al. [13] presented participants with 100ms long sinusoid rhythms based on tribal cadences and achieved ∼24.4% classification performance on a total of 24 classes. Stronger performance was shown (∼83.2%) in a 3 class RNN model using spoken vowels (∼0.5 sec) with featured extracted EEG inputs [14]. Another study used longer stimuli (∼10 sec) of 8 varying types of vocalizations and was able to achieve ∼61% performance without any

feature extraction on EEG passed to DenseNet. Yu et al. [15] improved the performance to ∼81% by incorporating canonical correlation analysis between DenseNet and pre-trained VGG model that extracted audio features of the experimental stimuli. Most recently, Sonawane et al. [16] improved on these approaches and showed that longer and complex stimuli (∼ 2 mins of music) could be used to evoke EEG responses used as spectral 2D CNN inputs to classify song ID (∼85.0%).

In this paper, we investigate raw EEG input as a potentially efficient and readily available input representation. We created a raw EEG image representation defined by [Channels, Samples] dimensions. This approach presents state-of-the-art performance using a fraction of trainable parameters with indie/electronic song stimuli (4 mins each) that were unfamiliar to participants. Input representations are analyzed through Multidimensional Scaling (MDS) on channel order and compared with Power Spectral Density (PSD) feature extraction. Lastly, classification results are extended with a supporting dataset using pop Hindi songs as stimuli.

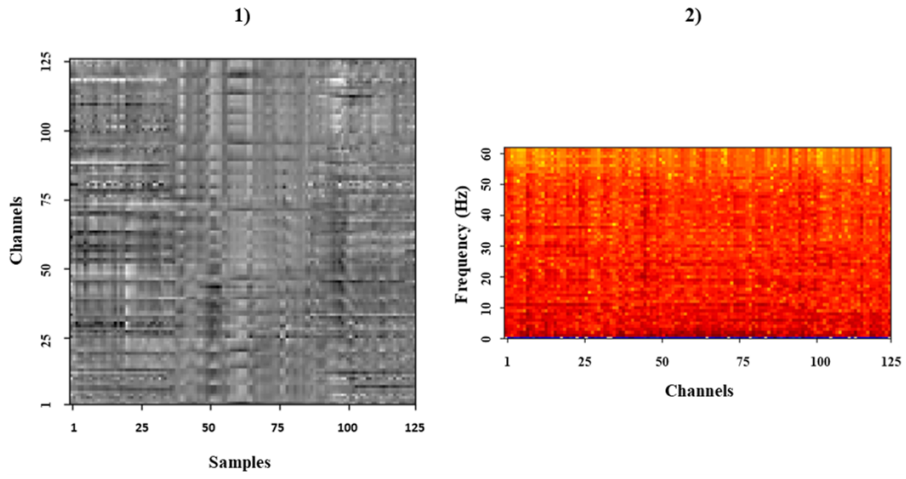

Figure 5.1: 1) the raw and 2) the PSD input representations of participant 1 at the 100th second of song 1. 2) one second PSD at 125 hz produced up to 63 Hz frequency components.

5.2 Methods

The first 4 minutes of all recordings were used and then split into 5-second chunks. To create training and test sets from the same distribution, we balance

the train and test data across time by assigning every other chunk to go either train or test at a 75/25 ratio. From there, the chunks were cut up into 1-second examples and the model presentation order was randomly shuffled. Input tensors were of [Batch Size × Samples× Channels × Depth] dimensions. Given a 1 sec input length the input representations were now square images at depth of 1 giving input tensor dimension values of [Batch Size × 125 × 125 × 1].

Given the training dataset $S = \cup_{i=1}^{n}\{x_i, y_i\}$ drawn i.i.d. from distribution D, we seek to learn a model that generalizes well. In particular, consider a family of models parameterized by $w \in W \subseteq \mathbf{R}^d$. We define the training set loss as follows: where $x_i \in \mathbf{R}^{125\times125\times1}$ which is a depth of 1 portrait of the EEG recordings. Then,

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\ell(w, x_i, y_i) \tag{6}$$

the input is fed to the CNN network and after a softmax layer the objective function to be minimized is as follows:

$$L(w) = -\frac{1}{n}\sum_{1}^{n} y_i \log(G(x)) \tag{7}$$

where $G(.)$ is the neural network followed by a softmax layer. A summary of the proposed architecture is illustrated in Table 5.1.

Since the input is like a grayscale image, we can apply 2D convolutional layers to extract features. The kernel size is fixed to 4 × 4 with a stride of 2. The convolutional layers have 32, 64, 128 filters, respectively. Since the final task is recognition, the network requires the most abstract representation. Therefore, we applied a Global Average Pooling (GAP) [Ebrahimpour et al., 2020] on top of the last convolutional layer. Since most of the parameters in a neural network comes in the fully connected layers (FC), GAP layer significantly reduces the parameters. The activation function on all convolution and fully connected layers (except the output layer) is fixed as linear Rectifier units and the network is initialized with the He [Ebrahimpour et al., 2020] initialization technique.

Layer Type	Filter Size	Input	Output
Conv2D	4×4	$125 \times 125 \times 1$	$63 \times 63 \times 32$
BatchNorm2D	-	-	-
Conv2D	4×4	$63 \times 63 \times 32$	$32 \times 32 \times 64$
BatchNorm2D	-	-	-
Conv2D	4×4	$32 \times 32 \times 64$	$16 \times 16 \times 128$
BatchNorm2D	-	-	-
GAP	16×16	$16 \times 16 \times 128$	$1 \times 1 \times 128$
FC1	-	128	100
BatchNorm1D	-	-	-
FC2	-	100	10

Table 5.1: Details of the proposed architecture. Input: The raw portrait of EEG signal. Output: The class labels associated to music genre.

Finally, transfer learning methods have proven useful for improving the speed of training and asymptotic performance of deep learning models. Since the input is a 2D image of the EEG with spatial structure, we hypothesized that transfer learning is feasible, even when transfer stimuli come from naturalistic images as opposed to EEG images. Therefore, we adopted the ResNet50 pre-trained on ImageNet to our task [Krizhevsky et al., 2012]. Because the input to ResNet50 is a color image, the raw EEG image was stacked three times to simulate what would be RGB channels.

5.3 Experimental Results

5.3.1 Datasets

We evaluated our method on two publicly available datasets: Naturalistic Music EEG Dataset – Tempo (NMED-T) and Naturalistic Music EEG Dataset – Hindi (NMED-H). In both datasets participants are recorded in a passive listening experiment [Dmochowski et al., , Losorelli et al., 2017]. The NMED-T dataset is comprised of 20 participants who listened to 10 songs to their full length (4:30-5:00 mins) in a randomized order. The dataset also included behavioral rating ques-

tions (scaled 1-9) on both how familiar the participant was with each song and how enjoyable they found each song after listening. Familiarity ratings across participants were very low on average. The experimenters also selected the songs to have unique tempos from one another. Uniqueness was defined by having a different Beats Per Minute (BPM) and varying low-frequency spectral peaks. Furthermore, all songs contained vocals with one song not in English.

The NMED-T dataset included both the cleaned and unprocessed signals, for which we opted to use the cleaned signals as to not include muscle artifacts. Specifics of their preprocessing steps can be found in their report [Losorelli et al., 2017]. Independent Component Analysis (ICA) was used to remove muscular artifacts and ocular components were computed to reject bad channels. The recordings were done using the Electrical Geodesics Inc. (EGI) GES300 system with an EGI Net Amps 300 amplifier and Net Station 4.5.7 acquisition software at a 1 kHz sampling rate [Tucker, 1993]. On the other hand, the NMED-H dataset included similar preprocessing except that it used Reliable Components Analysis (RCA) for channel selection. The main difference lies in the participant schema and stimuli, for which there were 12 different participants per condition and each stimulus was a full-length Hindi pop song. There was only a total of four different songs, and participants listened to the songs twice but in our model training, we only used data from the first listen. Lastly, guidance for cross-dataset compatibility has been published to better connect results [Dmochowski et al.,].

5.3.2 Comparisons

Results are generalized to random unheard examples across time and within participants. This follows the training approach of the visually evoked EEG response classification models [Spampinato et al., 2017]. Other EEG classification studies also do not do across participant generalization because of the high cost of acquiring a large number of new participants when one participant can give you sufficient recordings [Moinnereau et al., 2018, Sonawane et al., 2021].

The NMED-T dataset has a total of 10 class labels, and the NMED-H dataset

has 4 classes. We also added yet another classification task for enjoyment ratings. The NMED-T dataset has enjoyment ratings scaled from 1-9. The enjoyment ratings were applied as targets for the classification of rating. However, enjoyment ratings were unequally distributed. Thus, we decided to group ratings into three broad groups: low, medium, and high enjoyment classes for a 3-class classification model. Standard sentiment classification studies also only use either binary or 3-class models [Bird et al., 2019], and the label reassignment here seeks to make the task relevant while also compromising with the dataset size limitation. Furthermore, several validation metrics are included to more strictly interpret the effects of reassigning labels, as well as to better understand the strengths and weaknesses of classification.

Models	**Accuracy%**	**Precision%**	**F1%**	**Kappa%**
NMED-T	88.69	88.85	88.67	87.43
NMED-H	97.09	97.36	97.05	96.13
ResNet-50	93.05	93.08	93.05	92.28
Enjoyment	90.12	90.15	90.10	83.87
Random	80.23	80.68	80.22	78.04
MDS	83.13	83.60	83.08	81.26
PSD	83.18	84.05	83.20	81.31
PSD-2	80.75	81.81	80.60	78.61
Res-PSD	94.12	94.13	94.12	93.46

Table 5.2: Grand performance summary of all our models. Top panel shows results for our models trained on raw input and classification of the song or its enjoyment. Bottom panel shows results for models trained on the different feature sets tested.

Table 5.2 presents the testing results of the different classification models across several performance metrics. Cohen's Kappa illustrates the strictest metric for all models. In this case, Cohen's Kappa metric is measuring the agreeability between the target labels of the songs (the name of the songs) and the predicted labels from the classification. The usefulness of this metric is that it considers as a ratio both the difference in true and predicted target labels over the random chance of agreeability [Cohen, 1960]. This is especially useful for the enjoyment rating

model. It seems to pick up on the 3-tier class reassignment used here because it has the biggest difference (7%) in Cohen's Kappa to other metrics, compared to only 2% for the other models. Top panel of Table 5.2 shows that all the models here outperform any prior EEG classification attempts.

Recent attempts at classifying auditory EEG responses have achieved similar performance (84%) all while relying on feature extraction or simple and short acoustic stimuli [Moinnereau et al., 2018, Sonawane et al., 2021]. Study [Sonawane et al., 2021] provides a strong example for treating EEG responses as images in classification tasks. They show that when the EEG recordings are treated as 1D time-series in CNN layers, performance stays at chance. This is something we were able to verify with our initial attempts at classification using the NMED-T dataset. Their strongest model (84.96%) is a consequence of extracting frequency components with a PSD analysis. In our simple 2D CNN model we show that we can achieve 88.69% accuracy without normalizing the data or using spectral analysis. Table 5.3 illustrates a comparative summary across relevant studies considering their best performance and details of the datasets used to achieve that performance. The most comparable study is by [Sonawane et al., 2021] and we focus here on how our approach expands on it. Table 5.3 shows that the main model here is trained on EEG responses with the longest stimuli by a large margin on the order of minutes. Other studies have also attempted to not include feature extraction steps but the stimuli length in their experiments are significantly smaller than in this study (in the order of seconds) [Stober et al., 2014, Yu et al., 2018]. In comparison to [Sonawane et al., 2021], the EEG responses in NMED-T were to unfamiliar stimuli, which in a prior study it has been shown to be the harder case as classification performance drops when listeners are not familiar to the music stimuli [Hadjidimitriou and Hadjileontiadis, 2013a]. Our model also achieves this level of performance (88.69%), but with substantially smaller parameters (179,132 compared to 1,678,156) in [Sonawane et al., 2021].

