Université de Montréal

Characterizing and comparing acoustic representations in convolutional neural networks and the human auditory system

par

Jessica A. F. Thompson

Département de Psychologie
Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Sciences Cognitives et Neuropsychologie

le 30 avril, 2020

© Jessica A. F. Thompson, 2020

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée

Characterizing and comparing acoustic representations in convolutional neural networks and the human auditory system

présentée par

Jessica A. F. Thompson

a été évaluée par un jury composé des personnes suivantes :

Isabelle Peretz
(président-rapporteur)

Marc Schönwiesner
(directeur de recherche)

Yoshua Bengio
(codirecteur)

Karim Jerbi
(membre du jury)

Marcel van Gerven
(examinateur externe)

Simone Falk
(représentant du doyen de la FESP)

Résumé

Le traitement auditif dans le cerveau humain et dans les systèmes informatiques consiste en une cascade de transformations représentationnelles qui extraient et réorganisent les informations pertinentes pour permettre l'exécution des tâches. Cette thèse s'intéresse à la nature des représentations acoustiques et aux principes de conception et d'apprentissage qui soutiennent leur développement. Les objectifs scientifiques sont de caractériser et de comparer les représentations auditives dans les réseaux de neurones convolutionnels profonds (CNN) et la voie auditive humaine. Ce travail soulève plusieurs questions méta-scientifiques sur la nature du progrès scientifique, qui sont également considérées.

L'introduction passe en revue les connaissances actuelles sur la voie auditive des mammifères et présente les concepts pertinents de l'apprentissage profond. Le premier article soutient que les questions philosophiques les plus pressantes à l'intersection de l'intelligence artificielle et biologique concernent finalement la définition des phénomènes à expliquer et ce qui constitue des explications valables de tels phénomènes. Je surligne les théories pertinentes de l'explication scientifique que j'espére fourniront un échafaudage pour de futures discussions. L'article 2 teste un modèle populaire de cortex auditif basé sur des modulations spectro-temporelles. Nous constatons qu'un modèle linéaire entraîné uniquement sur les réponses BOLD aux ondulations dynamiques simples (contenant seulement une fréquence fondamentale, un taux de modulation temporelle et une échelle spectrale) peut se généraliser pour prédire les réponses aux mélanges de deux ondulations dynamiques. Le troisième article caractérise la spécificité linguistique des couches CNN et explore l'effet de l'entraînement figé et des poids aléatoires. Nous avons observé trois régions distinctes de transférabilité : (1) les deux premières couches étaient entièrement transférables, (2) les couches 2 à 8 étaient également hautement transférables, mais nous avons trouvé évidence de spécificité de la langue, (3) les couches suivantes entièrement connectées étaient plus spécifiques à la langue mais pouvaient être adaptées sur la langue cible. Dans l'article 4, nous utilisons l'analyse de similarité pour constater que la performance supérieure de l'entraînement figé obtenues à l'article 3 peuvent être attribuées aux différences de représentation dans l'avant-dernière couche : la deuxième couche entièrement connectée. Nous analysons également les réseaux aléatoires de l'article 3, dont nous concluons que la forme représentationnelle est doublement contrainte par l'architecture et la forme de l'entrée et de la cible. Pour tester si les CNN acoustiques apprennent une hiérarchie de représentation similaire à celle du système auditif humain, le cinquième article compare l'activité des réseaux «freeze trained» de l'article 3 à l'activité IRMf 7T dans l'ensemble du système auditif humain. Nous ne trouvons aucune évidence d'une hiérarchie de représentation partagée et constatons plutôt que tous nos régions auditifs étaient les plus similaires à la première couche entièrement connectée. Enfin, le chapitre de discussion passe en revue les mérites et les limites d'une approche d'apprentissage profond aux neurosciences dans un cadre de comparaison de modèles.

Ensemble, ces travaux contribuent à l'entreprise naissante de modélisation du système auditif avec des réseaux de neurones et constituent un petit pas vers une science unifiée de l'intelligence qui étudie les phénomènes qui se manifestent dans l'intelligence biologique et artificielle.

mots clés : apprentissage profond, audition, neurosciences computationnelles, IRMf, parole, analyse de similarité

Abstract

Auditory processing in the human brain and in contemporary machine hearing systems consists of a cascade of representational transformations that extract and reorganize relevant information to enable task performance. This thesis is concerned with the nature of acoustic representations and the network design and learning principles that support their development. The primary scientific goals are to characterize and compare auditory representations in deep convolutional neural networks (CNNs) and the human auditory pathway. This work prompts several meta-scientific questions about the nature of scientific progress, which are also considered.

The introduction reviews what is currently known about the mammalian auditory pathway and introduces the relevant concepts in deep learning. The first article argues that the most pressing philosophical questions at the intersection of artificial and biological intelligence are ultimately concerned with defining the phenomena to be explained and with what constitute valid explanations of such phenomena. I highlight relevant theories of scientific explanation which we hope will provide scaffolding for future discussion. Article 2 tests a popular model of auditory cortex based on frequency-specific spectrotemporal modulations. We find that a linear model trained only on BOLD responses to simple dynamic ripples (containing only one fundamental frequency, temporal modulation rate, and spectral scale) can generalize to predict responses to mixtures of two dynamic ripples. Both the third and fourth article investigate how CNN representations are affected by various aspects of training. The third article characterizes the language specificity of CNN layers and explores the effect of freeze training and random weights. We observed three distinct regions of transferability: (1) the first two layers were entirely transferable between languages, (2) layers 2–8 were also highly transferable but we found some evidence of language specificity, (3) the subsequent fully connected layers were more language specific but could be successfully finetuned to the target language. In Article 4, we use similarity analysis to find that the superior performance of freeze training achieved in Article 3 can be largely attributed to representational differences in the penultimate layer: the second fully connected layer. We also analyze the random networks from Article 3, from which we conclude that representational form is doubly constrained by architecture and the form of the input and target. To test whether acoustic CNNs learn a similar representational hierarchy as that of the human auditory system, the fifth article presents a similarity analysis to compare the activity of the freeze trained networks from Article 3 to 7T fMRI activity throughout the human auditory system. We find no evidence of a shared representational hierarchy and instead find that all of our auditory regions were most similar to the first fully connected layer. Finally, the discussion chapter reviews the merits and limitations of a deep learning approach to neuroscience in a model comparison framework.

Together, these works contribute to the nascent enterprise of modeling the auditory system with neural networks and constitute a small step towards a unified science of intelligence that studies the phenomena that are exhibited in both biological and artificial intelligence.

keywords: deep learning, audition, computational neuroscience, fMRI, speech, similarity analysis

Table des matières

Liste des tableaux

Table des figures

Glossary of Terms and Acronyms

AC: auditory cortex

AN: auditory nerve

ANN: artificial neural network

AVCN: antereoventral cochlear nucleus

BOLD: blood oxygenation level-dependent

characteristic frequency: frequency that has the lowest threshold intensity to elicit neural response in a specific neuron or voxel

CN: cochlear nucleus

CNN: convolutional neural network

DBN: deep belief network

DCN: dorsal cochlear nucleus

DL: deep learning

efferent: describes neural signals or connections from the central nervous system towards the sensory or motor system (i.e. top-down, feedback)

fMRI: functional magnetic resonance imaging

HMO: hierarchical modular optimization

HRF: hemodynamic response function

IC: inferior colliculus

ICC: central nucleus of the inferior colliculus

ICX: external nucleus of the inferior colliculus

ILD: interaural level difference

IT: inferior temporal cortex

ITD: interaural time difference

linearly separable: patterns that lie on opposite sides of a hyperplane

machine learning: a field of statistics and computer science concerned with learning from data/examples

MGB: medial geniculate body

MLP: multilayer perceptron

MTF: modulation transfer function

NE: neural encoding

PCA: principle component analysis

PVCN: posteroventral cochlear nucleus

RF: receptive field

RL: representation learning

RSA: representational similarity analysis

SC: superior colliculus

SI: system identification

SOC: superior olivary complex

STRF: spectro-temporal receptive field

supervised: describes learning algorithms that use ground truth labels to

SVM: support vector machines

tonotopy: The topographic mapping of frequency information within the auditory system

tuning curve: Threshold intensity across frequency required to stimulate a neuron/voxel above sponanteous firing rate/activity.

V1: primary visual cortex

V2: secondary visual cortex

V4: fourth visual area

VCN: ventral cochlear nucleus

voxel: three-dimensional volume pixel, typically measured with magnetic resonance imaging

Remerciements

I would like to thank the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council, les Fonds de Recherche du Québec–Nature et Technologies, the Erasmus Mundus Student Exchange Network in Auditory Cognitive Neuroscience, Mitacs, and Nuance Communications for their funding during my PhD. I have been lucky to have the expert guidance of my supervisors Prof Marc Schönwiesner and Prof Yoshua Bengio as well as my collaborators Prof Elia Formisano, Prof Federico De Martino, and Dr Daniel Willett, who made this work possible. I am grateful for my colleagues at the International Laboratory for Brain Music and Sound Research (BRAMS), especially Prof Robert Zatorre, Prof Virginia Penhune, Dr Megha Sharda, Prof Dominique Vuvan, Dr Pauline Tranchant and Dr Sebastian Paquette for their mentorship and camaraderie, and Johanne David, who made things happen. I am grateful for my colleagues at the Quebec Artificial Intelligence Institute (Mila), especially those who contributed to the Neuroscience and Biology and Neural-AI reading groups, which helped me to find a community of like-minded researchers. I am grateful for my colleagues in Acoustic Research at Nuance, especially Dr Shou-Chun Yin and Dr Kinfe Mengistu, who helped me to navigate Nuance's inhouse tools and datasets. I am grateful to my colleagues at the Department of Cognitive Neuroscience at Maastricht University, especially Dr Valentin Kemper, Dr Faruk Gülban, Dr Niels Disbergens, and Dr Roy Haast, who assisted me with data collection, and Prof Giancarlo Valente, who helped with data analysis. I would also like to thank the participants who suffered through my long fMRI sessions and listened to my strange stimuli. I am grateful for the Montreal Artificial Intelligence and Neuroscience (MAIN) community and conference and everyone who has been helping to make Montreal a hub for research at the intersection of AI and neuroscience, especially Prof Karim Jerbi and Prof Pierre Bellec, who have been dedicated to training and supporting the next generation of neuro-AI researchers. I would also like to thank Elizabeth Dupre who has been endlessly supportive, from recommending open-source analysis pipelines to sending me inspirational quotes. My favorite was "A word after a word after a word is power." from Margaret Atwood. Lastly, I would like to thank all my friends and family, especially my parents, who always encouraged me, and Tomas Matthews, who makes me excited for life everyday, even days when I need to write my dissertation.

Introduction

0.1. Motivation

Sound waves enter the ear and actuates the nervous system via vibrations of the ear drum. From there, a cascade of sensory processing is needed to extract behaviourally relevant information from the sensory information collected at the periphery. Similarly, in machine hearing, microphones record auditory vibrations and computational models must extract and transform the task-relevant information. One goal of auditory computational neuroscience is to characterize the neural computations underlying human auditory perception such that they can be implemented in machine hearing systems. Currently, the best performing machine hearing systems use task-optimized deep neural networks (DNNs), which also consist of a cascade of sensory processing stages, inspired by neural information processing (Yu and Deng, 2015; Hinton et al., 2012; Deng et al., 2013). Machine hearing algorithms don't necessarily need to mimic the human brain, but they are ultimately trying to accomplish the same tasks that humans perform, and so will be subject to the same task constraints and fundamental limits of sensory acuity. In both machine and animal hearing, it may be useful to consider sensory computation as process of 'untangling' the relevant factors of variation (DiCarlo and Cox, 2007). A shared goal of computational neuroscience and deep learning is to understand how and under what conditions this untangling occurs. Machine hearing systems provide a window into how such untangling *can* occur, not necessary how it does occur in animals. However, if we assume that there exist some fundamental principles that govern sensory processing in biological and artificial systems, the study of machine hearing systems may identify candidates of such principles. This presents an opportunity for a virtuous cycle wherein the study of artificial systems can inform and generate hypotheses for neuroscience and the study of brains can inform the development of new artificial systems. The characterization and comparison of artificial and biological neural pathways has the potential to yield insights into the abstract computations underlying human auditory perception.

0.2. The mammalian auditory system

Animal research in the last century has made significant progress in describing the physiology of several structures along the auditory pathway and how these structures represent and transform behaviourally relevant acoustic information.

0.2.1. Auditory representation in the brainstem, midbrain and thalamus

When sound pressure waves enter the ear, they vibrate the tympanic membrane which in turn passes the vibrations to the cochlea via the ossicles. The cochlea amplifies and converts these vibrations into neural signals. The cochlea consists of two liquid-filled chambers, the basilar membrane and the tectorial membrane. The basilar membrane acts as a mechanical frequency analyzer to decompose complex sounds into frequency components. This systematic representation of sound frequency along the length of the cochlea is referred to as tonotopy (Hall, 2008).

Two types of cells lie along the basilar membrane: inner hair cells and outer hair cells. The inner hair cells are the actual sensory receptors and their projections make up 95% of the fibers in the auditory nerve (AN). The outer hair cells receive most of their input from efferent axons of cells in the superior olivary complex (SOC). The outer hair cells are thought to sharpen the frequency resolution of the cochlea by actively changing the stiffness of the tectorial membrane at particular locations (Hall, 2008).

The AN carries sensory information from the inner hair cells to the cochlear nucleus (CN). The response time of this transduction mechanism is so fast that frequencies up to 3kHz in humans can be represented in a one-to-one fashion, meaning that action potentials will be generated at the same rate as the incoming sound pressure waves. This ability of hair cells to follow the waveform of low-frequency sounds results in *phase locking*. This temporal information from the two ears is crucial for the evaluation of interaural time differences, which is one of the primary cues used for sound localization and the perception of auditory space. However, this temporal coding or *volley theory* of auditory information transfer cannot account for the perception of spectral information above 3kHz. *Place coding* or *labeled-line coding* refers to an alternative coding mechanism that encodes frequency information by preserving the tonotopy of the cochlea at higher levels in the auditory pathway. A single AN fiber transmits information about only one part of the audible frequency spectrum. Electrophysiology can be used to measure response properties of specific fibers, such as their tuning curve and characteristic frequency. The topographic organization of characteristic frequency is preserved as signals ascend the auditory pathway (Hall, 2008).