In any standard image classification task, it matters how the input images are represented. The use of data augmentation such as rotations, mirroring transforms, and noise can help regularize model. On the other hand, we also know that if we

are classifying categories such as 'dogs', we want the face, legs, and tail to be in the correct spatial order. The concern with our input representation is whether the default channel order did not distort our proposed cortical portrait. Bottom panel of Table 5.2 shows a summary of testing the input representation format.

Studies	Accuracy(%)	Class Size	Stimuli Length	Feature Extraction	Stimuli Type
Sonawane et al (2021)	84.96	12	2 mins	Yes	Music
Moinnereau et al (2018)	83.20	3	0.5 secs	Yes	Spoken Vowels
Yu et al (2018)	61.00	8	10 secs	No	Vocals
Stober et al (2014)	24.40	24	100 ms	No	Sinusoid Rhythms
Our Study	**88.44**	10	**4 mins**	No	Music

Table 5.3: Summary of studies that try to classify EEG responses to an ID label of complex auditory stimuli.

5.3.3 Input Representations

Channel order was tested by training models with randomly shuffled channels and with channel groupings based on MDS as seen in Table 5.2 with models 'Random' and 'MDS'. In short, the MDS analysis consisted of taking the root mean square (RMS) of the channel amplitude envelopes in the training data to create a difference matrix (pairwise Euclidean). The matrix was then projected into a 1-dimensional embedding through MDS manifold learning, which allowed for rank ordering channels based on their dissimilarity. Table 5.2 shows that MDS ordered channel training had 4% less accuracy, while random ordered channels had a significant loss in performance 13%. With this we can have confidence that our proposed raw EEG portrait with the default channel ordering is adhering to crucial spatial relationships for classification. The concept behind testing channel order is inspired by [Saeed et al., 2021] Channel Reordering Module (CHARM), which helps make sense of heterogeneity of electrode channels. CHARM is a generalized model for taking in channel data and extracting components then reprojected into a meaningful order for a new 2D representation of the data.

Furthermore, we also evaluated the impact of raw input vs. spectral representation on the NMED-T dataset (visualized in Figure 5.1). The periodogram function was passed to all our 1 second raw EEG inputs at 125 Hz with Fast-Fourier Trans-

form (FFT) windows being the same size as the input length as well as passing the function to 2 seconds EEG inputs [Bartlett, 1950]. This was done because PSD approximation gives back frequency components up to half the maximum sampling rate of your input, and the 2 second EEG examples give us a representation with the same shape $\mathbf{R}^{125\times125}$ as our raw representation. This gives us confidence that differences in performance are between input representations and not model hyperparameters. In Table 5.2 we can see that PSD representation from 1 second EEG has better performance than the 2 second examples as seen in models 'PSD' and 'PSD-2' respectively. The performance here is still 5% less than our main model trained on raw inputs, showing us that feature extraction is not needed for classification tasks.

5.3.4 Transfer Learning

Fine-tuning of the ResNet50 is done with raw and PSD representations where a pretrained ResNet50 model on images commonly found on the internet is used as a foundation to then train with the EEG data on top. Both models in Table 5.2, 'ResNet-50' and 'ResPSD', achieve test performance up to 93%. The high performance here is surprising because the input representations used here are significantly different than the original training images, as seen in Figure 5.1. These results support our main point that EEG data can be effectively processed through standard computer vision methods to allow for stronger performance and more efficient models. We can see that with EEG we have two variables that configure a cortical portrait; channel topography and spectral extraction. More testing is needed to further understand what configurations are more appropriate across deep learning tasks, but with our results we show an example of high performing end-to-end classification.

5.3.5 Validation & Generalization

Many publicly available EEG datasets have a participant size limitation because of how expensive it is to have long recordings of various people. This primarily

becomes an issue when trying to balance inputs to targets as well as how to assign train and test data. We did label reassignment on the sentiment classification task to handle the prior mentioned issue, and the data was separated into time chunks for the latter to avoid the train or test sets having too much of the beginning or end of the songs. This is also why our generalization was done using a hold out metric instead of cross-validation. In an attempt to be critical of our results, we also included precision, F1, and Cohen's Kappa as model metrics. We find that precision is higher than accuracy for all models, although not by much. We also see that F1 scores are not much lower than model accuracy, but that kappa does reflect a strict interpretation of our results. Further validation testing of the main 'NMED-T' model is seen from Table 5.2 through the use permutation tests. The first test looked at predicted performance when randomly permuting the test labels and the second test randomly permuted the model weights; in both tests, performance stayed at chance 10%.

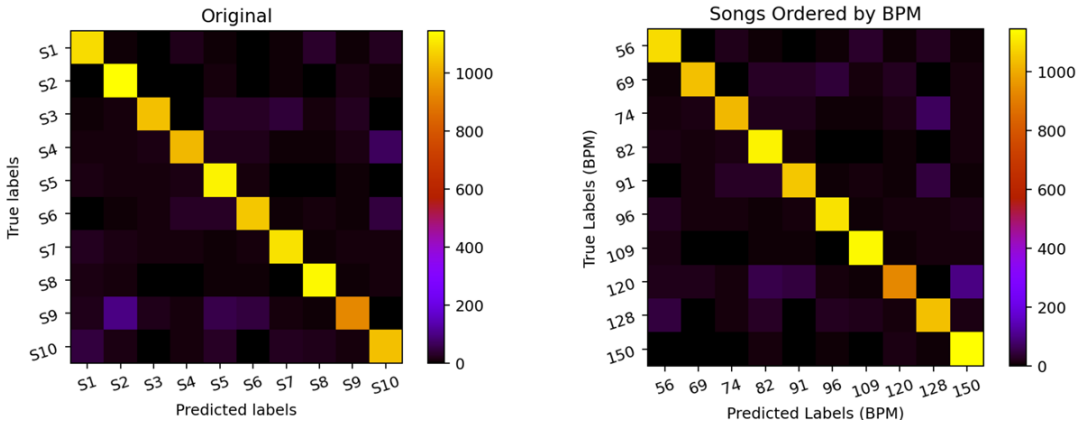

Figure 5.2: Left, the confusion matrix for the original model. Right, results from the original model sorted by ascending BPM.

Our results are interpreted here as providing a useful method for end-to-end classification of EEG responses to music listening, but alternatively we could also be seeing these results only because the dataset was designed to be distinguishing tempos. In other words, there is a competing alternative that the models here are not capturing overall acoustic features and only works because of the differences in BPM. [Stober et al., 2014] provides a good counterpoint to this because their

results explicitly test acoustic stimuli with the exact same BPM, and test performance across their models are fairly above chance. We perform an additional analysis to address this as well, by looking at the correlation of BPM across confusions as seen in Figure 5.2. Specifically, we took the confusion matrix of our best performing model and organized it by ascending BPM. It can be seen that confusion was random and not clustered along the diagonal. Here, the average BPM difference for confusions was 31.5 BPM, whereas the chance difference was 28.95 BPM. Although the stimuli were originally picked because of their differences in tempo, some only differ with 5 BPM, and genres fall under the similar 'indie-electronic' umbrella due to the researchers' focus on unfamiliar stimuli [Losorelli et al., 2017].

5.4 Conclusion

Traditionally, EEG data has been likened to time-series/sequence input representations in deep learning due to the analysis conventions in Cognitive Neuroscience studies [Glaser et al., 2020]. Recent relevant studies have leveraged CNNs by extracting frequency components of EEG [Moinnereau et al., 2018, Sonawane et al., 2021]. In this study, novel recognition results are presented which support the proposed method that EEG responses to full length music stimuli can achieve state-of the-art performance using the raw input without any feature extraction. In total, 9 models and 2 datasets are used to support our method, while the main dataset used (NMED-T) is the most difficult benchmark of EEG music classification. This is because the music stimuli are the longest by an order of minutes, the participants were unfamiliar with the stimuli, and the songs overlapped significantly in terms of genre. Despite these challenges, our experiments reveal that EEG responses to music can be processed end-to-end. The strength of this approach contrasts clearly with the previous Complexity Matching experimental paradigms as it does not need careful feature extraction, localization, and allows for more challenging stimuli that are more relevant to the act of music listening. The information retrieved via this method includes the identity of the music stimuli used given only 1 second of EEG data as well as the

classifying how much someone enjoyed it. These two features of the music stimuli may be limited to the discrimination of types of data, but the methodology builds a foundation for extending this to retrieve information closer to what is the HTS of the music. The next chapter will present results to that end via a deep regression modeling task.

Chapter 6

EEG2MEL: Reconstructing Sound From Brain Responses to Music

Information retrieval from brain responses to auditory and visual stimuli has shown success through classification of song names and image classes presented to participants while recording EEG signals. Information retrieval in the form of reconstructing auditory stimuli has also shown some success, but here we improve on previous methods by reconstructing music stimuli well enough to be perceived and identified independently. Furthermore, deep learning models were trained on time-aligned music stimuli spectrum for each corresponding one-second window of EEG recording, which greatly reduce feature extraction steps needed when compared to prior studies. The NMED-Tempo and NMED-Hindi datasets of participants passively listening to full length songs were used to train and validate Convolutional Neural Network (CNN) regressors. The efficacy of raw waveform versus power spectrum inputs and linear versus mel-spectrogram outputs were tested, and all inputs and outputs were converted into 2D images. Quality of reconstructed spectrograms was assessed by training classifiers which showed 81% accuracy for mel-spectrograms and 72% for linear spectrograms (10% chance accuracy). Reconstructions of auditory music stimuli were discriminated by listeners at an 85% success rate (50% chance) in a two-alternative match-to-sample task. The strength

of this methodology adds a new type of evidence for how much of the structural information of an acoustic stimuli is retained in brain responses given that a perceptually recognized reconstruction can be heard.

6.1 Introduction

Reconstructing stimuli from brain responses has been explored across several related subdisciplines spanning from signal processing methodology, Cognitive Neuroscience, and deep learning modeling. These methodologies stray away from the Complexity Matching approach to understanding Hierarchical Temporal Structure (HTS) as they often rely on parametric solutions to retrieving information of acoustic stimuli from brain responses. The success of these approaches is that they output results more easily interpretable, as both researchers and others can actually judge for themselves the quality of a reconstruction by listening to it. Therefore, if a reconstruction is good enough to sound like the original stimuli, then it makes it easier to propose that brain responses contain enough information, including HTS, for the original signal to be recovered. A notable example comes from the culmination of experimental studies showing how to retrieve a Frequency Following Response (FFR) from averaged brainstem recordings to short (<1 second) presentations of speech and music [Coffey et al., 2019, Skoe and Kraus, 2010]. This type of recording only needs three electrodes, one centered on the scalp, a reference on the earlobe, and one grounding on the forehead. With this set up, participants are presented a sound such as a violin playing with their eyes closed repeatedly for about 1000 - 3000 times. The strength in this approach comes from the quality of FFRs to be able to contain the fundamental frequency, harmonics, and onset aligned amplitude from the auditory stimulus. This allows for researchers to playback an FFR and hear a lower resolution version of the original stimulus. Such a methodology provided an impressive proof-of-concept for speech and music retrieval, but it also depends on various experimental conditions. First, careful stimuli selection is necessary to retrieve FFRs as it needs to consider stimuli that have clear amplitude bursts, strong onsets, lower fundamental frequencies (<300 Hz),

and short durations. Furthermore, the retrieved signal is averaged from hundreds of repeated presentations used to isolate a specific response such as in brainstem activity.

Another exemplary approach in the space of signal processing and Cognitive Neuroscience comes from auditory tracking experiments where researchers have been successful in retrieving the speech amplitude from the stimuli [Synigal et al., 2020]. Experiments had participants listening to 180 secs of an audiobook split up into 25 separate trials. What makes this approach strong is the simplicity behind the modeling procedure where it takes as input both low frequency and high-gamma power spectrum signals from Electroencephalogram (EEG) channels and puts them through a linear model that decodes the input and then maps to the stimulus target. Despite the modeling simplicity, the approach faces a trade off with the data processing needing to be more complicated. EEG signals need to first be split into two signals that separate activity from low and high frequencies while also being iteratively integrated across channels into the model over several time lags from 0-250 ms. Nevertheless, this research moves forward the methodology of neural decoding and acoustic information retrieval from cortical activity towards a more naturalistic scenario where it does not require mass averaging of presentations and stimuli presentations are longer.