The AN innervates the three divisions of the CN: the antereoventral cochlear nucleus (AVCN), the posteroventral cochlear nucleus (PVCN) and the dorsal cochlear nucleus (DCN). The tonotopy from the cochlea is reproduced in each of the three sections of the CN (Hall,

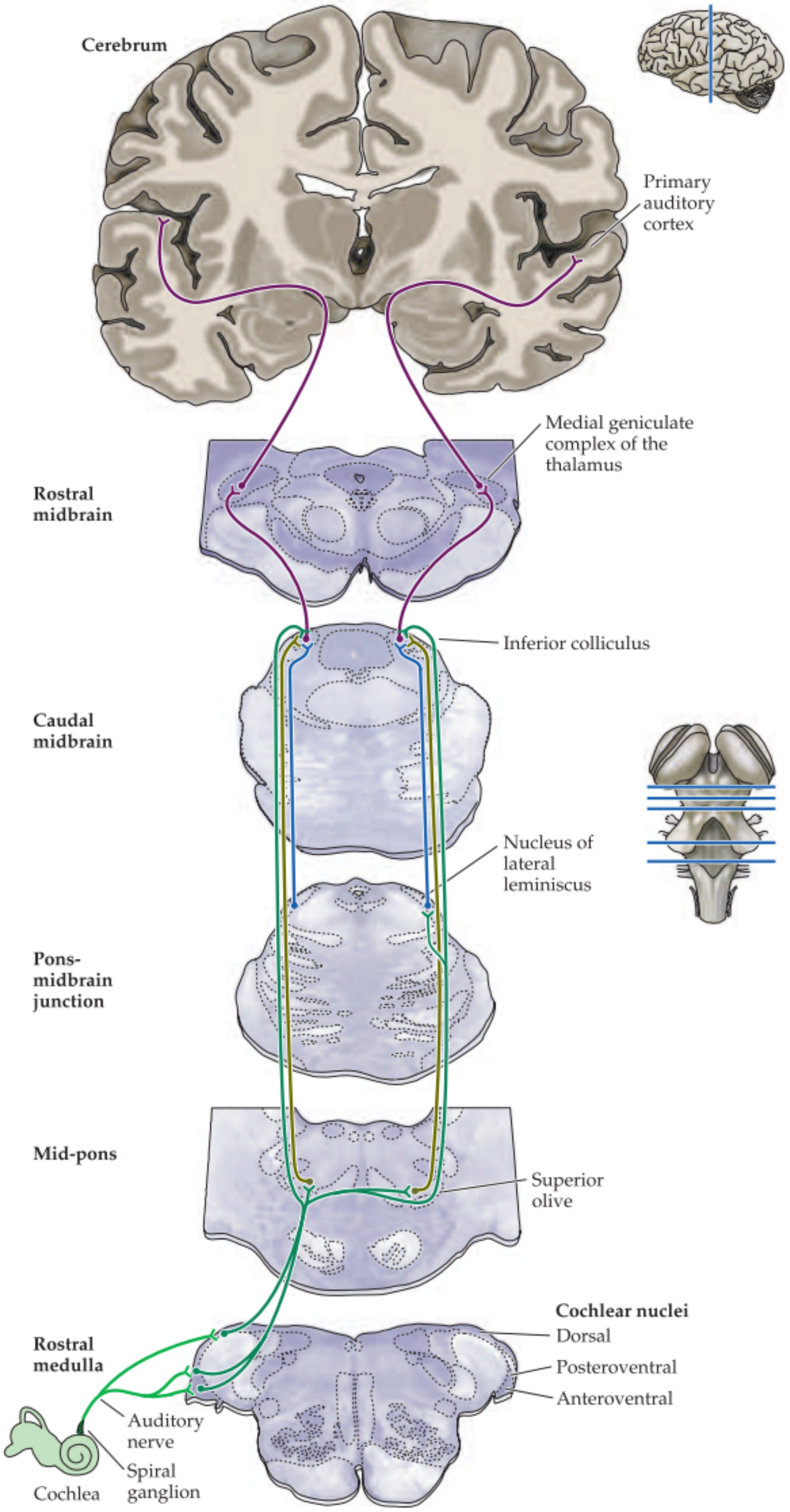


Figure 1. The Auditory Pathway. Afferent pathways from the cochlea up to auditory cortex are shown. Copied from Hall (2008)

2008). The division of the ventral cochlear nucleus (VCN) represent two main pathways to extract and enhance frequency and timing information: the sound localization path (AVCN) and the sound identification path (PVCN) (Young and Oertel, 1999). The AVCN provides input to the SOC, the first structure where information from both ears is combined, where interaural time differences (ITDs) and interaural level differences (ILDs) are mapped for each frequency band separately (Carr, 1993). Broadly tuned cells in the PVCN compute estimates of a level invariant spectral representation of sound (May et al., 1998). Both of these information streams are carried to the central nucleus of the inferior colliculus (ICC) where temporal and spectral information are both topographically but mutually orthogonally mapped (Langner, 1992). Frequency-specific ITD and ILD maps are combined to create a map of sound localization in the external nucleus of the inferior colliculus (ICX). The temporal resolution of these differences is on the order of 10 microseconds. This auditory space map is subsequently aligned with the retinotopic map of visual space and motor map of gaze in the superior colliculus (SC) (Hyde and Knudsen, 2000). For a detailed description of the neural mechanism underlying sound localization, see Hall (2008) and Eggermont (2001).

The medial geniculate body (MGB) in the thalamus is the auditory relay station between the IC and auditory cortex (AC) and receives convergent inputs from the separate spectral and temporal pathways in lower areas. Consequently, the MGB is the first structure where cells are found to respond to specific spectro-temporal patterns. While preserving a tonotopic organization (Imig and Morel, 1985), cells in the MGB are also selective to specific combinations of frequencies and specific time intervals between frequencies (on the order of milliseconds) (Hall, 2008). The STRF of an auditory neuron refers to that cell's preferred pattern in frequency and time. As we ascend the auditory pathway, STRFs become more complex and lose temporal precision (see Figure 2).

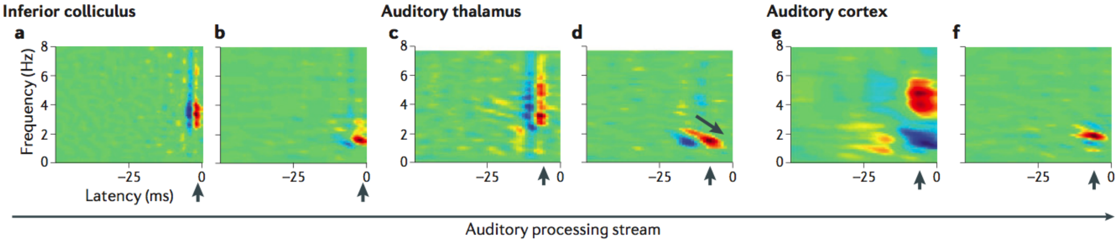


Figure 2. STRFs at different levels of the auditory system. In IC, some neurons have STRFs that are narrowly tuned in time, but broadly tuned in frequency (a) or narrowly tuned in both time and frequency (b). Thalamic neurons have greater latencies than IC neurons (c) and also demonstrate selectivity to patterns in time, e.g. a descending tone sweep (d). STRFs are much slower and more complicated in auditory cortex (e, f). Copied from Theunissen and Elie (2014)

Although midbrain and thalamic auditory areas have been well studied in several animal models, our knowledge about these structures in humans is poor due to the small size of the IC and MGB and the spatial resolution of non-invasive imaging methods. Recently, however, the development of new acquisition sequences for ultra-high field neuroimaging has facilitated the functional mapping of these two areas in humans. De Martino et al. (2014a) found that both the MGB and the IC were tuned to contralateral locations. They also found a single frequency gradient in IC and two frequency gradients in the MGB.

0.2.2. The auditory cortex

In the monkey and several other species, the AC is organized hierarchically into several primary or *core* areas that receive inputs from the thalamus and are surrounded by non-primary or *belt* and *parabelt* regions (Kaas and Hackett, 2000; Rauschecker et al., 1995). While it is generally agreed upon that the human auditory cortex is also arranged hierarchically with information passing from primary to secondary areas, the details of this model break down in the human because of considerable anatomical differences. Currently, there is no standard scheme for the functional parcellation of the human AC (Moerel et al., 2014). Although the precise delineation and functional organization of human auditory cortex remains an open question, new parcellation methods using pattern recognition and ultra-high field neuroimaging have recently been proposed (De Martino et al., 2014b; Moerel et al., 2014; Schönwiesner et al., 2014).

Despite cortical differences across species, several common representational mechanisms have been identified. As in lower auditory areas, neurons in auditory cortex also exhibit frequency selectivity and a tonotopic organization (Humphries et al., 2010). The exact number and orientation of tonotopic gradients varies across species and is still a topic of debate in human research (Moerel et al., 2014). Auditory cortical neurons have preferences for specific spectro-temporal patterns (DeCharms et al., 1998). Physiological and psychoacoustic studies suggest that the cortical representation of sound involves the explicit encoding of spectral and temporal modulations through dedicated modulation-detectors (Viemeister, 1979; Santoro, 2014). Consequently, STRFs in auditory cortex are often parameterized by modulations in both frequency and time. In the visual domain, it has been shown that cortical cells in primary visual cortex (V1) respond selectively to specific patterns of sinusoidal gratings and can be modeled as spatial modulation frequency filters (De Valois et al., 1979). The auditory equivalent of a grating is the *dynamic ripple*, a complex broadband sound with a sinusoidal spectral envelope that drifts along the logarithmic frequency axis over time (Kowalksi et al., 1996). Such stimuli can be used to calculate spectro-temporal modulation transfer functions (MTFs), whose 2-dimensional Fourier transform gives a STRF. Schönwiesner and Zatorre (2009) used dynamic ripples and fMRI to find voxels tuned to combined spectro-temporal modulations in the primary and secondary auditory cortex. The resulting spectro-temporal modulation maps were highly reliable within subjects and highly variable across subjects, highlighting the importance of building personalized models. It has been shown that STRFs in primary auditory areas are highly context-dependent and demonstrate rapid plasticity, meaning that the response properties of an auditory cortical cell can change drastically in response to top-down control signals (Mesgarani et al., 2009; David et al., 2012).

0.2.3. Neural processing of natural sounds

Most of the work summarized up to this point used simple, synthesized stimuli that could be systematically varied. Conversely, much has also been learned from the neuroethological approach, which employs natural and behaviourally relevant sounds as stimuli. Recent work on the neural processing of natural sounds can be thought of as a merger between the neuroethological and classical neurophysiological approaches facilitated by the analysis of the statistical properties of natural sound and the use of machine learning methods to take into account this statistical structure when estimating neural response characteristics.

0.2.4. The statistics of natural sound

Theunissen and Elie define natural sound as "environmental sounds that are not generated by human-made machines, such as the sounds of footsteps, wind, fire and rain; all animal vocalizations, including human speech; and other sounds generated for communication by animals, such as stridulation in crickets, buttress drumming by chimpanzees and instrumental music by humans." (2014, p. 356) Describing the statistics of these types of sounds helps us to understand what makes them special and to determine whether the auditory system evolved to process them optimally. For instance, it has been observed that perceptually relevant physical characteristics of isolated natural sounds follow a power law. More specifically, certain fluctuating physical characteristics of natural sounds follow a $1/f$ relationship, where f is frequency. This relationship does not hold for the sound spectrum itself, but it holds for other slower properties such as loudness or pitch height. This relationship also holds for the sound cepstrum (the power spectrum of the log of the sound spectrum) (Singh and Theunissen, 2003), which is related to the timbre of a sound (Müller and Ewert, 2010). Dependencies have also been observed in the modulation power spectrum (frequencies of temporal and spectral modulations in the spectrogram). For example, many animal vocalizations are dominated by relatively slow sounds with fine harmonic structure (see Figure 3) (Theunissen and Elie, 2014).

These statistical characteristics of natural sounds have several implications. Firstly, the power law relationship means that natural sounds have correlations over multiple timescales. This clearly separates natural sounds from signals that are completely random or uncorrelated, such as white noise, and signals that are dominated by a single correlation time, such as a perfect sine wave. Second, the properties that exhibit the power law relationship also tend to be those that are of perceptual relevance. For example, we are unable to perceive

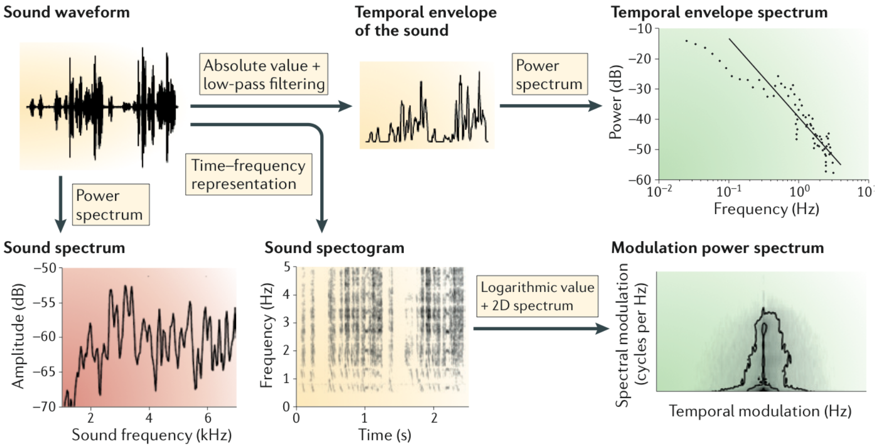


Figure 3. Natural Sound Statistics Various acoustic features extracted from recordings of a zebra finch song. Green graphs indicate common characteristics of natural sound and orange graphs indicate those that are specific to each sound class. Copied from Theunissen and Elie (2014)

the details of a sound pressure waveform, but we perceive attributes such as intensity fluctuations, rhythm, and timbre, whose modulations follow a $1/f$ relationship (Theunissen and Elie, 2014). It has also been shown that sounds with the same statistical properties as natural sound elicit higher information rates compared to synthetic sounds that lack some of these properties (Hsu et al., 2004). Functional neuroimaging studies have also shown that natural sounds elicit broader responses than synthetic stimuli (Moerel et al., 2012). Moerel et al. (2012) suggest that natural sounds may be optimal for studying the functional architecture of higher order auditory areas since they engage auditory neurons in meaningful and behaviourally relevant processing.

Natural sound statistics may also be related to the frequency tuning of mammalian auditory nerve fibers and efficient coding schemes. It has been suggested that the shape of the filters measured at the auditory nerve is optimal for representing the independent components of animal vocalizations and environmental sounds; the lower-frequency narrow-band filters efficiently represent animal vocalizations and the higher-frequency broad-band filter efficiently represent environmental sounds (Theunissen and Elie, 2014). Interestingly, human speech contains both of these components. Human speech, though obviously very behaviourally relevant now, cannot have had any effect on the evolution of vertebrate hearing. However, one can assume that human speech evolved under the constraints of the existing auditory and vocalization systems. It is not surprising then that human speech exhibits properties of both animal vocalizations and environmental sounds such as steady-state harmonically related frequencies, frequency modulations and noise bursts (Eggermont, 2001). In other words, it seems likely that the physical characteristics of speech evolved to be optimally represented in the auditory system (Lewicki, 2002).

0.2.5. Functional neuroimaging of natural sound perception

FMRI is a noninvasive functional neuroimaging technique that provides an indirect measure of brain activity. The FMRI signal is called the blood oxygenation level-dependent (BOLD) signal and is sensitive to local changes in deoxyhemoglobin which has been shown to reflect local changes in brain activity (Ogawa and Tank, 1992). As a noninvasive measure with high spatial resolution, fMRI has been extensively used in human studies of auditory perception and cognition. The traditional goal has been to localize or compare the magnitude of activity in regions of the brain that are responsible for a specific perceptual or cognitive attribute. To do this, experimental conditions are constructed such that their neural responses can be contrasted to isolate task-related activity. Univariate statistical analyses are performed within each voxel and subsequently corrected for multiple comparisons. Individual brains are aligned to a common anatomical template to calculate group-level effects. In auditory neuroscience, this approach has helped to identify and characterize regions that are responsible for the perception of voices (Belin et al., 2000), the comprehension of speech (Rodd et al., 2005), the perception and enjoyment of music (Peretz and Zatorre, 2005) and auditory-motor interaction (Zatorre et al., 2007), among many other auditory capacities. From a computational neuroscience perspective, this approach provides a broad road map for the information processing path involved in complex human auditory perception, but lacks the ability to probe the computational mechanisms at a finer scale. Some may argue that this is a limitation of fMRI itself (Mole and Klein, 2010). However, recent work that combines computational modeling with functional neuroimaging suggests that fMRI can indeed be a useful tool for studying computational mechanisms a finer scale.

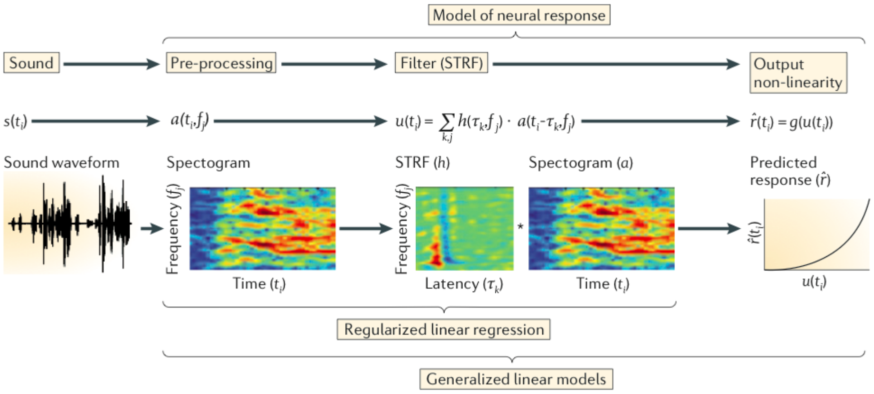


Figure 4. Stimulus-response characterization Stimulus-response functions can be calculated using natural sounds and regularized linear regression. Copied from Theunissen and Elie (2014)

As was already discussed, fMRI can be used to estimate stimulus-response characteristics using synthesized sounds such as pure tones and dynamic ripples. STRFs and MTFs can

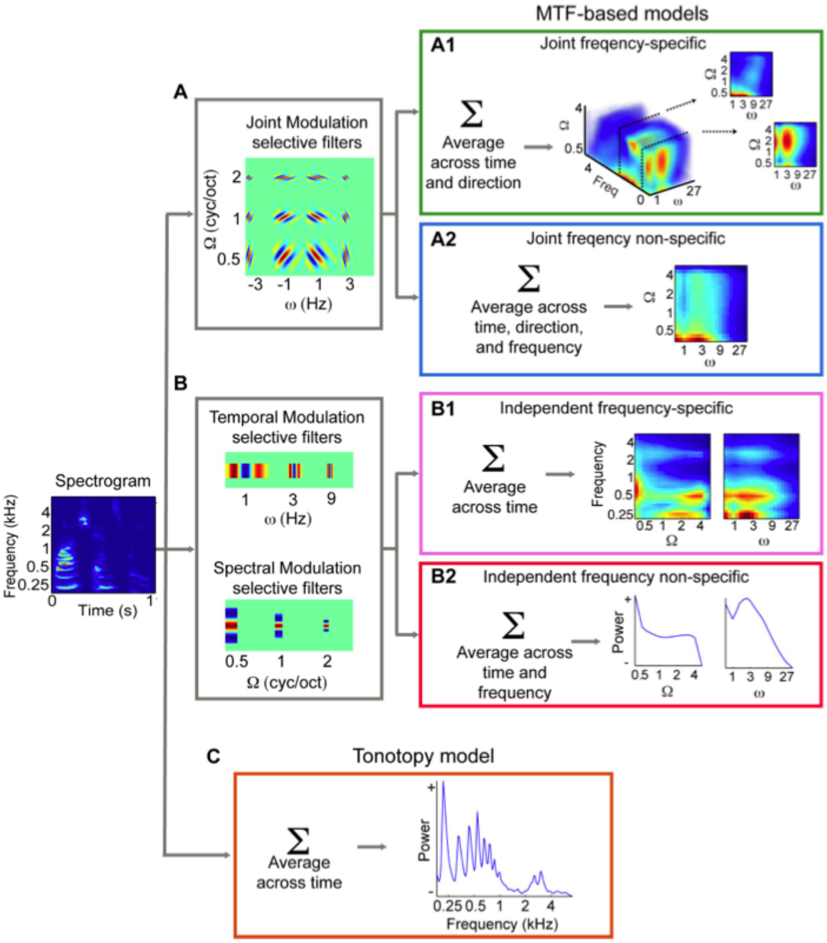


Figure 5. Overview of candidate encoding models tested in Santoro et al. (2014). Features based on simple tonotopy, spectral modulations, temporal modulations, and joint spectro-temporal modulations were extracted from the stimuli. Copied from Santoro et al. (2014)

also be calculated using natural sound stimuli and regularized linear regression, as pictured in Figure 4. An extension of this approach, referred to as system identification (SI) (Wu et al., 2006), neural encoding (NE) (Naselaris et al., 2011) or model-based fMRI (O'Doherty et al., 2007), has been used as a way of evaluating and comparing computational models of sensory information processing in the brain. NE treats the problems of sensory receptive field estimation as a regression problem and aims to build quantitative models that describe how a neuron or voxel will respond to a potential stimulus (Wu et al., 2006). A NE approach involves collecting fMRI responses to a large number and variety of sounds. Acoustic features are then extracted from these stimuli according to a computational model of neural information processing or *encoding model* (see Figure 5 for an example of the types of candidate encoding models that have been tested). As pictured in Figure 6, models are evaluated based on their ability to predict neural activity evoked by natural stimuli. This approach can be used in a hypothesis driven manner, where specific hypotheses about neural information processing are embedded in the design of the computational model. It has been suggested that the NE approach can also be used to generate hypotheses about sensory coding, even in the absence of a prior theoretical or quantitative model (Wu et al., 2006).

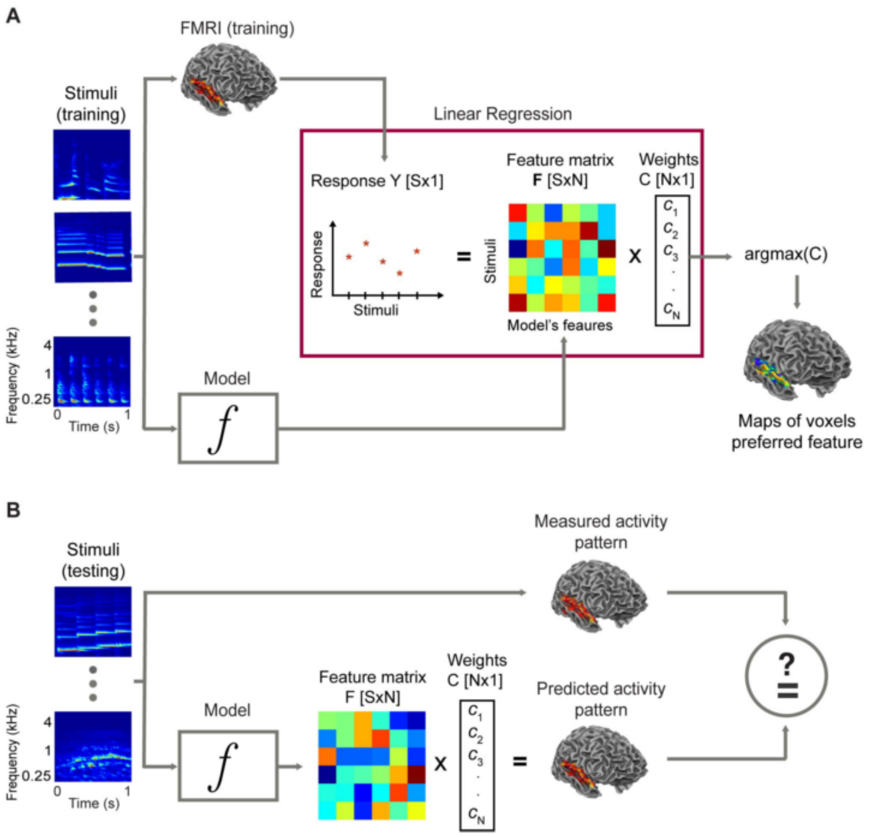


Figure 6. Schematic of model estimation and evaluation in NE. (a) Encoding model parameters for each voxel are estimated from a wide variety of stimulus-response pairs. (b) Model performance is evaluated based on their ability to predict fMRI responses to an independent set of stimuli. Copied from Santoro et al. (2014)

The NE approach was recently used by Santoro et al. (2014) to study role of spectro-temporal modulations in the cortical encoding of natural sounds. They used ultra-high field fMRI (7-Tesla) to measure brain activity while subjects listened to a variety of natural sounds. They found that the cortical encoding of natural sounds involves multiple representations of the sound spectrogram at different degrees of spectral and temporal resolution. Specifically, they found that posterior/dorsal auditory regions prefer coarse spectral information with high temporal precision, while anterior/ventral regions prefer fine-grained spectral information with low temporal precision. The authors suggest that this multi-resolution view of natural sounds may be a crucial computational mechanism for flexible and behaviourally relevant sound processing (Santoro et al., 2014).

0.3. Representation Learning

In the preceding sections, I have summarized the current understanding of the representational transformations performed by the auditory system. In the field of machine learning, the domain of representation learning (RL) refers to a family of methods that learn useful transformations of some raw input signal. When such a RL system involves more than one intermediate representation, this is referred to as deep learning (DL). Sometimes RL systems are used only for feature extraction and another machine learning system will be used to make predictions. This is called *feature learning* (Lee et al., 2009a). Other times, inferences will be made directly by the RL system. Recently, DNNs have significantly outperformed other methods on complex tasks such as speech recognition (Hinton et al., 2012), visual object recognition (Krizhevsky and Hinton, 2012) and natural language processing (Collobert and Weston, 2008), which has led to increased research and development in this area. Much of RL, especially that which evolved from the tradition of artificial neural networks (ANNs), was inspired by early computational neuroscience models of neurons and neural populations. In this regard, RL systems are more biologically plausible than other machine learning systems (Bengio et al., 2013). In the following sections, I will describe several common classes of RL models and training algorithms.

0.3.1. Component analysis

Principle component analysis (PCA) is perhaps the simplest example of representation learning. PCA identifies the axes of maximum variation in a given dataset. PCA learns a linear transformation $h = f(x) = W^T x + b$ of input $x \in \mathbb{R}^{d_x}$, where W is a $d_x \times d_h$ matrix whose columns define the orthogonal directions of greatest variance in the training set. The resulting d_h features (components of representation h) are decorrelated. The data can then be transformed to use these 'principal components' as a basis set. Dimensionality reduction is often performed by dropping the principal components that explain the least variance. This new representation can be more convenient to work with than the original representation.

0.3.2. Sparse coding

One way to investigate neural coding of natural stimuli is to analyze various ways of encoding signals such as natural images and audio. We can consider two broad coding strategies: efficient, compact coding and sparse, distributed coding. PCA-based dimensionality reduction is an example of efficient coding because it represents as much information as possible in as few dimensions as possible. An alternative coding strategy, sparse coding (a.k.a. minimum-entropy coding) is motivated by the fact that natural stimuli tend to have a sparse structure and can be represented with a small number of descriptors out of a large set. A sparse code is a high-dimensional representation with only a small number of non-zero elements.

A very influential paper from Olshausen and Field (1996) showed that a coding strategy that maximizes sparseness is sufficient to account for receptive field (RF) properties of V1 neurons, namely spatial localization, orientation and bandpass. Further evidence for sparse codes in V1 was provided by physiological work showing that responses elicited by natural stimuli are more sparse than those elicited by synthetic stimuli (Vinje and Gallant, 2000). Recent work in the auditory domain also showed that sparse codes for speech predicted spectro-temporal RFs in the IC (Carlson et al., 2012). Sparse coding has many theoretical and practical benefits. From the point of view of efficient resource use, having neurons primarily inactive reduces metabolic consumption. Representations that are distributed across a small number of active neurons is helpful for downstream computation because it forces individual neurons to carry more explicit information and increases the signal-to-noise ratio (Stansbury, 2014).

0.3.3. Perceptron

The perceptron, invented in 1958, was the first algorithmically described neural network (Rosenblatt, 1958; Haykin, 2009). The perceptron is the simplest ANN that can classify linearly separable patterns. Understanding this simplest form will help us to understand the more complicated deep networks described in subsequent sections. The perceptron is a model of a single neuron consisting of a linear combiner v and nonlinear hard limiter activation function, diagrammed in Figure 7. Inputs $x_1 \ldots x_m$ are linearly weighted by weights $w_1 \ldots w_m$ and bias b. The hard limiter input of the neuron is

$$v = \sum_{i=1}^{m} w_i x_i + b \tag{0.3.1}$$

The hard limiter applies a signum function to produce an output $y(x)$ that is -1 if the hard limited input is negative and 1 if it is positive. The perceptron parameters (weights and bias) are adapted iteratively according to a supervised error-correction learning algorithm. On each iteration, the weights and bias are adjusted with the following learning rule

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta[d(n) - y(n)]\mathbf{x}(n) \tag{0.3.2}$$

where the vector $\mathbf{w}$ includes both the weights and the bias term, n indexes the iteration, $d(n)$ is the desired output, $y(n)$ is the output produced by the perceptron and η is the learning rate, which controls how much the parameters will change on each iteration. In the case of two linearly separable classes, this algorithm has been proven to converge to a solution. See Haykin (2009) for the full perceptron convergence theorem. The perceptron can be extended to multi-class scenarios by adding additional neurons (Haykin, 2009).