Deep learning methods of decoding stimuli from brain responses have been successfully implemented through various classification models. In the image stimuli domain, researchers collected passive cortical responses of people watching images from ImageNet [Spampinato et al., 2017]. They trained the EEG responses in a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN). Model validation demonstrated that brain responses could be used to correctly identify an image class at 84% performance that is being perceived (2.5% chance), i.e., if a participant was watching an image of a panda the model could correctly identify that the image being perceived was of a panda. In the auditory domain, recent classification studies have developed methods to efficiently use EEG responses in Convolutional Neural Networks (CNNs) to correctly classify the name of the song a participant is listening to. This was shown to work strongly (85%) when

participants listened to music they were familiar with (8.33% chance), and if the input of the models were formatted as 2D representations of the power spectrum density across channels [Sonawane et al., 2021]. Using different datasets, another study replicated the prior study's results and further evaluated the efficacy of EEG data being used as images in computer vision models implemented through a custom version of AlexNet architecture and in transfer learning with a pretrained ResNet model. The researchers trained models with the power spectrum density EEG representations and compared it to the raw EEG representations. Training on the raw input representation was able to show that not only was end-to-end classification possible, but that it could also achieve State-of-the-Art performance at 88.6% with 10% chance on unfamiliar music [Ramirez-Aristizabal et al., 2022].

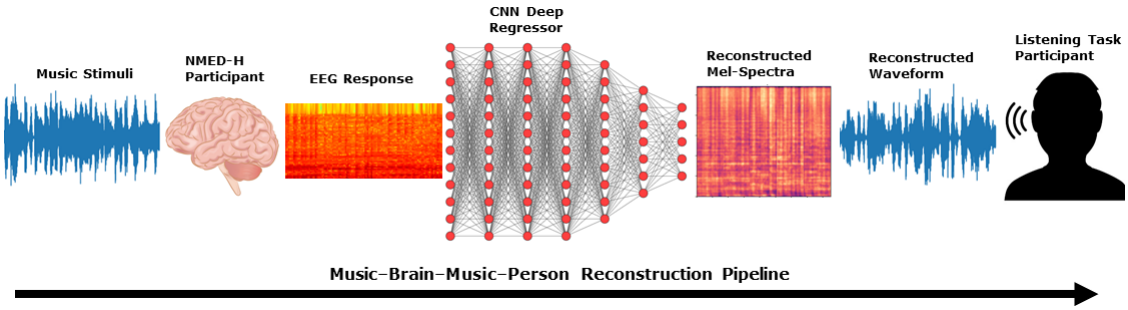

Figure 6.1: Visualization of music reconstruction in our study. Brain responses from music listening are processed by deep regressors and retrieved music is played back to new participants.

Studies moving beyond classification approaches and into trying to reconstruct the stimuli have seen some initial success with the use of generative models and careful feature extraction. Following the work in [Kavasidis et al., 2017], the researchers transferred weights from their classification models to use as encoders in a Variational Autoencoder (VAE) and Generative Adversarial Network (GAN). They successfully recreated the image classes that participants were looking at with an EEG response as input to the network. For example, if a participant was looking at images of pandas, then the models would recognize that it is a panda and output an image of a panda. The novelty of generative models comes from the ability to not just have outputs that are semantically similar, as their study claims, but to also generate a distribution of images not directly presented to par-

ticipants. This was made possible by the transfer learning of weights trained to classify image classes, which when fine-tuned to new examples created training procedures of generative models estimating the density of image classes. For the purposes of direct information retrieval this also becomes a limitation because the outputs are estimations from image classes rather than a pixel-by-pixel reconstruction. This means that if a participant was looking at a picture of a panda who was sitting down and profiled to the left, then the model would not necessarily pick up on those features but return a sampled panda image from the estimated image class fitted by the model. In music retrieval, another study had found a way to leverage the acoustic features of the music stimuli and use it as inputs for a multi-view approach in a deep Variational Canonical Correlation Analysis (VCCA) model [Ofner and Stober, 2018]. Despite the lack of quantitative model validation analyses, this study demonstrated the possibility of using deep learning to be able to reconstruct music spectra from long EEG responses. In their qualitative analyses, they confirmed that reconstructed spectra were able to retain acoustic features from the original stimuli such as pitch, timbre, and tempo. Furthermore, it allowed for an approach that would move beyond stimuli classes and attempt to consider time-aligned activity in the stimulus presentation to be retrieved from brain responses. Here, we present results that advance the methodology of music retrieval and stimuli reconstruction from brain responses by training models that have time aligned EEG responses to the music spectra target. This is done without needing to integrate acoustic features from the stimuli as input to the model or depending on a multi-step feature extraction process from the EEG. We also present a series of quantitative validation methods to measure the success of music reconstruction, including feedback from participants listening to the reconstructed music from our model outputs. Figure 6.1 outlines the scope of information retrieval, starting with music stimuli presented to participants being encoded in their EEG responses and mapped to the time-aligned stimuli spectrogram by a deep regressor which outputs a reconstructed music spectrogram that is inverted into a waveform and presented to a new person.

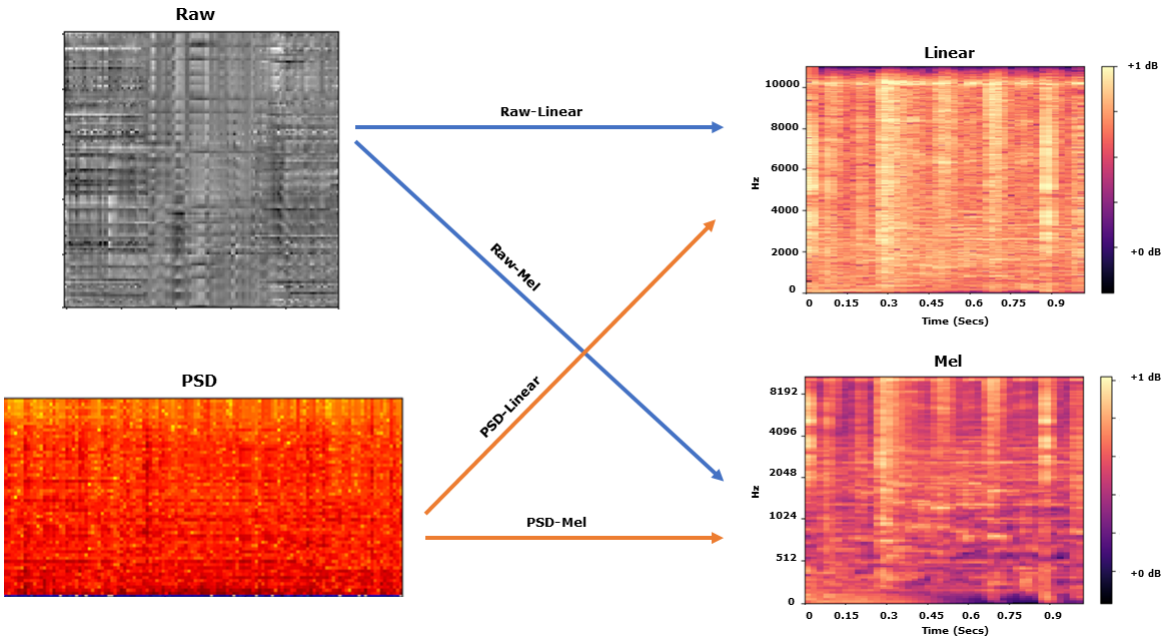

Figure 6.2: On the left are the two types of input representations we test, and on the right the two types of target representations for a total of 4 model combinations as labeled by each arrow. All representations come from Participant 1 at the 100th second.

6.2 Methods

6.2.1 Datasets

Here we train and validate models with the Naturalistic Music Electroencephalogram Dataset – Tempo (NMED-T) and Naturalistic Music Electroencephalogram Dataset – Hindi (NMED-H) [Dmochowski et al., , Losorelli et al., 2017]. Both datasets collect long recordings of passive brain responses to participants listening to music. The NMED-T contains recordings from twenty participants listening to ten different songs that were selected by the researchers because of the lack of familiarity and differences in tempo. On the other hand, the NMED-H has recordings from twelve participants listening to four pop songs in Hindi. These publicly available datasets have been central to various signal processing studies along with some studies looking into music retrieval from cortical activity [Ofner and Stober, 2018, Vinay et al., 2021]. To aid with replication we provide

guiding python notebooks as examples to our methods and analyses for others to follow. The supplementary materials also provide examples for others to listen to and visually inspect. Both datasets provide preprocessed versions of the data which include standard corrections for faulty channels, line-noise filtering, and muscle artefact corrections [Dmochowski et al., , Losorelli et al., 2017]. Here we use those preprocessed versions of the data, as our focus is on passive cortical signals188and not on any correlated motor behaviors that may show up as muscle artefacts.

Layer Type	Filter Size	Input	Output
Conv2D	4×4	$63 \times 125 \times 1$	$63 \times 125 \times 8$
BatchNorm2D	-	-	-
Conv2D	4×4	$63 \times 125 \times 8$	$63 \times 125 \times 16$
BatchNorm2D	-	-	-
Conv2D	4×4	$63 \times 125 \times 16$	$63 \times 125 \times 32$
BatchNorm2D	-	-	-
Conv2D	4×4	$63 \times 125 \times 32$	$63 \times 125 \times 64$
BatchNorm2D	-	-	-
Conv2D	4×4	$63 \times 125 \times 64$	$32 \times 63 \times 128$
BatchNorm2D	-	-	-
MaxPool2D	2×2	$32 \times 63 \times 128$	$17 \times 32 \times 128$
Flatten	-	$17 \times 32 \times 128$	69632
FC1	-	69632	128
BatchNorm1D	-	-	-
FC2	-	128	5632
Reshape	-	5632	44×128

Table 6.1: Architecture used in our deep regressors. This specific model was trained on the NMED-H dataset with a spectral input and mel-spectra music target.

6.2.2 Model Training

Our modeling approach contrasts with previous publications showing success with decoding and reconstructing complex stimuli, where generative models such as VAEs and GANs reconstructed image classes and where VCCA was able to reconstruct music stimuli spectra at the expense of needing a multi-view of extracted features [Ofner and Stober, 2018, Kavasidis et al., 2017]. We choose a straightforward approach with a sequential CNN based regressor mapping EEG input directly to the time aligned music spectra. Table 6.1 shows a summary of how the model architecture is constructed; the models contain five convolutional layers with the last convolutional layer being the first layer that reduces the dimensionality of the input. A Max Pooling layer with a small pool size is chosen over a Global Average Pooling layer used previously during EEG classification [Ramirez-Aristizabal et al., 2022] using the same dataset, because it was necessary to limit shrinkage since the output layer was the size of the spectral target. Regularization in the model was implemented through dropout layers and maintaining the size of intermediate fully connected layer as small as possible. Dropout layers between convolutional layers were kept with 10% dropout and the dropout layer between the fully connected layers was kept at 15%. The amount and strength of dropout layers were tested demonstrating no evident effect with less layers and weaker layers, stronger values were not helpful between convolutional layers but did show some improvement between fully connected layers. Activation and Kernel L2 regularization was tried but opted out due to showing signs of a vanishing gradient. Number of filters for convolutional layers were kept at a base 8 while increasing by a power of 2 for every subsequent layer. Increasing the number of filters showed training loss outpacing validation loss resulting in overfitted model runs while decreasing the number of filters made training loss stagnate too early which resulted in underfitting. For all intermediate layers a Rectified Linear-Unit activation was used with a linear output activation. Non-monotonic activation functions Swish and Mish were tried due to their success in improving image processing in deep networks, but they did not present any evident advantage over ReLu during training. Lastly, we use Adaptive Moment Estima-

Inputs (1 secs)	**Targets (1 secs)**	**Accuracy**
PSD (63,125)	Mel-Spec (44,128)	80.80%
PSD (63,125)	Lin-Spec (44,1025)	72.28%
Raw (125,125)	Mel-Spec (44,128)	46.07%
Raw (125,125)	Lin-Spec (44,1025)	37.53%

Table 6.2: Summary of reconstruction models' output classification across the four representation combinations. Each representation is shown with its data shape for 1 sec in parentheses.

tion (Adam) as an optimizer with a 0.0015 learning rate and initialized weights with a He uniform distribution which have shown advantages in similar training procedures [Ramirez-Aristizabal et al., 2022, Ebrahimpour et al., 2020].