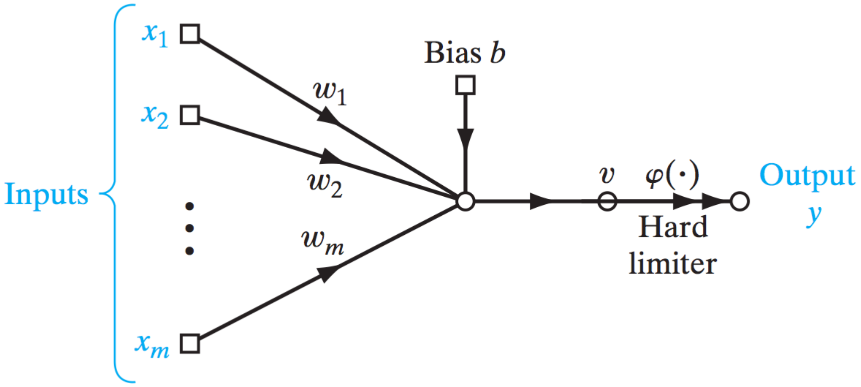


Figure 7. Signal flow graph of the perceptron. Inputs $x_1 \ldots x_m$ are linearly weighted by weights $w_1 \ldots w_m$ and bias b. The sum of these weighted inputs (linear combiner v) is passed through a signum function (hard limiter) to produce an output that is -1 if the hard limited input is negative and 1 if it is positive. Copied from Haykin (2009).

0.3.4. Feed forward neural network/multilayer perceptron

The term MLP refers to a neural network with one or more hidden layers. In this case, the use of the word "perceptron" is misleading since such a network is not actually a perceptron, but the term MLP persists none the less. An MLP, shown in Figure 8, consists of layers of neurons (or *units*) that include a nonlinear activation function that is typically differentiable. Common activation functions include the logistic and the hyperbolic tangent functions. One or more of these layers are *hidden* from both the input and output, meaning that they are unknown, latent representations of the input. The hidden layers act as feature detectors and allow for the network to solve nonlinear classification problems, unlike the perceptron (Haykin, 2009).

An MLP can be trained using the back propagation algorithm, which consists of forward phase and a backwards phase. In the forward phase, the synaptic weights are fixed and the input signal is passed through the network producing an output signal. In the backward phase, an error signal is calculated by comparing the output produced in the forward phase to the desired output. This error signal is propagated back through the network, making adjustments to the synaptic weights as it goes. As the learning progresses, the hidden layers will discover useful representations of the training data, i.e. the hidden layers perform a nonlinear transformation on the input data into a new *feature space*. Specifically, the induced local field or *preactivation function* $v_j(n)$ of hidden unit j at iteration n can be written as

$$v_j = \sum_{i=0}^{m} w_{ij}(n) y_i(n) \tag{0.3.3}$$

where i indexes the m units in the previous layer that are connected to unit j and $y_i(n)$ is the output of unit i on iteration n. We can define an error signal $e_k(n)$ produced at the

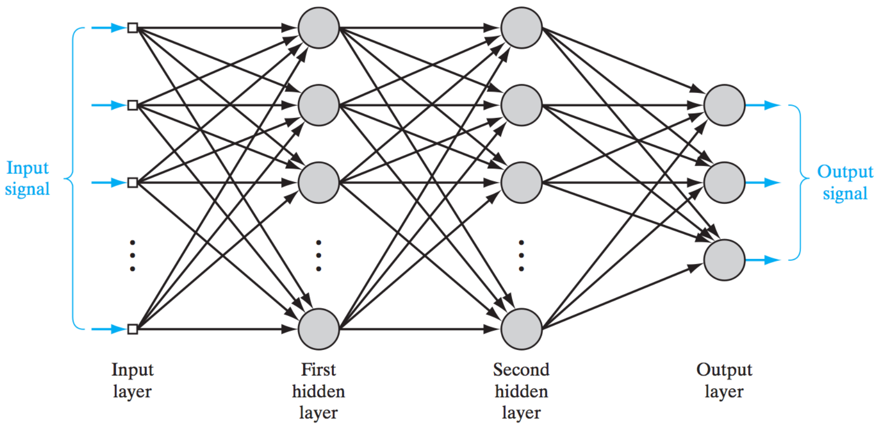


Figure 8. Architectural graph of a MLP with two hidden layers The input signal passes through two fully connected hidden layers (all units in one layer are connected to all units in the subsequent layer) before reaching the output units. Each hidden layer represents a latent representation of the input signal. Copied from Haykin (2009).

output of unit k as

$$e_k(n) = d_k(n) - y_k(n) \tag{0.3.4}$$

where $d_k(n)$ is the kth element of the desired output vector $\mathbf{d}(n)$. For the purposes of the backpropagation algorithm, we will be concerned with the *instantaneous error energy* of neuron k, defined by

$$\xi_k(n) = \frac{1}{2}e_k^2(n) \tag{0.3.5}$$

and the *total instantaneous error energy* of the whole network, defined as

$$\xi(n) = \sum_{k \in C} \xi_k(n) \tag{0.3.6}$$

$$= \frac{1}{2}\sum_{k \in C} e_k^2(n) \tag{0.3.7}$$

where the set C includes all output units. In the case of online learning, where the network parameters are updated after every training example, the cost function to be minimized will be $\xi(n)$. In the batch learning scenario, we will minimize the average instantaneous error energy or *empirical risk* $\xi_{av}(N)$, averaged over all training examples in batch N. In the online case, on each iteration n, the correction $\Delta w_{ji}(n)$ is applied to $w_{ji}(n)$ where

$$\Delta w_{ji}(n) = -\eta \frac{\partial \xi(n)}{\partial w_{ji}(n)} \tag{0.3.8}$$

where η is the learning rate of the back propagation algorithm. This amounts to gradient descent in weight space, i.e. seeking changes in the parameterization of the network that reduces $\xi(n)$ (Haykin, 2009).

0.3.5. Convolutional neural network

A CNN is a special class of MLP designed to be invariant to certain types of information. For example, in the context of object recognition from natural images, one may wish to recognize two-dimensional shapes regardless of their spatial location, scale, orientation or other forms of distortion within an image. The CNN uses several structural constraints to accomplish this. Firstly, each neuron receives inputs from only a local receptive field in the previous layer. Second, each convolutional layer of the network is composed of several *feature maps*, within which neurons are constrained to share their input weights. Each feature is convolved with the input. This has the effect that a feature will be detected regardless of where it occurs in the input signal. This property is called *shift invariance*. Weight sharing also reduces the number of free parameters, which makes the network easier to train. Third, each convolutional layer is followed by a computational layer that performs local averaging or pooling and subsampling. This reduces the resolution of the feature map and the sensitivity to shifts and other distortions (Haykin, 2009).

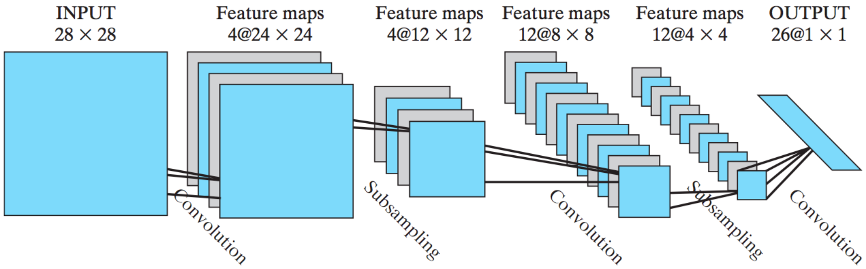


Figure 9. Convolutional network for image processing Layers alternate between convolutional layers, whose units extract features from local receptive fields in the layer below, and feature pooling or subsampling layer that combine or sample features extracted in the previous convolutional layer. Copied from Haykin (2009).

0.4. Modeling neural representations using statistical features of natural images

Given the recent successes and biological plausibility of DNNs as well as the increased use of machine learning methods for neuroimage analysis, it has been suggested that DL could be useful for neuroscience research (Hinton, 2011). One domain where DL methods have enjoyed considerable success is in visual object recognition, a task that is also very familiar to visual neuroscientists. The primate visual system is extremely good at determining the category of a visually presented object. Neurophysiological studies have shown that this ability is mediated by the IT, where neurons respond selectively to high-level visual object categories (Hung et al., 2005). How the brain arrives at this categorical representation is not well understood and presents a prime opportunity for use of DL methods in neuroscience research.

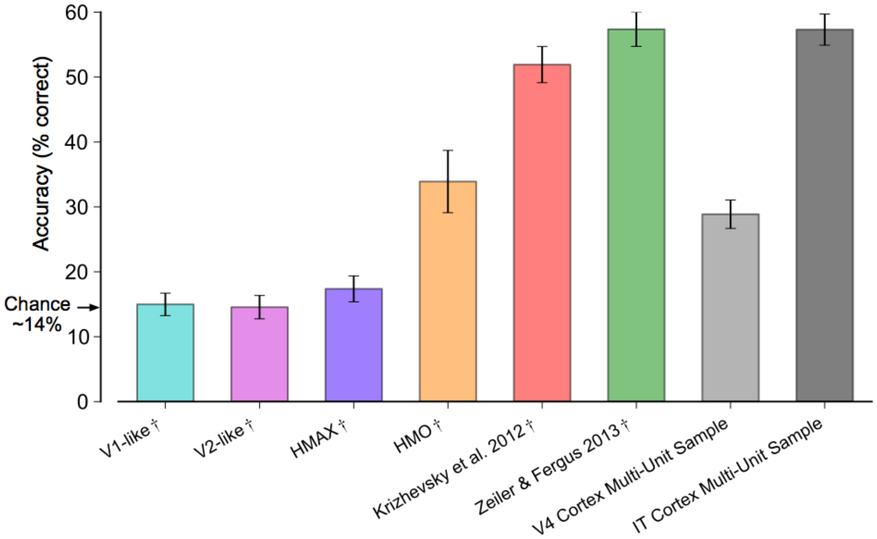


Figure 10. Linear-SVM generalization performance of neural and model representations. The performance of an SVM trained on representations produced by the Zeiler and Fergus (2014) CNN matches that of multi-unit recordings from IT. Copied from Cadieu et al. (2014).

Cadieu et al. (2014) recently reported a series of experiments designed to compare the performance and self similarity structure of representations learned by several DNNs to that of brain activity recorded in the IT of macaque monkeys during visual stimulation. The authors compared the neural recordings to three CNNs and several other biologically relevant representations on an 8-class visual object classification task. Models of V1, secondary visual cortex (V2) and the ventral visual stream, respectively, performed at or near chance when an support vector machines (SVM) was trained to predict visual object categories from the representations they produced. Figure 10 shows the accuracy achieved by the different models as well as by measurements made from V4 and IT. The IT measurements achieves high generalization accuracy and is only matched in performance by the Zeiler and Fergus representation. The authors also compared the models on their representational geometry—the structure of pairwise distances between each stimulus—presented as a dissimilarity matrix in Figure 11. This type of analysis is called a representational similarity analysis (RSA) (Kriegeskorte et al., 2008). While the simpler models produced representations whose similarity structure was not at all correlated with the neural measurements, the CNNs produced representations whose similarity structure was significantly correlated with that of IT neurons. They also found that the intermediate layers of the hierarchical modular optimization (HMO) model were highly predictive of neural responses in V4 (Yamins et al., 2014). Their

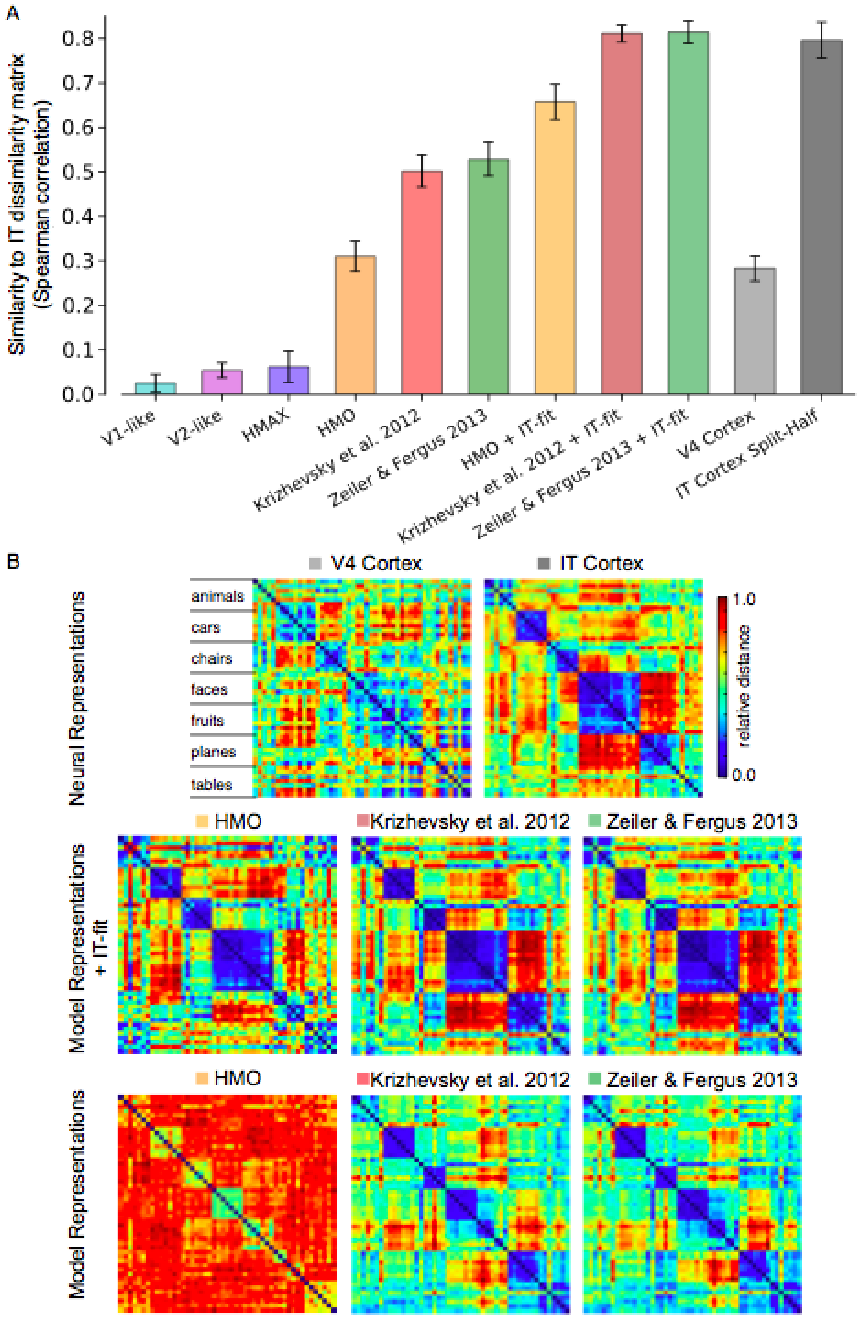


Figure 11. Object-level representational similarity analysis comparing model representations and neural data. Copied from Cadieu et al. (2014). A) The dissimilarity structure of the representations produced by the two state-of-the-art CNNs was highly correlated to the dissimilarity structure of recordings in IT. B) Dissimilarity matrices are shown for several models and neural recordings from IT and V4.

results show that performance-optimized hierarchical neural network models can learn representations that are similar to those found in higher level visual cortex (Cadieu et al., 2014).

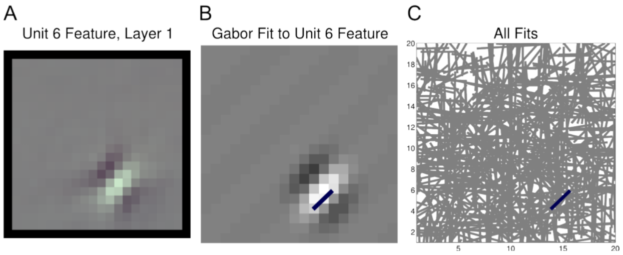


Figure 12. Fitting Gabor functions to 1st layer features. (A) The receptive field of a particular unit in Layer 1. (B) A Gabor function, summarized by a black line, was fit to the receptive field in A. (C) The receptive fields of a population of first layer units can be summarized by plotting the individual lines corresponding to their best-fitting Gabor filters. Copied from Stansbury (2014).

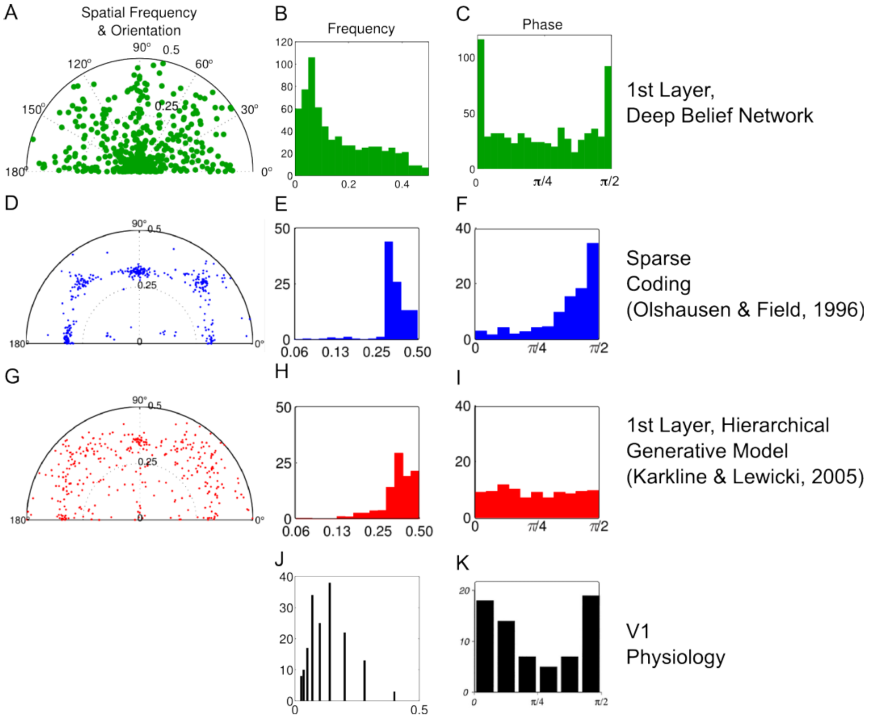


Figure 13. Comparing the properties of 1st layer features to other theoretical and physiological results. The distributions of frequency and phase of V1 receptive fields are better mimicked by the first layer of a deep belief network than by other models that learn similar features. Copied from Stansbury (2014).

Related work by Stansbury (2014) described a similar analysis with unsupervised DL. Stansbury presents a new visualization technique based on fitting Gabor filters to first layer features. Each first layer unit is fit to a Gabor function which is then summarized by a line. In this way, many first layer features can be plotted in the same image, as shown in Figure 12. Stansbury fit Gabor filters to the first layer units of a deep belief network (DBN) trained on natural scenes as well as a sparse coding model (Olshausen and Field, 1996), another hierarchical generative model (Karklin and Lewicki, 2005) and the receptive fields of V1 neurons. He compared the statistical properties of these Gabor filters, namely their spatial frequency, orientation, and phase. He found that the first layer features exhibited frequency and phase distributions that more closely resembled those of V1 neurons than sparse coding or the Karklin and Lewicki model (Figure 13). After extensive analysis and visualization of the representations learned at the three levels of the DBN, Stansbury built neural encoding models using the representations learned at each layer of the network. He found that features learned in the lowest layer of the DBN accurately characterize the responses of V1 cells, but not V2 cells. He also found that the representations learned in the higher levels of the DBN provide a better characterization of V2 neural responses.

The work by Cadieu et al. (2014) and Stansbury (2014) can be seen as an extension of the work on the statistics of natural sound and NE approach to neural signal analysis. Instead of characterizing the statistical properties of natural stimuli, Stansbury characterizes statistical features learned from such signals with DNNs. Both works demonstrate that DNNs learn representations that are more similar to patterns of activity in visual cortex than other models. However, these experiments differ from the classic NE in the way that hypotheses about neural coding are tested. For example, in the NE experiments reported in Santoro et al. (2014), for each computational model tested, one knows exactly what stimulus information is being encoded (e.g. joint spectro-temporal modulations). In the case of Stansbury (2014) and (Yamins et al., 2014), where they use features extracted at the 2nd or 3rd layer of a deep network, it is less clear exactly what information is being used in the linearized encoding model (although we can continue to work toward such an understanding with the types of analyses and visualizations presented in Stansbury (2014)). Instead, what is known is what computational architecture was used to extract these features and how such architectures were trained. In these experiments, hypotheses about neural information processing are embedded in the design of the computational architecture and training algorithms used to learn the features, rather than in the design of the features themselves. In the subsequent chapters of this dissertation, similar approaches are adapted to the auditory domain to understand how acoustic representations are influenced by architecture, task and training and to evaluate the correspondence of such learned representations to the human auditory system.

0.5. Organization of the Thesis

This thesis has a number of scientific and meta-scientific goals. The scientific goals are (1) to characterize intermediate representations in the human auditory system, (2) to characterize intermediate representations in DNNs, and (3) to compare representations in DNNs and the human auditory system. The interpretation of DNN-to-brain comparison is thus supplemented by the accompanying characterization of the network representations. The meta-scientific goals are to unify computational neuroscience and deep learning science and to address the methodological and philosophical issues that interfere with this unification.

The first scientific goal is addressed in the second and fifth articles of this thesis. Article 2 asks how well can cortical responses to sound be modeled as a linear function of spectrotemporal modulations. The fifth article provides a characterization of which DNN layers are most similar to ROIs throughout the human auditory pathway. The second scientific goal is addressed by the third and fourth articles. The third article characterizes the language specificity of DNN layers and both the third and fourth articles investigate how DNN representations change when modifying various aspects of networks training. The third scientific goal is addressed by the fifth article, which investigates whether representations in deep acoustic models learn a similar hierarchical structure as in the human auditory system.

The meta-scientific goals are primarily addressed in the first article which asks how would a unified science of intelligence (combining computational neuroscience and artificial intelligence) progress? What would be the form of scientific explanations of phenomena at this intersection? Article 1 and 5 both consider the merit of using DNNs as models of sensory processing. The meta-scientific question of how best to compare neural representations, be they in DNNs or in animal brains, is considered in the forth and fifth articles. Together, these works contribute to the nascent enterprise of modeling the auditory system with DNNs and constitute a small step towards a unified science of intelligence that studies the phenomena that are common to biological and artificial intelligence.

Bibliography

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403:309–312.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Cadieu, C., Hong, H., and Yamins, D. L. K. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12):e1003963.

Carlson, N. L., Ming, V. L., and Deweese, M. R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS computational biology*, 8(7):e1002594.

Carr, C. E. (1993). Processing of temporal information in the brain. *Annual review of neuroscience*, 16:223–43.

Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki.

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6):2144–9.

De Martino, F., Moerel, M., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E. (2014a). Spatial organization of frequency preference and selectivity in the human inferior colliculus. In *International Society for Magnetic Resonance in Medicine*, Milan.

De Martino, F., Moerel, M., Xu, J., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E. (2014b). High-Resolution Mapping of Myeloarchitecture In Vivo: Localization of Auditory Areas in the Human Brain. *Cerebral Cortex*.

De Valois, K. K., De Valois, R. L., and Yund, E. W. (1979). Responses of striate cortex cells to grating and checkerboard patterns. *The Journal of Physiology*, 291:483–505.

DeCharms, R. C., Blake, D. T., and Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing research*, 157(1-2):1–42.

Hall, W. C. (2008). The Auditory System. In Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W., Lamantia, A.-S., McNamara, J., and White, L., editors, *Neuroscience*, pages 313–342. Sinauer Associates, Sunderland, MA, 4th edition.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson Prentice Hall, Upper Saddle River, New Jersey, third edition.

Hinton, G. E. (2011). Machine learning for neuroscience. *Neural systems & circuits*, 1(1):12.

Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Vanhoucke, V., Nguyen, P., Sainath, T. N., Kingsbury, B., and Senior, A. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, (November):1–27.

Hsu, A., Woolley, S. M. N., Fremouw, T. E., and Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of Neuroscience*, 24(41):9201–11.

Humphries, C., Liebenthal, E., and Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, 50(3):1202–11.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–6.

Hyde, P. S. and Knudsen, E. I. (2000). Topographic Projection From the Optic Tectum to the Auditory Space Map in the Inferior Colliculus of the Barn Owl. *The Journal of Comparative Neurology*, 421:146–160.

Imig, T. J. and Morel, A. (1985). Tonotopic Organization in Ventral Nucleus of Medial Geniculate Body in the Cat. *Journal of Neurophysiology*, 53(1):309–40.

Kaas, J. H. and Hackett, T. a. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11793–9.

Karklin, Y. and Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural computation*, 17(2):397–423.

Kowalksi, N., Depireux, D. A., and Shamma, S. A. (1996). Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex : I . Characteristics of Single Unit Responses to Moving Ripple Spectra. *Journal of Neurophysiology*, 76:3503–3523.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. in Systems Neuroscience*, 2.

Krizhevsky, A. and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research*, 60(2):115–142.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–63.

May, B., Prell, G. L., and Sachs, M. (1998). Vowel Representations in the Ventral Cochlear Nucleus of the Cat: Effects of Level, Background Noise, and Behavioral State. *Journal of neurophysiology*, 79:1755–1767.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of Neurophysiology*, 102(6):3329–39.

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(41):14205–16.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8(July):1–14.

Mole, C. and Klein, C. (2010). Confirmation, Refutation, and the Evidence of fMRI. In Hanson, S. J. and Bunzl, M., editors, *Foundational Issues in Human Brain Mapping*, chapter 9, page 99–111. MIT Press, Cambridge, MA and London, England.

Müller, M. and Ewert, S. (2010). Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–10.

O'Doherty, J. P., Hampton, A., and Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104:35–53.

Ogawa, S. and Tank, D. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(July):5951–5955.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13):607–609.

Peretz, I. and Zatorre, R. J. (2005). Brain organization for music processing. *Annual review of psychology*, 56:89–114.

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207):111–114.

Rodd, J. M., Davis, M. H., and Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral cortex (New York, N.Y. : 1991)*, 15(8):1261–9.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization. *Psychological Review*, 65(6):386–408.

Santoro, R. (2014). *The Computational Architecture of the Human Auditory Cortex*. PhD thesis, Maastricht University.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLOS Computational Biology*, 10(1):e1003412.

Schönwiesner, M., Dechent, P., Voit, D., Petkov, C. I., and Krumbholz, K. (2014). Parcellation of Human and Monkey Core Auditory Cortex with fMRI Pattern Classification and Objective Detection of Tonotopic Gradient Reversals. *Cerebral Cortex*, pages 1–12.

Schönwiesner, M. and Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14611–6.

Singh, N. C. and Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394.

Stansbury, D. E. (2014). *Modeling neural representation using statistical features of natural scenes*. PhD thesis, University of California, Berkeley.

Theunissen, F. E. and Elie, J. E. (2014). Neural processing of natural sounds. *Nature reviews. Neuroscience*, 15(6):355–66.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, 66(5):1364–1380.

Vinje, W. E. and Gallant, J. L. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, 287(5456):1273–1276.

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual review of neuroscience*, 29:477–505.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24.

Young, E. D. and Oertel, D. (1999). The Cochlear Nucleus. In Shepard, G., editor, *The Synaptic Organization of the Brain*. Oxford University Press.

Yu, D. and Deng, L. (2015). *Automatic Speech Recognition: A deep learning approach*. Signals and Communication Technology. Springer.

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nature reviews. Neuroscience*, 8(7):547–58.

Zeiler, M. and Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision-ECCV 2014*.

First Article.

Scientific Explanation in Neuroscience and Artificial Intelligence

by

Jessica A. F. Thompson[1]

This article was or will be submitted in Neurons, Behavior, Data Analysis, and Theory.

Résumé. De nombreux débats entourant l'utilisation des réseaux de neurones profonds (DNN) en tant que modèles de réseaux de neurones biologiques reviennent à débattre de ce qui constitue un progrès scientifique en neurosciences computationnelles. Afin de discuter de ce qui constitue le progrès scientifique, il faut avoir un objectif en tête (progrès vers quoi?). Un de ces objectifs à long terme est de produire des explications scientifiques sur les capacités intelligentes (par exemple, la reconnaissance faciale, le raisonnement relationnel). Je soutiens que les questions philosophiques les plus pressantes à l'intersection de l'intelligence artificielle et biologique concernent finalement la définition des phénomènes à expliquer et ce qui constitue des explications valables de tels phénomènes. À ce titre, je propose qu'une fondation dans la philosophie de l'explication scientifique puisse étayer les discussions futures sur les mérites des DNN en tant que modèles. Vers cette vision, je passe en revue plusieurs des théories de l'explication scientifique les plus pertinentes et commençons à décrire les formes d'explication possibles pour les phénomènes à l'intersection de l'intelligence artificielle et des neurosciences.