The first four minutes of all recordings were used and cut up into five second chunks. To balance the train and test set distributions across time, we assign every other chunk to either train or test at a 75/25 ratio. Then all chunks were split into 1 second examples and randomly shuffled for training and validation. Five second chunks were also useful in securing consecutive 1 second examples to be reconstructed. These reconstructed five second music spectra were inverted into waveforms and used as examples in a behavioral experiment to validate the quality of brain to music reconstruction. Generalization stays within participants and across unseen chunks of time that are balanced to be sampled from the beginning, middle, and end of the songs. Other EEG studies also keep generalization within participants [Moinnereau et al., 2018, Sonawane et al., 2021] and a recent study has shown that weak correlations across participants in the NMED-T could be why generalizing to unseen participants is difficult without many recorded participants in the dataset [Pandey et al., 2022].

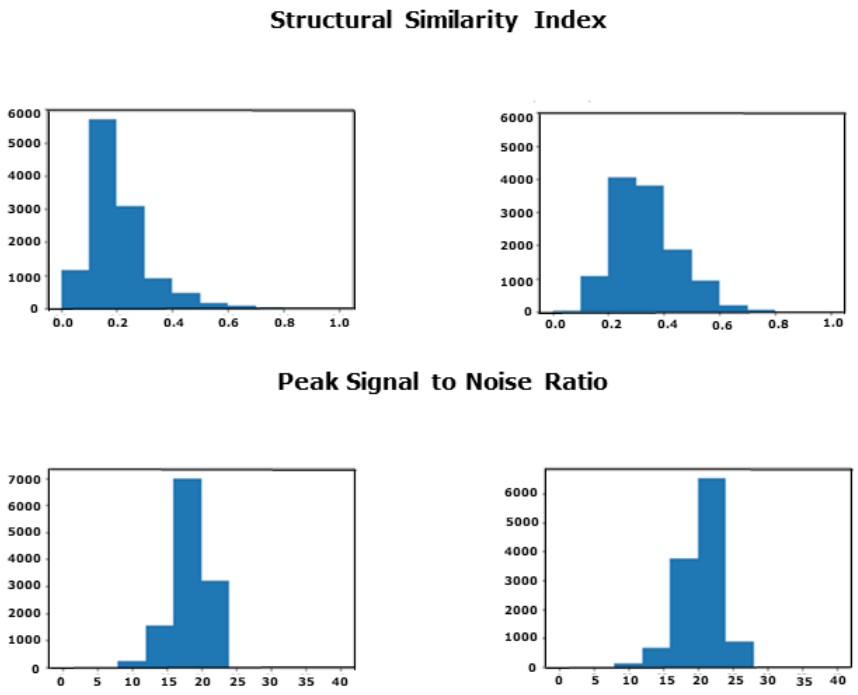

Figure 6.3: Distributions of SSI and PSNR scores across target representations, along with their mean score.

6.3 Experimental Results

6.3.1 Representations

Using the NMED-T, four reconstruction models were trained and evaluated depending on their varying input and target representations. Figure 6.2 outlines the combinations between input and target representations along with their visual qualities. Each arrow in the figure is a model showing the designated input to target mapping. Prior classification studies have shown a preference for input representations either being raw or a Power Spectral Density (PSD), and their ability to boost performance in CNNs [Sonawane et al., 2021, Ramirez-Aristizabal et al., 2022]. As a regression task, we also find it important to compare target representations since linear and mel-spectrogram representations have tradeoffs. The linear spectrum scales low frequency activity equally to higher frequency, which allows it to be more easily inverted back into a listenable waveform. On the other hand, the mel-spectrogram applies a non-linear scaling across frequency ranges that has made it

easier to map in classification tasks [Gururani et al., 2018, Chillara et al., 2019], but also adds more noise during the inversion to a listenable waveform. The reconstruction models were all trained until they reached a plateau in their loss with varying epoch ranges (100-300) needed for the models to converge their training and validation performance. Given that the mean-squared error (MSE) loss was only indicative of the convergence of the model during training and not a metric that informs how image reconstruction compares across models, we decided to classify the output as an objective quantitative measure. Simply put, the four trained regressors were used to create new datasets from their outputs given the original NMED-T as input and keeping the same train-test split. Then those new output datasets were passed through a classifier that would attempt to classify the name of the song from each spectral image reconstruction. This type of validation has been used before in the image stimuli domain when trying to objectively test the quality of reconstructed image classes from generative models [Kavasidis et al., 2017]. The classifiers were CNN based following the architecture from a previous study where the EEG was classified into song name classes [Ramirez-Aristizabal et al., 2022]. Table 6.2 demonstrates a summary of the results from the classified outputs across modeling approaches. The raw input representation showed to not give the best results in our deep CNN regressors, but it did remain well above the chance performance rate of 10%. Meanwhile, the mel-spectrogram worked better across both the raw and PSD input representations with performance comparable to studies where only EEG classification was conducted [Moinnereau et al., 2018, Stober et al., 2014, Yu et al., 2018].

Classifying outputs to their stimuli classes reveals fidelity of semantically relevant reconstructions. This was especially useful in the image stimuli domain when the models being evaluated focused at the stimuli class level [Kavasidis et al., 2017]. Furthermore, their results of classification stayed well above chance (2.5% chance) for forty classes with performance 40%. The results from Table 6.2 also stand well above chance for 10 classes (10% chance). Given that our training method maps EEG to the target stimuli directly rather than estimating stimuli class densities in the model's latent space, common image reconstruction metrics were used to see

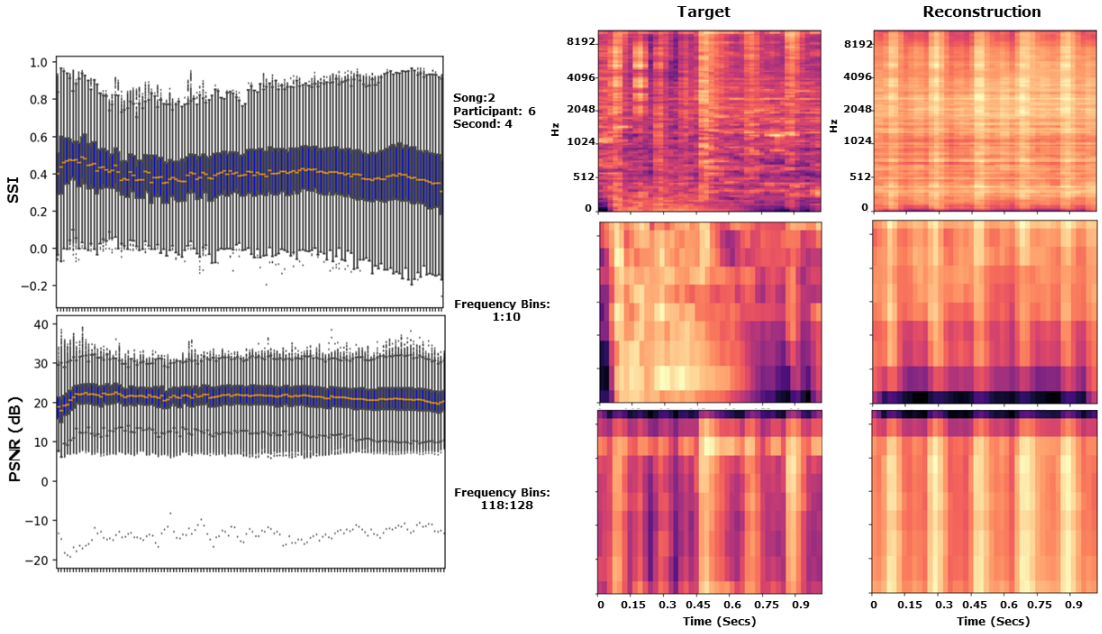

Figure 6.4: Distributions of SSI and PSNR scores across target representations, along with their mean score.

how different each reconstruction was from the target image. Structural Similarity Index (SSI) and Peak Signal to Noise Ratio (PSNR) were chosen because of the ability to compare the reconstruction of perceived image features and how much noise the reconstructions contain from a lossy compression comparison respectively [Hore and Ziou, 2010]. Figure 6.3 summarizes the results from both metrics across the test set of the image reconstructions from the regressor models trained on the PSD input representation along with their linear and mel-spectrogram targets. SSI measures, as a normalized ratio, the similarity between images where an SSI score of 1 points to the images being identical. The mean SSI is higher in the mel-spectrogram reconstruction than with the linear-spectrogram by 14%. PSNR on the other hand, measures as a ratio the logged maximum image power over its mean squared error which gives a relative difference between images and not a normalized score. We also see in Figure 6.3 that mel-spectrogram has a higher mean PSNR which further validates how the mel-spectrogram makes the EEG to music spectra reconstruction closer to the target.

To better understand what musical features the model learned from mapping

to the mel-spectrogram, we took SSI and PSNR scores across frequencies. Figure 6.4 demonstrates the results of this analysis; for every example we paired up each 1 second frequency bin from target to model output to calculate both an SSI and PSNR value. Boxplots were created for all 128 frequency bins, and Figure 6.4 shows that frequencies vary around a 0.4 SSI score. Meanwhile, the PSNR plot shows notably lower values for the first 5 frequency bins. We show a visual example of this difference by taking a reconstruction and zooming into the ten lowest and highest frequency bins. This visual comparison demonstrates how many more pixel differences the lower frequencies have when compared to the higher frequencies. This went against our expectations as we believed it would be easier to map to lower frequencies because of pattern regularity. Further testing is needed as future studies adopt this methodology, but we speculate that this could be attributed to these model architectures not explicitly learning temporal dependencies such as in an RNN.

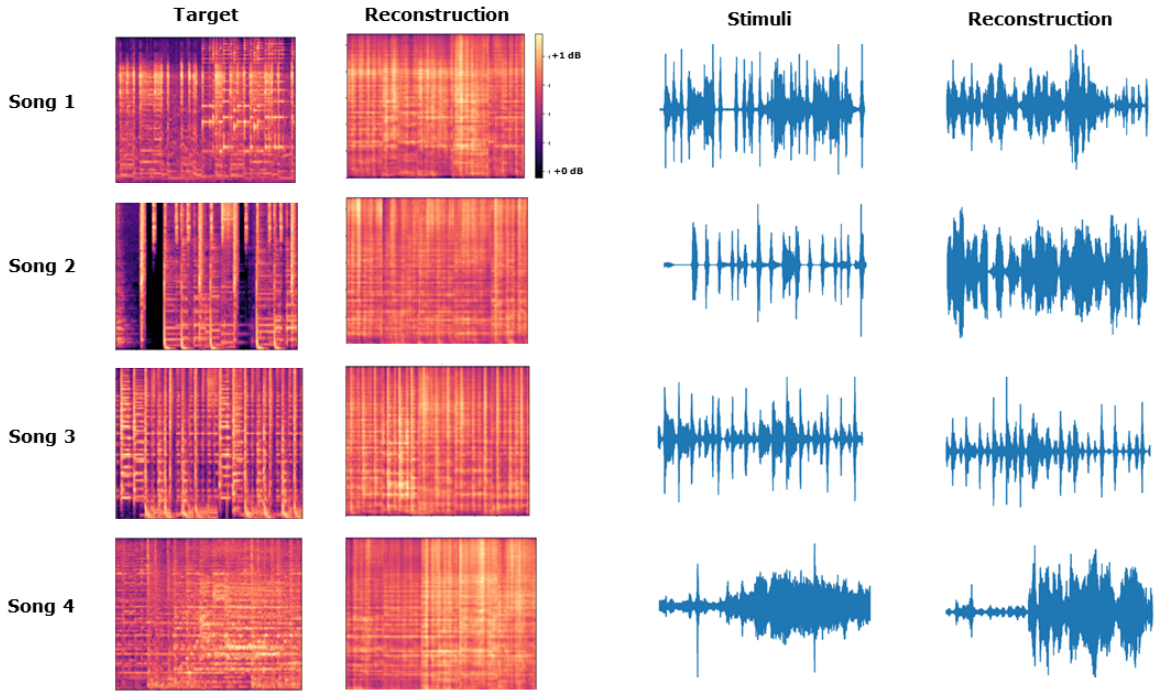

Figure 6.5: Five second examples of model spectra predictions (left) and their reconstructions from spectra to sound wave (right). Examples come from the 10th second across randomly selected participants.