Mots clés : philosophie des sciences, explication scientifique, explication en neuroscience, explication causale

Abstract. any of the debates surrounding the use of deep neural networks (DNNs) as models of biological neural networks amount to debates over what constitutes scientific progress in computational neuroscience. In order to discuss what constitutes scientific progress, one must have a goal in mind (progress towards what?). One such long term goal is to produce scientific explanations of intelligent capacities (e.g. face recognition, relational reasoning). I argue that the most pressing philosophical questions at the intersection of artificial and biological intelligence are ultimately concerned with defining the phenomena to be explained and with what constitute valid explanations of such phenomena. As such, I propose that a foundation in the philosophy of scientific explanation can scaffold future discussions about the merits of DNNs as models. Towards this vision, I review several of the most relevant theories of scientific explanation and begin to outline candidate forms of explanation for phenomena at the intersection of artificial intelligence and neuroscience.

Keywords: M

1. Introduction

Neuroscience is constantly evolving as new methods to collect, analyze and model neural measurements are being developed. One such development has been the use of deep neural networks (DNNs) as models of biological neural networks, in particular the ventral stream of the primate visual system. This approach has gained popularity during a data-driven era of neuroscience where emphasis has been placed on collecting and integrating more (more cells, more regions, more trials) and better (higher resolution, higher signal-to-noise ratio) data than ever before. However, it has also become clear that data alone can't push neuroscience forward. The data is important but what is the data *for*?

One approach has been to build models that are able to predict neural activity while an animal is experiencing some task or stimuli. Traditionally, a model would be designed given what is already known about the system of interest and the researchers hypotheses about neural function. Modern DNNs, on the other hand, though originally inspired by biological neural networks, were designed to solve computer vision problems independent of any knowledge or specific hypotheses about neural function. That such networks trained to recognize objects in images learned representations that were similar to those found in the primate ventral stream caused much debate in the neuroscience community. Jim DiCarlo, one of the leading researchers in this area, has described his approach as turning the scientific problem of neuroscience into an engineering one where the primary goal is to optimize the accuracy of predictive models (DiCarlo, 2018). Competitions such as the Algonauts project (Cichy et al., 2019) and BrainScore (Schrimpf et al., 2018) seek to identify the models that achieve the best score on standardized tasks, akin to engineering competitions such as Kaggle. The approach of 'Predict, then Simplify' prescribes first building a predictive model and then trying to explain why it successful (Kubilius, 2017).

Critics of this approach have essentially claimed that these DNN similarity results don't count as scientific progress, at least not in the same way that the results of traditional modeling studies do, because the models themselves are "uninterpretable" (Kay, 2018). The approach has been also criticised as replacing one black box with another, i.e., modeling one thing we don't understand with another that we also don't understand (Middlebrooks, 2019). This criticism implies that no scientific progress has been made. How can representations learned in DNNs tell us anything about the brain when they don't encode specific hypotheses about neural function?

Many of the conversations alluded to above start by asking *what would it mean to understand the brain?*: e.g., the paper, "What does it mean to understand a neural network?" by Lillicrap and Kording (2019) or the Challenges and Controversies session, "What it would mean to succeed at understanding how cognition is implemented in the brain" at the 2018 conference on Computational Cognitive Neuroscience. These questions are framed as if their answers would help us to reason about the merits of using DNNs as models. Here I propose a reformulation of these questions to better serve that purpose. Instead of asking 'How to understand the brain' or 'How to understand a neural network?', let us focus our efforts on 1) defining the specific phenomena to be explained and 2) how the relevant classes of phenomena ought to be explained.

The second part of the reformulation seeks to emphasize the importance of scientific explanation, not just scientific understanding. Providing understanding might be one goal of scientific explanation (De Regt, 2017), but philosophers have identified several other desirable attributes, e.g. truthfulness, predictive power, or usefulness for future scientific efforts. At the individual level, understanding has been discussed as a type of personal, cognitive achievement state (Grimm, 2010). For example, when an individual comes to understand a language, another person, a proof, or a scientific theory, that individual has transitioned from a cognitive state of not understanding to understanding. One may distinguish here between understanding, a cognitive achievement state, and the *sense of understanding*, the subjective experience of understanding. One may also speak of understanding as an accomplishment of a group, where a scientific field, for example, might be said to understand a phenomena of study. In all cases, understanding is subjective and may change over time. An individual or a group may come to understand something and later update their understanding to take into account new information. On many accounts, scientific explanation, on the other hand, is typically thought to be more objective and unchanging. Explanations can be good or bad (sometimes equated with true or false) and their goodness is primarily a function of their relationship to the phenomena they are supposed to explain, rather than to the subjective perspective of scientists. For example, Europeans in the 19th century understood the superiority of the white race, based on explanations provided by colonialist scientists of the time (Saini, 2019). Hopefully, respected scientists today no longer find such explanations to provide the same understanding. While our understanding has evolved, the explanations have always been faulty. The explanations of colonialist race scientists are as poor now as they were two centuries ago. In retrospect, we can look back and analyze the signatures of these faulty explanations—what were they missing? where did they go astray?—and the social factors that influenced the understanding they provided.

One may disagree with my particular framing of the distinction between understanding and explanation. What matters here is to have a distinction between the human, subjective, cognitive state provided by an explanation and the somewhat more objective goodness or badness of an explanation. I find it useful to assign the subjective, cognitive achievement state to the word *understanding*, and let the objective component live in the word *explanation*. Both are clearly important for scientific progress. My goals here are to argue against equating understanding and explanation and to warn against neglecting scientific explanation in favor of understanding. Certainly we want to gain understanding, but we don't *only* want to understand—we want that understanding to be robust. If we want to effectively debate how our science will progress, we must consider the explanations we will produce not just the understanding they will provide.

The first part of my proposed reformulation focuses on defining the specific phenomena to be explained. I will argue that classes of similar phenomena, which may span distinct scientific domains, should be explained similarly. Here I will focus on phenomena that lie at the intersection of neuroscience and artificial intelligence, which I will define as phenomena that occur to some degree in both artificial and natural intelligence, e.g. learning in distributed networks, visual object recognition, language translation, navigation. According to this definition, the project of comparing representations in convolutional neural networks to firing rates in the primate ventral visual pathway constitutes research at the intersection of AI and neuroscience where the common phenomenon is the capacity to recognize objects in images. Here, as in most experimental psychology, a demonstration of a psychological capacity is operationalized by task performance. We may also consider phenomena at lower levels, e.g. patterns of neural dynamics or selectivity, that can be observed during learning, decision making, or action in both biological and artificial systems.

In the next section, I continue to justify this reformulation of the relevant philosophical questions when reasoning about scientific progress at the intersection of neuroscience and AI. I assume no background in philosophy of science, so this section also includes a gentle but woefully incomplete and superficial introduction to different ways of doing philosophy. Then, I review a subset of relevant theories of scientific explanation. I hope that this summary will help readers to identify what notions of scientific explanation are reflected in their own work and the work of their peers [1]. Lastly, I outline some desiderata for a theory of scientific explanation tailored to the phenomena at the intersection of AI and neuroscience. I hope that this paper provides a base from which to launch productive discussions about the nature of a unified science of intelligence.

2. Asking the Right Questions

To begin, let us assume that science eventually makes progress towards its goals. This doesn't imply that science proceeds directly to its goals or that all its goals are achievable, but it does imply that the activity of doing science is more than just going in circles—that science moves towards something. Some applied sciences will have very clearly defined goals, such as improving patient outcomes for medical research or enabling the development of new technology. In the absence of clear applied research goals, one of the primary goals of fundamental science, of science for science's sake, is to explain, i.e., to provide explanations of the phenomena that are the focus of scientific study.

So then, if we want to understand how fundamental science progresses, we need to know what constitutes a scientific explanation. This has been a central question for philosophers of science over the past century. To develop a theory or model of scientific explanation is to characterize the structure of scientific explanations and to define the criteria that must be met in order for a phenomenon to be successfully explained. Around the mid twentieth century, philosophers who tackled this question sought to identify a universal and objective logic of scientific explanation. Much of this work can be characterized as *armchair philosophy* because it reflects the belief that philosophy of science can pass judgment on and discern the rules of science from a non-scientific view point. Philosophers looked to physics as the

[1] To those who might protest that their work does not reflect a philosophical stance, I agree with **?** when they write, "there is no escape from philosophy. Every scientist takes a philosophical position, either tacitly or explicitly, whenever they state that a result is "important," "fundamental," or "interesting." This is because such assertions are always a judgment from outside of science. There is no "interesting" variable inherent to the data that can be objectively plotted on a graph—abstract reasoning and normative claims cannot be substituted by, or obtained from, data." (**?**, pg. 485)

model science and tried to develop general theories of explanation that would account for all scientific explanations. Proposals included that explanations are deductive arguments based on laws of nature (deductive-nomological model) and that explanations are collections of statistical relevance relationships (statistical relevance model). These works viewed science as providing an objective window onto truth and assumed that it must have a clear and universal set of rules. This enterprise is largely considered to have failed to achieve its goal since each theory has a number of counter examples for which it is unable to account [2].

In contrast, the field of science studies centers the fact that science is performed by bias-laden humans in a particular social and historical context which will effect what questions get asked, how science is performed, and how its results are presented and received. An extreme view on explanation within this tradition is that scientific explanation is simply whatever scientists find to be explanatory at a given time and place—that there is no objective component of explanation independent from its social and historical context. This view is unsettling to many scientists and philosophers who want to believe that there is something special about science compared to other ways of knowing. Wesley C. Salmon writes,

> First, we must surely require that there be some sort of *objective* relationship between the explanatory facts and the fact-to-be-explained. . . Second, not only is there the danger that people will feel satisfied with scientifically defective explanations; there is also the risk that they will be unsatisfied with legitimate scientific explanations . . . The psychological interpretation of scientific explanation is patently inadequate (Salmon, 1984, pg. 13).

Similarly, Carl Craver writes, "All scientists are motivated in part by the pleasure of understanding. Unfortunately, the pleasure of understanding is often indistinguishable from the pleasure of misunderstanding. The sense of understanding is at best an unreliable indicator of the quality and depth of an explanation"(Craver, 2007, pg. 21).

The extreme psychological interpretation says that scientific explanation is nothing more than a consensus of scientific understanding—whatever the field agrees is explanatory at a particular moment. On the other extreme, some accounts would say that what constitutes an explanation is exclusively determined by the physical mechanism producing the phenomenon to be explained and has nothing to do the psychology of scientists. Many contemporary perspectives will fall somewhere in between, acknowledging that there must be an objective component but also recognizing the human element of scientific explanation, e.g. that the explanation must be expressed in human-readable language or mathematics. For the purposes of this article, we need only commit to there being a non-negligible objective component of scientific explanation and it is this objective component on which I wish to focus. There is

[2] However, that doesn't mean that these older theories are now irrelevant. Some explanations may still fit nicely into one of these theories and some of the component ideas have been revised and incorporated into more contemporary theories.

an objective goodness or badness (which may or not correspond to proximity to truth [3]) accorded to an explanation that is independent of the understanding that it yields. A scientist may momentarily consider a good explanation of some phenomenon and discard it, thinking it implausible for whatever reason, but its goodness remains even after the scientist rejects it. Different theories of scientific explanation may identify this goodness or badness with different attributes, e.g. predictive adequacy, unification, usefulness.

In the later part of the twentieth century, after several failed attempts at a universal theory of scientific explanation, philosophers began to adopt a more context-specific approach. Many found that theories originally developed to account for explanation in physics did not transfer well to the biological sciences and began to develop new theories specific to particular scientific disciplines. This coincided with an increase in empirical philosophy, which borrows qualitative methods from fields like history, anthropology, and psychology, such as interviews and field observations, to build theories based on observations rather than based solely on reason and introspective conceptual analysis (as in the armchair philosophy mentioned above). This empirical approach may include the analysis of historical explanations that have either passed or failed the test of time. If an explanation is robust, makes accurate predictions, enables the development of new technology, and/or leads to new productive research, then we may take it to be good. Empirical philosophers develop theories to account for their observations of good and bad explanations and of how scientists went about developing or discovering these explanations.

There are empirical philosophers of science today dedicated specifically to neuroscience who develop theories of explanation in neuroscience. But, given the highly interdisciplinary nature of neuroscience, I am skeptical about the feasibility of a single theory of explanation to account for all explanation in neuroscience. Neuroscience is not clearly separated from its sister sciences: biology, physics, psychology, AI, etc. Thus, the search for a universal theory of explanation in neuroscience may be as ill-fated as the original quest for a theory of explanation for all science. I think the spirit of the idea of field-specific theories of explanation is that similar phenomena ought to be explained similarly, assuming that phenomena within a branch of science will be more similar than phenomena from different disciplines. This assumption does not seem to hold for neuroscience where, for example, one phenomenon might be closer to biophysics and another closer to psychology than the two are to each other. Rather than organizing our theories of explanation around objects of study (in this case, the brain) or the departmental silos of our academic institutions, I propose we organize our theories around classes of similar phenomena, regardless of which specific scientific discipline the phenomena belong to.

[3] Pragmatic anti-realist accounts of scientific explanation may reject the notion that explanations can be true or false, but maintain that they can be good and bad, identifying their goodness with their usefulness. (c.f. Bas van Frassen)

This organization can be especially unifying at the intersection of neuroscience and AI where the behaviour of artificial systems are designed to mimic human and animal behaviour. This mimicry brings into alignment the scientific goals of AI and much of cognitive computational neuroscience, which both seek to identify the computational components (algorithms, procedures, mechanisms, cost functions) that underlie cognitive abilities. Take for instance the growing sub-discipline sometimes referred to as 'science of deep learning' or 'understanding deep learning'. The ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena solicited 'contributions that view the behavior of deep nets as natural phenomena, to be investigated with methods inspired from the natural sciences like physics, astronomy, and biology'. Similar research has also been referred to as 'artificial neuroscience' (Metz, 2018) or 'synthetic neurophysiology' (Kriegeskorte, 2015), to highlight the similarity of both methodologies and goals (e.g. ablations/lesioning, comparisons of task performance, characterization of neural selectivity patterns). Therefore I propose we consider theories of explanation that are specific to the classes of phenomena we associate with animal intelligence, such as learning and cognitive abilities, irrespective of whether they manifest in biological or artificial systems.

Within this phenomenon-specific framework, it becomes very important to clearly identify and delineate the phenomena to be explained. How a scientist conceptualizes the phenomenon to be explained may bias them towards one form of explanation or another. This is especially apparent in the cognitive sciences where there are many different perspectives on the nature of mind and cognition. Is cognition computation? Are cognitive agents embodied dynamical systems? If explanations are phenomena-specific, we cannot completely separate the ontological question (What is cognition?) from the epistemological question (How do we explain cognition?). Thus, a commitment to a particular theory of explanation in neuroscience may also suggest a related commitment to a theory of cognition. For example, the information processing view of vision, as exemplified by pioneers like David Marr (1982), may bias vision researchers towards functional explanations, which proceed by decomposing a phenomenon into its component operations and showing how those operations are organized to exhibit the phenomena to be explained (Cummins, 1975). Alternatively, a radical embodied perspective, like that espoused by Chemero (2009), might bias vision researchers towards dynamical explanations and away from explanations that rely on representations.