6.3.2 Spectra to Music

The above-mentioned metrics were useful in evaluating the success of the spectral image reconstructions, but here we are also interested in whether that allows for the retrieval of perceptually interpretable music reconstruction. This means that our spectral image reconstructions should be good enough to then be processed by common out of the box signal processing libraries that invert spectrograms to soundwaves and be identified by listeners as matching the original stimulus. For a practical experimental design and for testing the generalization of our methodology, we decided to train models and produce reconstructed spectral outputs from the NMED-H. Instead of a total of ten songs like in the NMED-T, the four songs from the NMED-H made it easier to test reconstructions across time slices and recorded participants; the training procedure in the deep CNN regressors stayed the same. Models trained on this dataset converged sooner and provided a steeper gradient descent during training which could be attributed to the familiarity of songs from participants [Hadjidimitriou and Hadjileontiadis, 2013b]. Because the best performing models using the NMED-T used the PSD and mel-spectrogram as input-target representations, we focused our training on the NMED-H with that specific representation pairing. Figure 6.5 provides a visual summary of the quality of reconstructions from the NMED-H. As a tradeoff for efficient modeling procedures and feature mapping within the proposed methodology, processing the outputs to listenable waveforms faces several steps of lossy transformations. The first comes from the actual deep learning models themselves where the input to target mapping assumes that the music signal is hidden/scrambled in cortical activity and attempts to recover that into a spectral representation. Second, those model outputs must be denormalized and set to a decibel range. Lastly, the spectrograms are transformed into waveforms by approximating the Short-Time Fourier Transform (STFT) magnitude from the mel power spectrum and reconstruction of the phase is done using Griffin Lim Algorithm (GLA). The mel-spectrogram in comparison to the linear is more lossy in the last two steps because mel is scaled as a non-linear function that represents human auditory perception across frequencies. Nevertheless, our modeling procedure finds it easier to map mel targets, so

we commit to lossy tradeoffs and adjust to produce listenable examples to present to participants during the behavioral evaluation. During the denormalization step we use the following transformation:

$$\text{Mel dB} = (\text{mel}^T \times \text{Max dB}) - \text{Max dB} + \text{ref dB} \tag{8}$$

In Equation 1, mel^T is the mel-spectrogram we aim to denormalize set as a matrix transpose and Max dB is set as a constant to 100 while ref dB is set as a constant within a song but varies across songs. To avoid presenting jarring noise to participants, ref dB becomes a free parameter set to [46 dB, 46 dB, 52 dB, 43 dB] for songs 1-4 respectively. The parameter value was set simply by randomly sampling 5 second examples and listening to how they would sound with values between 40 dB – 60 dB. The evaluation of which value produced the least jarring example was decided by the researcher; therefore we encourage any other studies following this methodology to play around with their own constraints relevant to their own studies. In Figure 6.5 it is demonstrated how well the waveforms come out despite the lossyness of the inversion process. The reconstructed spectrograms in Figure 6.5 shows how information across frequencies are recovered and how that translates to alignment in amplitude from target to reconstructed waveform.

The produced listenable examples were then presented to participants in a listening task to evaluate whether their quality was good enough to be identifiable. The experiment was a two-alternative AB-X task in which a participant had to pick between sounds A or B that matches the unlabeled sound X. The target sound X was a 5 second example from one of the four song stimuli in the NMED-H. Sounds A or B contained a non-corresponding foil and a corresponding 5 second reconstruction from our deep regressor model that was taken from spectra to a waveform. Participants were tested on a total of 24 trials plus 4 practice trials at the start of the experiment. The reconstructed model output produced a total of forty-eight examples which then half of the examples were used for one version of the experiment and the other half for a second version. Each version of the experiment controlled for balancing presentations across time by picking examples that belonged to the beginning, middle, and end of the songs. Furthermore, the

experiment also contained the same number of examples per song while controlling for each trial to have reconstructions coming from different participants from the NMED-H. Trials were all randomized and foils balanced to minimize the repetition of reconstructions from a specific song or participant. During the AB-X discrimination task, participants were given a max of 30 minutes to finish and were allowed to listen to the sound samples as many times as they liked before making their choices, while also not taking longer than 1 minute per presentation. This experiment collected responses from a total of 16 participants with 8 participants for each version of the task. On average, participants had an 85% success rate (50% chance) with max performance of 95.83% and minimum performance of 66.67%. The average performance lines up with output classification, adding evidence of not just robust image reconstruction but also the perceived interpretability of listeners when inverting the spectrum.

6.4 Conclusion

Here we show that it is possible to reconstruct music presented to participants using their EEG responses as input to a quality that allows others to correctly identify it when they listen to the music recovered from a brain response. The EEG to music stimuli mapping approach allows for reconstructions to preserve time-dependencies, during concatenated examples, necessary in the perception of music rather than taking an average of the stimuli class. Furthermore, no heavy feature extraction from the EEG or stimuli signals are needed in this process and only requires the EEG input to be transformed into a power spectrum for high quality reconstructions. With this we have shown that a computer vision approach has allowed for the processing of EEG responses to long naturalistic music stimuli as well as the ability to recover information across all frequency components. This is something that goes beyond the capabilities of complexity matching paradigms, as these straightforward methods do not just statistically measure correlated dynamics to the stimuli but actually recover the signal itself. The question to whether this is a result of the strength of deep learning models, the

extent to how much information is contained in brain responses, or both can now be asked in a more encapsulating way via training these reconstruction models. This is because it releases many experimental constraints and reduces the differences between researchers applying their own feature extraction steps. Therefore, making it easier to simply collect more data and learn about how data from different brains can be understood. In the final chapter, this last point will be reviewed and synthesized given the data and results discussed so far that, although may have been antithetical to the original aims and purposes, will argue this as a stronger epistemic path if not a natural consequence to the underlying questions we seek to explore.

Chapter 7

The Synthesis

Chapter 1 outlined a thesis for how a Complex Systems & Dynamics approach in Cognitive Science has made predictions on how the temporal structure of complex systems resonate via principles of Maximal Information Transfer, specifically in the case of speech and music. Empirical validation of these predictions has been demonstrated through the measurement of Complexity Matching effects. A core theoretical prediction included validation through empirical measurements of Complexity Matching effects of signals such as music and speech to corresponding neuronal activity, encapsulating Maximal Information Transfer of these related perceptual systems. In reference to the original theoretical grounding, the empirical results presented in this dissertation help extend and deconstruct these predictions. Methodological advancements and constraints have shaped how we continue to ask these questions, putting forth new motivation for the exploration of temporal structure of acoustic stimuli and brain responses. Like in many domains, Machine/Deep Learning approaches have found themselves relevant at moving forward the conversation where experimental paradigms have often been stuck. From neural networks back to neural networks, a self-referential epistemic loop is formed arguing for the deconstruction of how scientific questions are explored. In this chapter, what this means for how a Complex Systems & Dynamics camp of Cognitive Science when it studies speech and music is discussed.

7.1 The Limitations

In part of the original theoretical framework outlined by [West et al., 2008] on Maximal Information Transfer, it was predicted that signals such as speech and music should demonstrate Complexity Matching effects empirically. The original proposition was established through numerical analyses from stochastic neural network models. Both models were grounded in examples in the literature trying to explain emergent behavior of varying systems. The first model posited the collective behavior of a network of stochastic clocks [Bianco et al., 2008], while the second referenced the [Mirollo and Strogatz, 1990] stochastic model of neuronal synchronization based on the integrative theory of alpha oscillations [Başar, 2006] in the brain as measured by Electroencephalogram (EEG) studies. Why the model comparison made sense was because it was suggested that alpha rhythm acts as a nonlinear clock which could facilitate mechanisms of association in the brain as a gating function.

With this, the prediction suggested that because the stochastic complexity of those two network models matched, that a cooperation of non-ergodic renewal events emerged via their coupling. Such a finding was paired with emerging studies linking the inverse power-laws in natural systems to proposed network models of the human brain as showing high clustering and inverse power-laws of its edges [Holme and Kim, 2002]. In fact, the Allan Factor (AF) method borrows from Complexity Matching's emphasis of non-ergodic/non-Poisson events instead of continuous processes, through the implementation of capturing peak-amplitude events. This made sense in Chapter 2 when seeing the connection between linguistic units, which can be thought of as symbolic events of a production process, and the temporal structure measured via the AF method. Here the often-used nomenclature of a measured 'Hierarchy' makes sense in reference to symbolic units that construct linguistic hierarchies as well as musical syntax. For example, we can start with a simple note that is then grouped with others to create a motif that as the song goes on is repeated to outline phrasal structure within chord progressions and so on to build a story that is unwrapped over time. In [Falk and Kello, 2017], the modeling of the relationship between the speech and linguistic units show a

strong correlation, but the data being compared comes from scripted productions of speech and with more careful feature extraction of linguistic units at certain timescales. What we see in the data from Chapter 2 is that there is a gap between that carefully produced discourse between mothers & infants and the spontaneous conversations contained in the Buckeye corpus. What the physical speech signal carries is often all encapsulating to every utterance, including non-linguistic utterances, such as coughs or other interruptions, but in turn are still part of the production between interlocutors. Something that, via this neat modeled symbolic hierarchy, is difficult to always capture through feature extraction of the speech signal. These results put forth the question of: What is this 'hierarchy' really?

Results from Chapter 3 take this from the general realm of coordination between people and focuses on the production aspect of these signals. In [Schneider et al., 2020], they demonstrate that the HTS within participants as they produce speech across conditions is correlated. This puts an easy-to-understand case of people matching with themselves, as these long-range correlations within these signals should perhaps be self-referential if they are produced by the same system in similar conditions. But then, if we experimentally control for the production of these signals, what we see is something different. Through algorithmic and behavioral manipulation of these signals, the characteristic HTS of the speech changes. Algorithmic manipulations hint at the diminishing of prosodic embellishment that occurs, which lowers the variability of clustering in AF functions. Behavioral manipulations on speaking rate simulate restraints people face when they are impromptu pushed to speed up or slow their speech productions. This creates a rolling effect on the AF functions, which much like a seesaw, sees the variability go down on one end of the timescales and raises it to the other. Again, coming back to this notion of 'hierarchy'. Certain levels within that hierarchy seem to be emphasized one way or another, but does this mean that it still remains as a proper reference for the information within? In other words, can I simply refer back to a linguistic hierarchy and say that speaking rate is augmenting phrase structure and downplaying syllables? Bear in mind that the only variable being tested is speaking rate, and that participants are still reading from the same speech

excerpts, without any words or syntactic demarcations being manipulated.

If signals produced by interlocutors can show a convergence in their temporal structure through the coupling of their common ground in conversation, then can the same be expected of perceptual systems including the cortex? Let's take the results from Chapter 2 and juxtapose it to the paradigm of EEG to acoustic stimuli Complexity Matching, to better understand this question. Given that experimental controls and parameters allow this, we can have an experiment that couples a speaker and a listener, similar to the coupling between interlocutors. If the speaker adjusts the scaling of their speech via speaking rate, then we would expect such a correspondence to be reflected in the measured brain response. Similarly, if we present to someone music that has a scaling temporal structure uniquely different from the 'ideal' $1/f$ and if the perceptual system of the listener is coupled to the stimuli production, then we can also expect that somewhere in the cortical activity we will be able to see that correspondence. Evidence for that was not able to come out in the study presented in Chapter 4. Furthermore, studies who found better success focused on only using classical music, which is characteristic of the $1/f$ scaling, and through their own methods found evidence for Complexity Matching. Could it be that we should only expect a $1/f$ to $1/f$ resonance in related EEG paradigms? Or is it that, how I outline in Chapter 4, that such results happen to be the simplest case and further methodological developments need to occur for such a result to be discovered? I argue here that the latter is a more sensible scenario. This is given the limiting factors of feature extraction, repeated measures, and localization.