Returning to the original reformulation goal, the relevant philosophical questions are often phrased as being about how to 'understand the brain' or a 'understand a neural network'. Firstly, I encourage shifting the focus from understanding to explanation. Lillicrap and Kording (2019) use 'understanding' similar to how I use it here, associating it with a cognitive achievement. They emphasize compactness and compressibility in their proposal for what it means to understand a neural network because humans are only able to argue about compact systems. According to their view, any meaningful understanding of a neural network must be compressible into an amount of information that a human can consume, e.g. a textbook. The limits of human cognitive abilities constrain the understanding that can reasonably be sought. For example, humans cannot conceptualize the interactions of 100 trillion synapses simultaneously and so our scientific goals should not require such a feat. What, if anything, might this say about explanation? By some accounts of explanation, the limits of human cognition have no bearing on explanation because the goodness of an explanation is purely a function of its relationship to the phenomenon to be explained. For those who think that good explanations ought to be true, there is no reason to believe that a bias towards human understandability would lead scientists closer to truth; the truth is not necessarily easy to understand. By other accounts, one might consider understandability as a constraint; we don't just want good explanations, we want explanations that produce human understanding[4]. For my own part, I don't want to insist that good scientific explanations must yield immediate understanding because I recognize that our understanding is constantly changing as we develop new concepts and new mental instruments to help us manipulate them. I prefer to assume that good explanations exist, regardless of whether a human scientist can ever grasp them, and to treat the process of developing human understanding as a separate but related process. Consequently, there may be some phenomena whose explanations humans will never be able discover. I think it is important that our conception of explanation leaves this option open.

Secondly, I caution us against seeking a single answer for how to *explain the brain*. Science isn't in the business of explaining objects. Science may produce descriptions or characterizations of objects, but explains phenomena. This phrase, 'explain the brain', could be interpreted as short hand for 'explain all the phenomena that the brain is involved in'. This phrasing subtly reflects a commitment to the idea that all phenomena involving the brain can be explained in a similar fashion, i.e. a unitary theory of explanation in neuroscience. Alternatively, if one organizes theories of explanation around phenomena rather than objects of study, then we cannot discuss how to 'explain the brain'; we need to be more specific. The brain participates in a plethora of distinct phenomena at many different spatial and temporal scales. What specific phenomenon one seeks to explain will determine how it ought to be explained.

3. Scientific Explanation

Scientific explanation consists of the *explanandum*, which is the target of the explanation (the phenomenon to be explained), and an *explanans*, which does the explaining. An account of scientific explanation must distinguish between explanations and non-explanations. For example, a set of claims about the appearance of a particular species may be true, accurate

[4] or equivalently that one of the attributes that makes a scientific explanation good is that it produces understanding

and supported by evidence without being explanatory in any way. They are *merely* descriptive Woodward (2017). Explanations may still be descriptions, but they are descriptions that also explain. Explanations are thought to answer why-questions while non-explanatory descriptions might answer how- or what-questions.

3.1. Deductive-Nomological (DN) Model

According to the Deductive-Nomological (DN) model [5], An explanation is the deductive argument that shows that the explanandum is expected given the premises of the explanans, where the explanans successfully explains the explanandum only if:

(1) the explanandum is a logical consequence of the explanans, and

(2) the explanans relies on at least one law of nature in its explanatory logic.

The term *law* here is used to differentiate deterministic laws from other true generalizations that are only accidentally true (Hempel, 1965). Unfortunately, little agreement about how to define this notion of lawhood has emerged in the decades since the DN model was proposed. Additionally, many generalizations that are central to explanation in the special sciences (biology, psychology, economics, etc.) fail to satisfy any of the standard criteria for lawfulness. Are fundamental laws of nature only present in physics? If so, then how would the DN model apply to explanations in the special sciences?

Similarly, deductive-statistical (DS) explanation relies on statistical laws instead of deterministic ones, and inductive-statistical (IS) explanation will be "successful to the extent that its explanans confers high probability on its explanandum outcome" (Woodward, 2017). According to all DN variants, "the essence of scientific explanation can be described as *nomic expectability*–that is, expectability on the basis of lawful connection"(Salmon, 1989, pg. 57)

There are a number of well-known counter examples to the DN model where either a good explanation is not captured by the DN model or where a faulty explanation satisfies the DN model. Two main issues emerge from these counter examples:

— Explanatory asymmetries: derivation of an explanandum from a law and initial conditions can meet the critiera for a DN explantaion, while the reverse derivation of initial conditions from the explanandum and law is not explanatory, yet still satisfies the DN model. The DN model doesn't account for the fact that some explanations are directional. The classic example is that of a flagpole's shadow. The position of the sun relative to the flagpole will explain the length of its shadow and not vice versa.

— Explanatory Irrelevancies: A derivation may satisfy the DN model, while relying on a true generalization that is irrelevant to the explanandum. Consider this counter example from Salmon (1971): "John Jones avoided becoming pregnant during the

[5] also known as Hempel's model, the Hempel–Oppenheim model, the Popper–Hempel model, or the covering law (CL) model

> past year, for he has taken his wife's birth control pills regularly, and every man who regularly takes birth control pills avoids pregnancy." Assuming that John Jones is a cis-man without female reproductive organs, no generalizations concerning birth control will ever play a role in explaining why he does not get pregnant. Yet, this example satisfies the DN-model.

In both of the above examples, there is a causal story that seems to determine what can and cannot be explanatory. The DN model failed to capture the true causal factors of the *explanandum.*[6]

3.2. Statistical Relevance Model

The Statistical Relevance (SR) model (Salmon, 1971) is an attempt to capture the features of causal or explanatory relevance that elude the DN model variants. Statistical relevance here refers to conditional dependence. It is assumed that causal relationships are captured by statistical relevance relationships. The main claim is that explanatory properties are statistically relevant: "Given some class or population A, an attribute C will be statistically relevant to another attribute B if and only if $P(B|A.C) \neq P(B|A)$—that is, if and only if the probability of B conditional on A and C is different from the probability of B conditional on A alone" (Woodward, 2017). This tackles head on the problem of explanatory irrelevancies in the DN model. Notice though, that an explanation is no longer an argument, as in the DN model. Here, an explanation is a collection of information that is statistically relevant to the explanandum. A consequence of this model is that an explanation need not make an explanandum expected, as in the IS model:

— I-S model: an explanation is an *argument* that renders the explanandum *highly probable.*

— S-R model: an explanation is an *assembly of facts statistically relevant* to the explanandum, *regardless of the degree of probability* that results. (Salmon, 1971, pg. 11)

A high probability event (e.g. a biased coin toss landing on heads) and its alternative low probability outcome (landing on tails) are both explained by the same explanans (the bias of the coin and the action of tossing).

A limitation of the SR model is that it relies on the condition of objective homogeneity, which requires that there are no omitted variables that would affect the relevant probabilities. This condition is rarely met in most sciences. It may hold when studying quantum mechanics in controlled experiments, but likely will not when trying to explain phenomena like recovery from illness or juvenile delinquency. This is the same problem that makes it difficult to estimate causal effects from observational data.

[6] Further counterexamples and objections to the DN and IS models can be read in Salmon (1989)

Another problem with the SR model is the assumption that causal relationships are captured by statistical relevance relationships. Explicitly, the assumption is that if a cause is present, some conditional probability will increase. This statement is incorrect for two reasons:

— An increase in conditional probability may be spurious, which may or may not be due to omitted causal variables. For example, individuals who purchase life insurance may live longer, however, purchasing life insurance would not be an effective strategy to extend one's life (Cartwright, 1979).

— Causal relationships are underdetermined by statistical relevance relationships. Several causal structures may yield the same structure of conditional dependence between variables.

"…statistical relevance relations, in and of themselves, have no explanatory force. They have significance for scientific explanations only insofar as they provide evidence for causal relations" (Salmon, 1989, pg. 166). [7]

3.3. Causal Mechanical Model

The criticisms against the DN and SR family of theories for not capturing the causal aspects of explanation led some philosophers to focus explicitly on the causal nature of scientific explanation. The Causal Mechanical (CM) model was presented as an alternative. According to this model, scientific explanation is a matter of tracing the causal processes that lead to the explanandum. The CM model posits the following constraints on scientific explanation (as described in Craver (2007)):

— mere temporal sequences are not explanatory;

— causes explain effects and not vice versa;

— causally independent effects of common causes to not explain one another;

— causally irrelevant phenomena are not explanatory; and

— causes need not make effects probable to explain them.

Different types of causal explanations, corresponding to different types of mechanisms, can be organized in a hierarchical taxonomy. At the top level, one can distinguish between etiological and constitutive mechanisms.

(1) Etiological explanation: To explain in terms of antecedent causes, i.e., to "trace the causal processes and interactions leading up to [the explanandum]" (Woodward, 2002, pg. 44), e.g., the virus causes the flu, dehydration causes thirst. The explanandum is *produced* by the mechanism.

[7] This is the last I'll mention of the SR model, but I wanted to mention it because I have heard people in our community make claims about the explanatory power of statistical relevance relations. For example, I've heard that we don't need philosophy because we have Bayesian statistics (Nemenman, 2018) and I've observed researchers confounding scientific explanation with statistical explanation of variance.

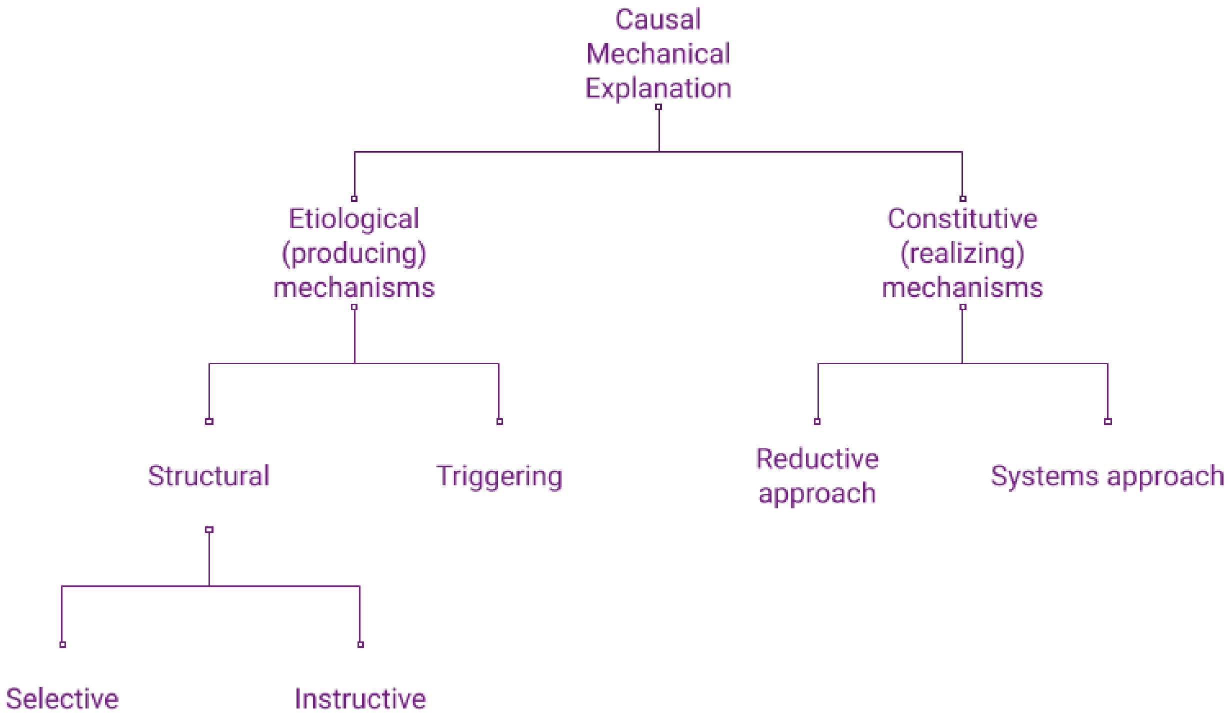

Causal Mechanical Explanation
Etiological (producing) mechanisms
Constitutive (realizing) mechanisms
Structural
Triggering
Reductive approach
Systems approach
Selective
Instructive

Figure 14. Taxonomy of Mechanisms

(2) Constitutive (or componential) explanation: To explain via description of causal relationships among component parts and their activities. The explanandum is *realized* by the mechanism.

We can further distinguish between structural and triggering etiological mechanisms. Structural mechanisms set up the necessary conditions such that a trigger will cause the explanandum. Structural mechanisms can be selective (like natural selection) or instructive (like pedagogy)(Craver, 2002). Within constitutive explanation, we can distinguish between the systems tradition and the reductive tradition. The systems tradition explains by decomposing a system into its parts and demonstrating how those parts are organized such that they exhibit the explanandum. Reduction is a loaded word with many different meanings in philosophy of science. By some accounts, the aforementioned systems approach is also reductive. All constitutive explanations might be said to be reductive in a sense since they explain via description of component parts. However, Craver (2007, pg. 108) identifies the reductive tradition with the specific view that 'explanation proceeds by constructing identity statements (or partial identity statements) between the kind-terms of the higher-level theory and those of the lower-level theory and then deriving the laws of the higher-level theory from the laws of the lower-level theory.' The systems approach, as described by Craver, requires no such derivation or one-to-one mapping and also allows for explanations to describe multi-level mechanisms.

One challenge to the CM model is that one needs a way of identifying only those causal relationships that are relevant to the explanandum (Hitchcock, 1995). Consider a billiard ball, which, after being struck by a cue stick, moves in a particular direction to make impact with another ball on the table. Let's imagine that the cue stick has also left a mark of blue chalk on the ball. The relevant causal factors to explain the motion of the balls on the table will not include the blue chalk mark, even though the transmission of the blue chalk mark is part of the sequence of causal processes and interactions that led to the explanandum. This example demonstrates the importance of counterfactuals in explanation. The presence of the chalk mark is not explanatorily relevant because the explanandum would be unaffected by it absence (Woodward, 2002).

Another potential critique has to do with the constraint that explanations trace continuous causal processes where causal processes are exclusively physical processes. The components of mechanisms must be physical objects, not abstract concepts. This presents challenges to formal sciences, like theoretical computer science, that are concerned with abstract quantities, or the social sciences where the causally relevant factors may not map easily to physical objects and their interactions. For example, is a social norm a physical object? Such criticisms of the CM model are often rebutted by claiming either that the relevant components of the mechanism are abstractions of physical objects (e.g. hunger is an abstraction of a bodily state), or, if no such abstraction is clearly defined, that only a how-possibly, not a how-actually model of the mechanism has been provided. Some (causal mechanist and otherwise) would argue that formal sciences that do not rely on empirical observation are not in fact true sciences to begin with, and so are irrelevant to discussions about scientific explanation.