For these questions to be better answered via Complexity Matching, it seems that a path towards experimental over parameterization would be needed. We have already seen that this could lead to a decrease in generalization, and goes away from naturalistic components. Instead, in Chapter 4 we see that a machine learning approach is able to rescue many of the marginally significant and null results from the study. Thus, moving the exploration of these questions in a different direction; towards acoustic information retrieval from brain responses. Chapters 5 & 6 not only make a general case for Deep Learning approaches, but it puts forth a new

way of thinking about the EEG data itself. A Computer Vision approach in Deep Learning allowed for modeling procedures to be strong in performance, efficient, and be by far more naturalistic than many relevant approaches. The progression from Chapter 2-6 in turn reflecting the deconstruction of theoretical predictions via the limitations and constraints in experimental paradigms. The deconstruction of one way of producing new testable hypothesis leading to another, in the natural progression of scientific advancements.

7.2 Anticipating

As it has been discussed, Complexity Matching effects are expected when systems are appropriately coupled. In the behavioral experiments and EEG experiments, what coupling means is not always simple, despite these studies showing evidence for it. Between interlocutors, there are several factors to consider involving physical, perceptual, and a coupling of mental states to facilitate the information exchange. In the EEG paradigms, the focus is on the perceptual coupling as the coupling between systems is between the participant and the physical implementation of the experiment, which could include the chair, speaker/headphones, the recording room, and the EEG system itself. The purpose of the physical implementation in the EEG recording room becomes crucial to the quality of data collected, as its organization is meant to facilitate the presentation of the stimuli. This is something that across research labs contains many discrepancies, but despite the differences have still been able to replicate results of participants being coupled to presented stimuli. Here the theoretical framework used to explain these cognitive processes, poses the perceptual loop of an acoustic stimulus to brain response as an anticipatory process. In the EEG experimental paradigm this would include the coupling between the recording and presentation system with the perceptive mechanisms of the participant. In this scenario we can describe the coupling between systems (the stimulus presentation system and the participant) as defining a third larger system representing the entirety of the experiment. The process here being anticipatory simply means that the stimulus presentation perturbs the

dynamics of the participant, and the phase transitions of future events can be in part traced back to the perturbations of the stimulus. In the strict sense, a strong anticipatory process [Stepp and Turvey, 2010] would mean that the system being measured (the brain) does not need to create a model of future events and predict its trajectory, but rather the trajectory of its dynamics is in a negative phase relationship to that of the stimulus, describing the coupling of complex oscillators (music waveform and neural activity).

To anticipate or 'to take before' or 'to follow a path before' presents an Occam's razor approach to the representation hungry problem of modeling these cognitive processes. This is not to say that all observed processes can be explained as strong anticipation, as we know that processes at certain levels of cognition explicitly predict or simulate future states using historical priors to model their environment. But rather, the theoretical stance taken here is meant to oppose the notion that this modeling of future states happens at all levels a la Clark's Radical Predictive Processing [Clark, 2015]. Clark's levels of Predictive Processing much like Dante's fourth circle of hell, is stuck in an epistemic greed where everything is considered a representation and every process a predictive one. Maximal Information Transfer proposes that complexity matching occurs without the need of a 'predictive' process or 'representation' of the system as it simply demonstrates that the extent of coupling between systems shapes the alignment or resonance of the produced temporal structure. The extent of coupling can be described as the overlapping bandwidth during communication, where anything outside of the shared bandwidth limits does not resonate. Early information retrieval studies can also frame their processes in a similar way. For example, the early studies measuring brainstem responses to speech vowels were able to develop experimental methods that would repeatedly present speech vowel stimuli, and the measured responses were good enough to average and reconstruct it into a listenable waveform strongly resembling the original stimuli [Johnson et al., 2005]. Despite the information retrieval coming from sub-cortical activity, it was easier to understand the physical coupling of acoustic stimuli and neural response. This is due to the well-studied anatomical relationships between the sound itself, and how that information physically

resonates in these listening pathways. For example, how the cochlea is shaped to vibrate at different frequencies we perceive to how electrical impulses on the brainstem reflect the amplitude dynamics of the incoming acoustic signal. A simple enough physical coupling from the vibrating air, to vibrating flesh, and up to the vibration of electrical signals occurs while not needing a framing of 'prediction' in the process.

The information transfer in these perceptual processes such as in the brainstem recordings, also act like a signal processing task where the researcher tracks how the information is processed and maintained as it passes through auditory pathways. From a modeling perspective, this is also happening with artificial neural networks acting as another artificial 'perceptual' pathway that processes the observed activity, and maps that data into an acoustic feature the researcher assumes is contained in the brain response. Therefore, this notion of anticipatory processes I have described works well with the proposed deep learning methods presented in Chapter 5 6. Given that the experimental paradigm works well enough to couple the acoustic stimulus dynamics to the perceptual pathways leading up to the measured cortical activity, the information retrieval modeling can simply clean up the signal back to how it looked like when it was an input. If this assumption of how acoustic information resonates in cortical activity is true, then treating EEG as images can work well as it bypasses the issues of feature extraction, localization, and repeated measures. This is because we can assume that the resonance exists in one form or another in cortical activity without needing to develop methods to find the needle in the needlestack. Also, because we do not need to treat this as an active inferential process where the salience of representations is key to the quality of the data, we also loosen up constraints in the experimental paradigm, like in the Multiscale Entropy complexity matching study where they depended on an active behavioral attention task. Ultimately this proposed model of coupled oscillators passively reflecting information of temporal structure, opens towards more efficient data collection and naturalistic stimulus presentations.

7.3 The AI Inclusion

The inclusion of Machine/Deep Learning analyses and methodologies has been a phenomenon existing across domains of experimental work from anywhere in medical imaging, bioengineering, and to psychological experiments [Kim et al., 2020, Orrù et al., 2020]. No strong argument is made here about any transformative epistemic influence of simply using Machine Learning analysis as added evidence in studies. This is something that scientists have been doing for a long time with classic methods of statistical modeling such as fitting generalized linear models to their data to see how their experimental operationalization can capture some relationship that may be generalized to other studies or replicate findings. On the other hand, the notion of molding your experimental paradigms around Machine Learning approaches is something that has been showing more momentum in the last 5 years. Like much of the history of developments in Machine Learning or Artificial Intelligence research, a fair amount of skepticism can be held to some studies following the hype, whether it would be for incentives involving the relevancy at a broader scale of the research or the funding itself. But filtering out studies that are either one-hit wonders or simply develop to publish catchy headlines, a more concrete paradigm has been recognized with well argued motivations, named as 'Data-Driven'. Such a methodology was discussed in a (2018) issue of 'Trends in Cognitive Sciences' by Jack, Crivelli, and Wheatley outlining the benefits of a Data-Driven paradigm for human psychological experiments [Jack et al., 2018]. The main argument that was presented for Data-Driven experiments was to relax theory driven constraints. That meant that experiments could now have less a priori assumptions, capture pan-cultural patterns that often times were overlooked through a western centric lens, and push for common grounds for comparisons through the investment of publicly available datasets.

A more recent review on this topic explicitly discusses the Machine Learning aspect of Data-Driven methods in psychological experiments [Vélez, 2021]. Some of the points brought up in that review that are of relevance to this discussion include the benefits from hosting databases and the democratization of Machine Learning. The latter point simply refers to the ever-expanding open-source and importable

libraries in Julia, R, and Python which allow for computational research to be accessible to a larger pool of researchers. Furthermore, the constraints of owning and managing hardware responsible for processing large datasets and training heavy models has been pushed towards a cloud computing based market, which highlights the convenient service of offloading information-technology work and letting the researcher simply be a researcher. For example, the models presented in Chapters 5 & 6 were developed primarily using Google Colab, which in some situations has allowed me to train Deep Learning models with trainable parameter counts in the millions from my phone using a coffee shop's Wi-Fi. To the prior point of benefits from a hosted publicly available database, it seems that such an endeavor is the most transformative for how researchers collect and analyze data. This not only argues for moving past small sample sizes in esoteric experimental paradigms, but it also frees up how a researcher can generate testable hypotheses. For example, the 'Emergence of Communication Lab' at UCLA led by Dr.Warlaumont hosts the HomeBank database of naturalistic day long audio recordings of infant interactions funded by a National Science Foundation (NSF) grant of resource implementations for data intensive research [VanDam et al., 2016]. With access to such a database both collaborators and critics alike can further connect with their research as they have this shared common ground. For these databases to function in such a manner, the collected data needs to have little parameterization restricting its dimensionality while also having a pipeline for the merging of target variables. Simply having raw recordings would put too much of a burden on any researcher from the outside looking in. Meanwhile, a pipeline for target variables would aid in the proliferation of models mapped to varying tasks, whether it would be classification, regression, or distribution fitting such as in generative models.

Such an approach would be of great benefit to Cognitive Neuroscience studies as it is often the case that parametric modeling of these data lacks generalization strength across participants due to how expensive it is to collect a larger number of participants. Data hosting applications such as OpenNeuro have made significant strides towards this end. Through their data hosting services, researchers can go to one place to search and download data/code from published studies. With

this, accessibility is facilitated, but the issue of small datasets still looms, perhaps to be serviced by future applications that would ideally allow the collaboration of experimentalists worldwide to develop large datasets of brain responses to the same sizable impact that databases like HomeBank have. As it has been argued so far, the Computer Vision methods developed in Chapters 5 & 6 also present an applicable use case for these ideal large EEG datasets given that the motto of Deep Learning data usage is 'more is better'. Efficient data processing in this case would allow for the collection of long recordings to be used for general purposes and not just for one specific task, as it has been shown that classification of sentiment, song name, and the image regression task have performed well. It would also be sensible to say that such an endeavor would add to the collaborative development of computational libraries aiding researchers in their data processing. This could one day bring back Complexity Matching studies with EEG to focus under a facilitated process that includes better data and stronger tools.

Lastly, the adaptation of these experimental paradigms does not stop with the development of Data-Driven academic research. Consumer based sectors have pushed for services dependent on the capitalization of human data through the proliferation of user interactions that facilitate data mining [Reyes, 2020]. Marketing goals of capturing consumer profiles to highly specialized consumer needs often referred as 'one customer one product' further exemplifies the creation of demand sold to consumers. In this case, substantial capital gains can be framed as the centralized ownership of 'goldmine' databases, where data science insights become the golden bullions sold as a premium. Such a market then is responsible for the explosion in human data which is moving to the precision of what has traditionally been expected from laboratory experiments via the integration of emerging technologies from Internet of Things [Yan et al., 2020] and the newly revamped buzzword 'Metaverse', which broadly encapsulates human computer interaction via an embodied internet [Sparkes, 2021]. These emerging technologies benefit from the increasing placement of precise sensors measuring everyday human behaviors such as heartrate, sentiment, eye-tracking, and posture. At the center of these Data-Driven user interactions exists the Artificial Intelligence systems that

they train which in turn train consumer behavior via implementations such as recommender systems. It is not farfetched to argue that such a scaling of mining precise human behavior would be able to explore difficult topics such as the anticipatory connection of temporal structure from acoustic stimuli to brain responses. But it is not the technical capability that may shape epistemic developments, but rather the material conditions of funding, as investments in public education decrease and the privatization of the academic sector increases [Price et al., 2012]. The exploration of the scientific questions discussed so far seem to be headed beyond a paywall into a behind-the-scenes environment guided by the privatization of observed human behavior. But as Scientists, Philosophers, and Academics it is our imperative to anticipate these conditions and clean up the noise hiding the long-range correlations of human exploration across time.

Bibliography

[Abney et al., 2014] Abney, D. H., Paxton, A., Dale, R., and Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, 143(6):2304.

[Abney et al., 2021] Abney, D. H., Paxton, A., Dale, R., and Kello, C. T. (2021). Cooperation in sound and motion: Complexity matching in collaborative interaction. *Journal of Experimental Psychology: General*.

[Allegrini et al., 2009] Allegrini, P., Menicucci, D., Bedini, R., Fronzoni, L., Gemignani, A., Grigolini, P., West, B. J., and Paradisi, P. (2009). Spontaneous brain activity as a source of ideal 1/f noise. *Physical Review E*, 80(6):061914.

[Allegrini et al., 2010] Allegrini, P., Paradisi, P., Menicucci, D., and Gemignani, A. (2010). Fractal complexity in spontaneous eeg metastable-state transitions: new vistas on integrated neural dynamics. *Frontiers in physiology*, 1:128.

[Almurad et al., 2018] Almurad, Z. M., Roume, C., Blain, H., and Delignières, D. (2018). Complexity matching: restoring the complexity of locomotion in older people through arm-in-arm walking. *Frontiers in Physiology*, 9:1766.

[Anderson, 2000] Anderson, J. R. (2000). *Learning and memory: An integrated approach*. John Wiley & Sons Inc.

[Baddeley, 1998] Baddeley, A. (1998). Recent developments in working memory. *Current opinion in neurobiology*, 8(2):234–238.

[Bak et al., 1987] Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, 59(4):381.

[Barabási, 2003] Barabási, A.-L. (2003). Linked: The new science of networks.

[Baronchelli et al., 2013] Baronchelli, A., Ferrer-i Cancho, R., Pastor-Satorras, R., Chater, N., and Christiansen, M. H. (2013). Networks in cognitive science. *Trends in cognitive sciences*, 17(7):348–360.

[Bartlett, 1950] Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37(1/2):1–16.

[Başar, 2006] Başar, E. (2006). The theory of the whole-brain-work. *International Journal of Psychophysiology*, 60(2):133–138.

[Beggs and Plenz, 2003] Beggs, J. M. and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *Journal of neuroscience*, 23(35):11167–11177.

[Bianco et al., 2008] Bianco, S., Geneston, E., Grigolini, P., and Ignaccolo, M. (2008). Renewal aging as emerging property of phase synchronization. *Physica A: Statistical Mechanics and its Applications*, 387(5-6):1387–1392.

[Biber et al., 1998] Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

[Bird et al., 2019] Bird, J. J., Ekart, A., Buckingham, C. D., and Faria, D. R. (2019). Mental emotional sentiment classification with an eeg-based brain-machine interface. In *Proceedings of theInternational Conference on Digital Image and Signal Processing (DISP'19)*.

[Boyland, 1996] Boyland, J. T. (1996). *Morphosyntactic change in progress: A psycholinguistic treatment*. University of California, Berkeley.

[Bryce and Sprague, 2012] Bryce, R. and Sprague, K. (2012). Revisiting detrended fluctuation analysis. *Scientific reports*, 2(1):1–6.

[Buiatti et al., 2007] Buiatti, M., Papo, D., Baudonnière, P.-M., and van Vreeswijk, C. (2007). Feedback modulates the temporal scale-free dynamics of brain electrical activity in a hypothesis testing task. *Neuroscience*, 146(3):1400–1412.

[Bybee and Scheibman, 1999] Bybee, J. and Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in english.

[Carpentier et al., 2020] Carpentier, S. M., McCulloch, A. R., Brown, T. M., Faber, S. E., Ritter, P., Wang, Z., Salimpoor, V., Shen, K., and McIntosh, A. R. (2020). Complexity matching: brain signals mirror environment information patterns during music listening and reward. *Journal of Cognitive Neuroscience*, 32(4):734–745.

[Chen et al., 2014] Chen, S.-H., Hsieh, C.-H., Chiang, C.-Y., Hsiao, H.-C., Wang, Y.-R., Liao, Y.-F., and Yu, H.-M. (2014). Modeling of speaking rate influences on mandarin speech prosody and its application to speaking rate-controlled tts. *IEEE/ACM transactions on audio, speech, and language processing*, 22(7):1158–1171.

[Chillara et al., 2019] Chillara, S., Kavitha, A., Neginhal, S. A., Haldia, S., and Vidyullatha, K. (2019). Music genre classification using machine learning algorithms: a comparison. *Int Res J Eng Technol*, 6(5):851–858.

[Clark, 2015] Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, 53:3–27.

[Coey et al., 2016] Coey, C. A., Washburn, A., Hassebrock, J., and Richardson, M. J. (2016). Complexity matching effects in bimanual and interpersonal syncopated finger tapping. *Neuroscience letters*, 616:204–210.

[Coffey et al., 2019] Coffey, E. B., Nicol, T., White-Schwoch, T., Chandrasekaran, B., Krizman, J., Skoe, E., Zatorre, R. J., and Kraus, N. (2019). Evolving perspectives on the sources of the frequency-following response. *Nature communications*, 10(1):1–10.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Delignières et al., 2016] Delignières, D., Almurad, Z. M., Roume, C., and Marmelat, V. (2016). Multifractal signatures of complexity matching. *Experimental brain research*, 234(10):2773–2785.

[Dellwo et al., 2003] Dellwo, V., Wagner, P., Solé, M., Recasens, D., and Romero, J. (2003). Relations between language rhythm and speech rate.

[Di Liberto et al., 2015] Di Liberto, G. M., O'sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465.

[Ding et al., 2016] Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158–164.

[Ding et al., 2017] Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81:181–187.

[Dmochowski et al.,] Dmochowski, A. M. N., Berger, J., and Naturalistic Music, E. Dataset—hindi,". *Latebreaking/demo*, 3.

[Dubois, 2003] Dubois, D. M. (2003). Mathematical foundations of discrete and functional systems with strong and weak anticipations. In *Anticipatory behavior in adaptive learning systems*, pages 110–132. Springer.

[Ebrahimpour et al., 2020] Ebrahimpour, M., Shea, T., Danielescu, A., Noelle, D., and Kello, C. (2020). End-to-end auditory object recognition via inception nucleus. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 146–150. IEEE.

[Eliazar and Klafter, 2010] Eliazar, I. and Klafter, J. (2010). Universal generation of 1/f noises. *Physical Review E*, 82(2):021109.

[Ellis, 2004] Ellis, D. P. W. (2004). Sinewave speech analysis/synthesis in matlab. Web Resource, available: http://www.ee.columbia.edu/ln/labrosa/matlab/sws/.

[Falk and Kello, 2017] Falk, S. and Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, 163:80–86.

[Farbood et al., 2015] Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., and Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in neuroscience*, 9:157.

[Fenk-Oczlon, 1989] Fenk-Oczlon, G. (1989). Word frequency and word order in freezes.

[Freeman and Holmes, 2005] Freeman, W. J. and Holmes, M. D. (2005). Metastability, instability, and state transition in neocortex. *Neural Networks*, 18(5-6):497–504.

[Glaser et al., 2020] Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., and Kording, K. P. (2020). Machine learning for neural decoding. *Eneuro*, 7(4).

[Golshan et al., 2020] Golshan, H. M., Hebb, A. O., and Mahoor, M. H. (2020). Lfp-net: A deep learning framework to recognize human behavioral activities using brain stn-lfp signals. *Journal of neuroscience methods*, 335:108621.

[Golumbic et al., 2013] Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–991.

[Grigolini, 2017] Grigolini, P. (2017). Call for papers: Special issue on evolutionary game theory of small groups and their larger societies.

[Gururani et al., 2018] Gururani, S., Summers, C., and Lerch, A. (2018). Instrument activity detection in polyphonic music using deep neural networks. In *ISMIR*, pages 569–576.

[Hachinski and Hachinski, 1994] Hachinski, K. V. and Hachinski, V. (1994). Music and the brain. *CMAJ: Canadian Medical Association Journal*, 151(3):293.

[Hadjidimitriou and Hadjileontiadis, 2013a] Hadjidimitriou, S. K. and Hadjileontiadis, L. J. (2013a). Eeg-based classification of music appraisal responses using time-frequency analysis and familiarity ratings. *IEEE Transactions on Affective Computing*, 4(2):161–172.

[Hadjidimitriou and Hadjileontiadis, 2013b] Hadjidimitriou, S. K. and Hadjileontiadis, L. J. (2013b). Eeg-based classification of music appraisal responses using time-frequency analysis and familiarity ratings. *IEEE Transactions on Affective Computing*, 4(2):161–172.

[Hausen et al., 2013] Hausen, M., Torppa, R., Salmela, V. R., Vainio, M., and Särkämö, T. (2013). Music and speech prosody: a common rhythm. *Frontiers in psychology*, 4:566.

[Haykin, 2012] Haykin, S. (2012). *Cognitive dynamic systems: perception-action cycle, radar and radio*. Cambridge University Press.

[Hofstadter, 1979] Hofstadter, D. R. (1979). *Gödel, escher, bach*. Basic books New York.

[Holme and Kim, 2002] Holme, P. and Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Physical review E*, 65(2):026107.

[Hopper and Bybee, 2001] Hopper, P. J. and Bybee, J. L. (2001). Frequency and the emergence of linguistic structure. *Frequency and the Emergence of Linguistic Structure*, pages 1–502.

[Hore and Ziou, 2010] Hore, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE.

[Horváth, 2015] Horváth, J. (2015). Action-related auditory erp attenuation: Paradigms and hypotheses. *Brain Research*, 1626:54–65.

[Hsü and Hsü, 1991] Hsü, K. J. and Hsü, A. (1991). Self-similarity of the" 1/f noise" called music. *Proceedings of the National Academy of Sciences*, 88(8):3507–3509.

[Hurst, 1951] Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799.

[Jack et al., 2018] Jack, R. E., Crivelli, C., and Wheatley, T. (2018). Data-driven methods to diversify knowledge of human psychology. *Trends in cognitive sciences*, 22(1):1–5.

[Jennings et al., 2004] Jennings, H. D., Ivanov, P. C., Martins, A. d. M., da Silva, P., and Viswanathan, G. (2004). Variance fluctuations in nonstationary time series: a comparative study of music genres. *Physica A: Statistical Mechanics and its Applications*, 336(3-4):585–594.

[Johnson et al., 2005] Johnson, K. L., Nicol, T. G., and Kraus, N. (2005). Brain stem response to speech: a biological marker of auditory processing. *Ear and hearing*, 26(5):424–434.

[Jones, 2010] Jones, M. R. (2010). Attending to sound patterns and the role of entrainment. *Attention and time*, 317:330.

[Jones et al., 2006] Jones, M. R., Johnston, H. M., and Puente, J. (2006). Effects of auditory pattern structure on anticipatory and reactive attending. *Cognitive psychology*, 53(1):59–96.

[Jun, 2003] Jun, S.-A. (2003). The effect of phrase length and speech rate on prosodic phrasing. In *proceedings of the XVth international congress of phonetic sciences*, pages 483–486.

[Kavasidis et al., 2017] Kavasidis, I., Palazzo, S., Spampinato, C., Giordano, D., and Shah, M. (2017). Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817.

[Keller and Mrsic-Flogel, 2018] Keller, G. B. and Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron*, 100(2):424–435.

[Kello et al., 2017] Kello, C. T., Bella, S. D., Médé, B., and Balasubramaniam, R. (2017). Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of The Royal Society Interface*, 14(135):20170231.

[Kim et al., 2020] Kim, M., Yan, C., Yang, D., Wang, Q., Ma, J., and Wu, G. (2020). Deep learning in biomedical image analysis. In *Biomedical information technology*, pages 239–263. Elsevier.

[KISH et al., 2001] KISH, L. B., HARMER, G. P., and ABBOTT, D. (2001). Information transfer rate of neurons: stochastic resonance of shannon's information channel capacity. *Fluctuation and Noise Letters*, 1(01):L13–L19.

[Kiyono et al., 2005] Kiyono, K., Struzik, Z. R., Aoyagi, N., Togo, F., and Yamamoto, Y. (2005). Phase transition in a healthy human heart rate. *Physical review letters*, 95(5):058101.

[Kragh, 2012] Kragh, H. (2012). *Niels Bohr and the quantum atom: The Bohr model of atomic structure 1913-1925*. OUP Oxford.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[Krug, 1998] Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English linguistics*, 26(4):286–320.

[Lawhern et al., 2018] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013.

[Lewis, 2005] Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and brain sciences*, 28(2):169–194.

[Lindenberg and West, 1990] Lindenberg, K. and West, B. J. (1990). *The nonequilibrium statistical mechanics of open and closed systems*. VCH.

[Lise and Paczuski, 2001] Lise, S. and Paczuski, M. (2001). Self-organized criticality and universality in a nonconservative earthquake model. *Physical Review E*, 63(3):036111.

[Livezey and Glaser, 2021] Livezey, J. A. and Glaser, J. I. (2021). Deep learning approaches for neural decoding across architectures and recording modalities. *Briefings in bioinformatics*, 22(2):1577–1591.

[Losorelli et al., 2017] Losorelli, S., Nguyen, D. T., Dmochowski, J. P., and Kaneshiro, B. (2017). Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music. In *ISMIR*, volume 3, page 5.

[Mandy and Lai, 2016] Mandy, W. and Lai, M.-C. (2016). Annual research review: The role of the environment in the developmental psychopathology of autism spectrum condition. *Journal of Child Psychology and Psychiatry*, 57(3):271–292.

[Marković and Gros, 2014] Marković, D. and Gros, C. (2014). Power laws and self-organized criticality in theory and nature. *Physics Reports*, 536(2):41–74.

[Marmelat and Delignières, 2012] Marmelat, V. and Delignières, D. (2012). Strong anticipation: complexity matching in interpersonal coordination. *Experimental brain research*, 222(1):137–148.

[Mirollo and Strogatz, 1990] Mirollo, R. E. and Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, 50(6):1645–1662.

[Moinnereau et al., 2018] Moinnereau, M.-A., Brienne, T., Brodeur, S., Rouat, J., Whittingstall, K., and Plourde, E. (2018). Classification of auditory stimuli from eeg signals with a regulated recurrent neural network reservoir. *arXiv preprint arXiv:1804.10322*.

[Moss et al., 2004] Moss, F., Ward, L. M., and Sannita, W. G. (2004). Stochastic resonance and sensory information processing: a tutorial and review of application. *Clinical neurophysiology*, 115(2):267–281.

[Musacchia et al., 2014] Musacchia, G., Large, E. W., and Schroeder, C. E. (2014). Thalamocortical mechanisms for integrating musical tone and rhythm. *Hearing research*, 308:50–59.

[Nolan and Jeon, 2014] Nolan, F. and Jeon, H.-S. (2014). Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130396.

[Nozaradan et al., 2011] Nozaradan, S., Peretz, I., Missal, M., and Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience*, 31(28):10234–10240.

[Ofner and Stober, 2018] Ofner, A. and Stober, S. (2018). Shared generative representation of auditory concepts and eeg to reconstruct perceived and imagined music. In *ISMIR*, pages 392–399.

[Orrù et al., 2020] Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in psychology*, 10:2970.

[Palmer and Hutchins, 2006] Palmer, C. and Hutchins, S. (2006). What is musical prosody? *Psychology of learning and motivation*, 46:245–278.

[Pandey et al., 2022] Pandey, P., Sharma, G., Miyapuram, K. P., Subramanian, R., and Lomas, D. (2022). Music identification using brain responses to initial snippets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1246–1250. IEEE.

[Patel, 2003] Patel, A. D. (2003). Language, music, syntax and the brain. *Nature neuroscience*, 6(7):674–681.

[Patterson et al., 2021] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

[Paxton and Dale, 2013] Paxton, A. and Dale, R. (2013). Argument disrupts interpersonal synchrony.

[Pease et al., 2018] Pease, A., Mahmoodi, K., and West, B. J. (2018). Complexity measures of music. *Chaos, Solitons & Fractals*, 108:82–86.

[Peng et al., 1994] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of dna nucleotides. *Physical review e*, 49(2):1685.

[Petrushin et al., 2018] Petrushin, A., Tessadori, J., Barresi, G., and Mattos, L. S. (2018). Effect of a click-like feedback on motor imagery in eeg-bci and eye-tracking hybrid control for telepresence. In *2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 628–633. IEEE.

[Pickering and Garrod, 2004] Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

[Pitt et al., 2005] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

[Polak, 2020] Polak, R. (2020). Non-isochronous meter is not irregular: A review of theory and evidence. In *15. Jahreskongress der Gesellschaft für Musiktheorie*, pages 365–379. Georg Olms.

[Power et al., 2012] Power, A. J., Mead, N., Barnes, L., and Goswami, U. (2012). Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. *Frontiers in Psychology*, 3:216.

[Price et al., 2012] Price, T. A., Duffy, J., and Giordani, T. (2012). *Defending public education from corporate takeover*. University Press of America.

[Qin et al., 2018] Qin, Z., Zhang, Z., Chen, X., Wang, C., and Peng, Y. (2018). Fd-mobilenet: Improved mobilenet with a fast downsampling strategy. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1363–1367. IEEE.

[Ramirez-Aristizabal et al., 2022] Ramirez-Aristizabal, A. G., Ebrahimpour, M. K., and Kello, C. T. (2022). Image-based eeg classification of brain responses to song recordings. *arXiv preprint arXiv:2202.03265*.

[Ramirez-Aristizabal et al., 2018] Ramirez-Aristizabal, A. G., Médé, B., and Kello, C. T. (2018). Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons & Fractals*, 111:175–179.

[Redfern et al., 1994] Redfern, P., Minors, D., and Waterhouse, J. (1994). Circadian rhythms, jet lag, and chronobiotics: an overview. *Chronobiology international*, 11(4):253–265.

[Reed, 2003] Reed, W. J. (2003). The pareto law of incomes—an explanation and an extension. *Physica A: Statistical Mechanics and its Applications*, 319:469–486.

[Remez et al., 1981] Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212(4497):947–950.

[Reyes, 2020] Reyes, M. (2020). *Consumer behavior and marketing*. BoD–Books on Demand.

[Richman and Moorman, 2000] Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049.

[Rimmele et al., 2015] Rimmele, J. M., Golumbic, E. Z., Schröger, E., and Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, 68:144–154.

[Rimmele et al., 2018] Rimmele, J. M., Morillon, B., Poeppel, D., and Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in cognitive sciences*, 22(10):870–882.

[Rosenberg et al., 1974] Rosenberg, J. F. et al. (1974). *Linguistic representation*, volume 6. Taylor & Francis.

[Saeed et al., 2021] Saeed, A., Grangier, D., Pietquin, O., and Zeghidour, N. (2021). Learning from heterogeneous eeg signals with differentiable channel reordering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1255–1259. IEEE.

[Schneider et al., 2020] Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., and Kello, C. T. (2020). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition*, 23(4):845–857.

[Silagadze, 1999] Silagadze, Z. (1999). Citations and the zipf-mandelbrot's law. *arXiv preprint physics/9901035*.

[Skoe and Kraus, 2010] Skoe, E. and Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and hearing*, 31(3):302.

[Slanzi et al., 2017] Slanzi, G., Balazs, J. A., and Velásquez, J. D. (2017). Combining eye tracking, pupil dilation and eeg analysis for predicting web users click intention. *Information Fusion*, 35:51–57.

[Sonawane et al., 2021] Sonawane, D., Miyapuram, K. P., Rs, B., and Lomas, D. J. (2021). Guessthemusic: song identification from electroencephalography response. In *8th ACM IKDD CODS and 26th COMAD*, pages 154–162.

[Spampinato et al., 2017] Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., and Shah, M. (2017). Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817.

[Sparkes, 2021] Sparkes, M. (2021). What is a metaverse.

[Stawicki et al., 2017] Stawicki, P., Gembler, F., Rezeika, A., and Volosyak, I. (2017). A novel hybrid mental spelling application based on eye tracking and ssvep-based bci. *Brain sciences*, 7(4):35.

[Stephen et al., 2008] Stephen, D. G., Stepp, N., Dixon, J. A., and Turvey, M. (2008). Strong anticipation: Sensitivity to long-range correlations in synchronization behavior. *Physica A: Statistical Mechanics and its Applications*, 387(21):5271–5278.

[Stepp and Turvey, 2010] Stepp, N. and Turvey, M. T. (2010). On strong anticipation. *Cognitive systems research*, 11(2):148–164.

[Stober et al., 2014] Stober, S., Cameron, D. J., and Grahn, J. A. (2014). Classifying eeg recordings of rhythm perception. In *ISMIR*, pages 649–654.

[Supratak et al., 2017] Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008.

[Synigal et al., 2020] Synigal, S. R., Teoh, E. S., and Lalor, E. C. (2020). Including measures of high gamma power can improve the decoding of natural speech from eeg. *Frontiers in human neuroscience*, 14:130.

[Tal et al., 2017] Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., and Golumbic, E. Z. (2017). Neural entrainment to the beat: The "missing-pulse" phenomenon. *Journal of Neuroscience*, 37(26):6331–6341.

[Tao et al., 2021] Tao, L., Wang, G., Zhu, M., and Cai, Q. (2021). Bilingualism and domain-general cognitive functions from a neural perspective: A systematic review. *Neuroscience & Biobehavioral Reviews*, 125:264–295.

[Teixeira Borges et al., 2019a] Teixeira Borges, A. F., Irrmischer, M., Brockmeier, T., Smit, D. J., Mansvelder, H. D., and Linkenkaer-Hansen, K. (2019a). Scaling behaviour in music and cortical dynamics interplay to mediate music listening pleasure. *Scientific reports*, 9(1):1–15.

[Teixeira Borges et al., 2019b] Teixeira Borges, A. F., Irrmischer, M., Brockmeier, T., Smit, D. J., Mansvelder, H. D., and Linkenkaer-Hansen, K. (2019b). Scaling

behaviour in music and cortical dynamics interplay to mediate music listening pleasure. *Scientific reports*, 9(1):1–15.

[Ten Oever and Martin, 2021] Ten Oever, S. and Martin, A. E. (2021). An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *Elife*, 10:e68066.

[Tierney and Kraus, 2015] Tierney, A. and Kraus, N. (2015). Neural entrainment to the rhythmic structure of music. *Journal of Cognitive Neuroscience*, 27(2):400–408.

[Tseng and Lee, 2004] Tseng, C.-y. and Lee, Y.-l. (2004). Speech rate and prosody units: Evidence of interaction from mandarin chinese. In *Speech Prosody 2004, International Conference*.

[Tucker, 1993] Tucker, D. M. (1993). Spatial sampling of head electrical fields: the geodesic sensor net. *Electroencephalography and clinical neurophysiology*, 87(3):154–163.

[VanDam et al., 2016] VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., and MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language*, volume 37, pages 128–142. Thieme Medical Publishers.

[Vélez, 2021] Vélez, J. I. (2021). Machine learning based psychology: Advocating for a data-driven approach. *International Journal of Psychological Research*, 14(1):6–11.

[Vinay et al., 2021] Vinay, A., Lerch, A., and Leslie, G. (2021). Mind the beat: detecting audio onsets from eeg recordings of music listening. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 231–235. IEEE.

[Voss and Clarke, 1975] Voss, R. F. and Clarke, J. (1975). '1/fnoise'in music and speech. *Nature*, 258(5533):317–318.

[Wang et al., 2019] Wang, J., Zhang, Y., Ma, Q., Huang, H., and Hong, X. (2019). Deep learning for single-channel eeg signals sleep stage scoring based on frequency domain representation. In *International Conference on Health Information Science*, pages 121–133. Springer.

[Watts, 1999] Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, 105(2):493–527.

[Wei et al., 2022] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

[West et al., 2008] West, B. J., Geneston, E. L., and Grigolini, P. (2008). Maximizing information exchange between complex networks. *Physics Reports*, 468(1-3):1–99.

[Yan et al., 2020] Yan, Y., Huang, C., Wang, Q., and Hu, B. (2020). Data mining of customer choice behavior in internet of things within relationship network. *International Journal of Information Management*, 50:566–574.

[Yu et al., 2018] Yu, Y., Beuret, S., Zeng, D., and Oyama, K. (2018). Deep learning of human perception in audio event classification. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 188–189. IEEE.

[Zatorre et al., 2002] Zatorre, R. J., Belin, P., and Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, 6(1):37–46.

[Ze et al., 2013] Ze, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE.

[Zipf, 1945] Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of general psychology*, 33(2):251–256.

[Zipf, 1949] Zipf, G. K. (1949). Human behavior and the principle of least effort: an introd. to human ecology.