4. Explanation in the Cognitive Sciences

Now, equipped with some background knowledge about scientific explanation in general, let us discuss explanation specifically in psychology and neuroscience.

4.1. The Deductive Nomological Model

Consider the field of psychophysics, which seeks to identify relationships between physical properties of sensory stimuli and human experience of those stimuli. For example, Fechner's law ($\Psi = c \log(I/I_0)$) states that the "intensity of a sensation (Ψ) is proportional to the logarithm of the intensity of the stimulus (I) relative to the threshold intensity (I_0)" (Bechtel and Wright, 2009, pg. 2). According to the DN model, such laws could be considered to explain individual percepts because they show that the percept is to be expected based on a more general regularity. However, these general regularities themselves are left unexplained. The DN model interpretation states that Fechner's law helps to answer why-questions like, Why do humans experience item A as twice as heavy as item B? The causal mechanical interpretation is that Fechner's law answers a how question (How is the perception of weight related to mass?) and the corresponding why-question (Why does Fechner's law hold?) is unanswered.

In general, the view that to explain is to make a phenomenon expected based on universal or statistical laws is referred to as *predictivism*. "On a very liberal interpretation of predictivism, any mathematical or computational model that predicts all the relevant features of the phenomenon in a wide range of conditions counts as an explanation" (Craver and Kaplan, 2011, pg. 269). The predictivist holds that phenomenological models, models that characterize or store a phenomenon, are explanatory by virtue of their ability to make accurate predictions about the phenomenon. Phenomenological models in neuroscience are called descriptive models because they "summarize data compactly" without addressing "the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry" (Dayan and Abbott, 2005).

In " 'How does it Work?' vs. 'What are the Laws?' ", Robert Cummins (2000) argues against the application of the DN model in psychology, i.e. that psychological phenomena cannot be explained by subsumption under law. He suggests that the laws of psychology are actually the phenomena to be explained, rather than features of an explanation. Laws in psychology are what are commonly referred to as effects, e.g. the McGurk effect, the spacing effect. Most efforts in psychology are dedicated to identifying and describing effects. This activity in and of itself can only be considered to explain to the extent that one believes that explanation consists of subsumption under law. He writes that the focus on effects in psychology is "fostered by a confusion between explanation and prediction" (Cummins, 2000, pg. 4). He argues instead that explanation and prediction are orthogonal: that one can predict without being able to explain (as in the ocean's tides) and that one can explain without being able to predict (as in stochastic or chaotic systems whose relevant initial states are unknown). Knowledge of the McGurk effect may enable the prediction of human perception, but according to Cummins, it does not explain that perception. Instead, the McGurk effect is a phenomenon to be explained, for example by reference to multi-modal interaction.

4.2. Functional Explanation

According to Cummins, the main explananda in psychology are *capacities*: "the capacity to see depth, to learn and speak a language, to plan, to predict the future, to empathize, to fathom the mental states of others, to deceive oneself, to be self-aware, and so on" (Cummins, 2000, pg. 8–9). He proposes that capacities are explained via functional analysis and realization. Functional analysis refers to the process of decomposing a capacity into a number of simpler subcapacities and their functional organization. Realization in this context refers to the requirement that the analysis must show how the behaviour of the parts of the system come together to enable the system to demonstrate the capacity to be explained (Cummins, 1975). However, according to a mechanist perspective, such descriptions only provide a how-possibly model, and provide no evidence that it corresponds to a how-actually model (Piccinini and Craver, 2011).

4.3. Causal Mechanical Explanation

Several contemporary philosophers claim that a version of the CM model accounts for explanation across several areas of neuroscience and the cognitive sciences. Carl Craver (2007) says neuroscience produces constitutive (or componential) causal mechanical explanations: "to explain a phenomenon, one has to know the mechanism that produces it, one has to know what its components are, what they do and how they are organized together (spatially, temporally and hierarchically) such that they give rise to the phenomenon to be explained" (Craver and Kaplan, 2011, pg. 269). Craver also deals with the issue of causal relevance, adopting a view based on manipulation: X is causally related to Y if that relationship is potentially exploitable to manipulate or control Y.

He enumerates six aspects of mechanistic explanation in neuroscience:

(1) The nature of the phenomenon to be explained: delineation, description, characterization

(2) The constitutive relationship between a phenomenon and its components: decomposition

(3) The difference between real components and useful fictions: distinguishing *as is* from *as if*

(4) The nature of capacities or activities: what are the actions undertaken by the parts of the system

(5) The nature of mechanistic organization: what matters is not just the sum of the parts but how they are organized to interact

(6) The nature of constitutive explanatory relevance: not all parts of a system are components of a mechanism. The explanatory relevance of each component must be established.

According to Craver, explanations in neuroscience describe mechanisms, span multiple levels and integrate multiple fields.

Applied to computational neuroscience, the mechanist would say that a model is considered to be explanatory only when it satisfies strict model-to-mechanism mapping (3M) requirements:

(1) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon

(2) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism. (Craver and Kaplan, 2011, pg. 272)

According to this view, a computational model is explanatory to the extent that it faithfully describes the physical mechanisms that realize a given phenomenon (Kaplan, 2011).

Causal mechanisms have also been posited to account for explanation in cognitive neuroscience (Bechtel, 2008). According to William Bechtel (2004), a mechanism producing a specific behaviour must first be decomposed structurally and functionally into component parts and component operations respectively. The localization goal of much of cognitive neuroscience amounts to mapping the component operations to the component parts. However, according to the mechanist perspective, functional analyses only provide 'sketches' of explanations. These sketches can potentially eventually be developed into full-fledged mechanical explanations, but are not yet explanatory in their own right. In this sense, functional analyses, such as those performed in cognitive psychology to decompose a cognitive capacity into subcapacities, constitute a first step towards ultimate mechanical explanations (Craver and Kaplan, 2011; Piccinini and Craver, 2011). This is contrary to the functionalist perspective on explanation in cognitive psychology which states that explanations of cognitive phenomena need not necessarily make reference to physical components—identification of the component operations and demonstrating an organization of those operations that yields the phenomenon to be explained is sufficient for the phenomenon to be explained. Both perspectives agree that functional analysis is useful, but they disagree on long term explanatory goals.

4.4. Dynamical Explanation

There exist several views on dynamical explanation and how it might fit into the landscape of theories of explanation. Dynamic models invoke the mathematical framework of dynamical systems theory to model complex systems with differential equations. It has been argued that dynamic explanation adheres to the DN model of explanation (Bechtel, 1998; Stepp et al., 2011). Similar to the psychophysics example in section 3.1, dynamical models can be said to identify the mathematical regularities that govern how a phenomenon unfolds over time. Others have proposed that dynamical explanations resemble mechanistic explanations: dynamical accounts are explanatory to the extent that they characterize underlying dynamic mechanisms (Zednik, 2011; Kaplan and Bechtel, 2011). It might be best, then, not to think of dynamical explanation as a distinct theory of explanation. One might argue that a dynamical model is explanatory or not by reference to one or more theories of explanation depending on the specifics of the model itself and the phenomenon to be explained.

4.5. Minimal Model Explanations

Lauren Ross (2015) has proposed that not all dynamical explanations appeal to mechanisms. She analyzes an example of explanation in dynamical systems neuroscience that is well characterized instead by Robert Batterman's theory of minimal model explanation.

Explanations are considered to answer why-questions in science. Batterman distinguishes between why-questions that ask why a phenomenon manifests in a particular situation and why-questions that ask why a phenomenon manifests generally or in a number of different circumstances. Minimal model explanations are concerned with the latter type while mechanistic explanations are concerned with the former. Ross discusses the example why-question: Why do neurons that differ drastically in the microstructural details all exhibit the same type of excitability? The *canonical model* approach to this question attempts to reduce the complexity of molecularly diverse neural systems to a single, abstracted model (the canonical model) that explains excitability. Ross argues that the canonical model approach is an example of a dynamical model in neuroscience that provides a minimal model explanation.

Minimal model explanation employs mathematical abstraction techniques to delineate a set of physically distinct systems that demonstrate some shared behaviour (Batterman and Rice, 2014). Batterman writes,

> explanation of universal behavior involves the elucidation of principled reasons for bracketing or setting aside as 'explanatory noise' many of the microscopic details that genuinely distinguish one system from another. In other words, it is a method for extracting just those features of systems, viewed macroscopically, that are stable under perturbation of their microscopic details (Batterman, 2001, pg. 43)

This theory of explanation differs from the causal mechanical account in that the explanation need not share relevant features with the phenomenon to be explained. Instead, the explanation must abstract away from specific features of the phenomenon to enable wider generalization.

5. Explaining Intelligence

What are the why-questions at the intersection of neuroscience and AI (neuro-AI)? When asked about the common goals of neuroscience, cognitive science and AI, some researchers have answered that the common goal is to "explain intelligence" or to uncover "the laws of physics for intelligence" (c.f. the discussion panels at the 2017 and 2018 Cognitive Computational Neuroscience (CCN) conference). Leading AI researcher Yoshua Bengio has often described his goals as being focused on discovering general principles of learning and intelligence, presumably that would govern both artificial and biological learning systems. These comments suggest two philosophical stances: (1) a commitment to scientific why-questions that are invariant to (or abstract over) the differences between artificial and biological intelligence, and (2) an appeal to the explanatory power of laws. The potential for DN model-style thinking to proliferate in AI research can also be evidenced by machine learning's relationship to physics. For example, 'Statistical Physics of Learning' was a subject area at the 2020 Neural Information Processing Systems conference. In general in machine learning theory, exact solutions, where phenomena can be precisely derived from idealized or abstracted models, are celebrated. In the previous sections, I reviewed a number of criticisms of the DN family of theories of scientific explanation, especially when applied to phenomena outside of physics, and a number of alternate views of scientific explanation that make no appeal to the explanatory power of laws. Therefore, I would like to consider how to maintain stance (1), while abandoning stance (2).

I propose that many why-questions at the intersection of neuroscience and AI are similar to the why-questions that the minimal models theory of explanation are said to address. The canonical model approach as discussed by Ross (2015), "explains why physically distinct neural systems all share the same behavior by showing that principled mathematical abstraction techniques—which preserve qualitative behavior—can be used to reduce all models of these distinct systems to the same canonical model" (pg. 15). For example, consider an AI system with human-level ability to recognize faces. A canonical model may explain why the AI system and a human demonstrate (or do not demonstrate) the same behaviour. We can also ask why-questions about learning in distributed networks, the answers to which would hold for some class of networks, regardless of whether they were implemented in cells or silicon.

On the other hand, some why-questions will be about a specific manifestation of a phenomenon (e.g. in human brains or DNNs or brains of a particular clinical population). In these cases, the why-question and its answer, appropriately stated, will not abstract over the relevant features that define the particular manifestation in question. Cartwright's distinction between theoretical and causal explanation in physics may be a useful parallel here. She suggests that theoretical explanations organize and unify diverse phenomena, without necessarily corresponding to physical reality (in fact, she suggests that the explanatory power of theoretical laws is at odds with their truthfulness). Causal explanations, on the other hand, "describe the concrete causal process by which a phenomenon is brought about" (Cartwright, 1983, pg.4), of which there can be only one correct account which corresponds to the true causal process. Batterman makes a similar distinction between why-questions that ask why a phenomenon manifests in a particular situation vs why-questions that ask why a phenomenon manifests generally (Batterman and Rice, 2014). Rather than two distinct categories, it seems to me that all why-questions require a clause of scope, where that scope can be more or less specific. Thus, perhaps the minimal models theory can be applied to seemingly more specific questions as well. For example, why-questions about why certain behaviours are exhibited by a particular architecture have a scope that includes all instantiations of that architecture, but not other architectures and not human brains.

To be precise, I propose that explanations answer why-questions that include both a description of the phenomena to be explained and the scope in which the answer should apply. An alternative view might say that the relevant scope is part of the definition of the phenomenon itself. I prefer to keep them separate because it allows us to discuss why-questions that are concerned with phenomena that are exhibited in both biological and artificial systems, while allowing the scope to be an additional, separate variable. For example, Leavitt and Morcos (2020) recently posed the question, Why does class selectivity emerge in deep neural networks trained on classification tasks? The phenomenon, the emergence of class selectivity, also occurs in animal brains, therefore we can additionally ask, Why does class selectivity emerge in rodent brains? or Why does class selectivity emerge in human brains? or Why does class selectivity emerge in both rodent and human brains? or Why does class selectivity emerge in both DNNs and human brains? In all cases, the why-question is asking, What is the common reason that this phenomenon (emergence of class selectivity) occurs in some set of observations, where the scope of that set is larger (e.g. all animals) or smaller (e.g. DNNs of a particular architecture). I propose that the why-questions at the intersection of neuro-AI are those about phenomena that occur in both artificial and biological intelligence, where the scope may or may not include both. This is a relatively broad definition as it includes both AI research that doesn't appear to care about the brain and brain research that doesn't appear to care about AI, depending only on their phenomenon of study, not the particular object in which it occurs.

As discussed in section 2, if what constitutes a valid scientific explanation is dependent on the phenomenon to be explained rather than on the field or object of study, then, to the extent that an artificial and biological system demonstrate the same phenomenon, what constitutes a valid explanation of that phenomenon will be the same in both, even if the content of the explanations differ. For example, analyzing which visual features an agent uses to make decisions about an image reflects a functionalist approach where detection of individual features are the component operations that combine to yield the decision—the phenomenon to be explained. Geirhos et al. (2019) showed that convolutional neural networks trained the ImageNet dataset are biased to recognize texture rather than shape, whereas humans privilege shape over texture when making decisions about object category. Although the content of the explanations differ (one prefers texture, the other shape), the same approach to explanation (decomposition into component operations) is applied to both artificial and biological intelligence.

Thus, when evaluating theories of scientific explanation for why-questions about phenomena that occur in both artificial and biological systems, we can insist that the theory of explanation be appropriate for both manifestations . This may lead us to eliminate some candidate theories that appear appropriate in one context but not the other. For example, causal mechanical theories of explanation are currently dominant in philosophy of neuroscience. However, such theories do not seem particularly well-suited to explain the behaviour of AI systems. For AI systems implemented on digital computers, the explanatorily relevant factors are often abstract, theoretical entities. An explanation of a particular capacity exhibited by an artificial system should be invariant to whether the system was trained on one graphical processing unit or another. Thus, I conclude that causal mechanical theories of explanation as currently conceived will not account for explanations of phenomena at the intersection of neuroscience and AI. Relaxing the strictness of the model-mechanism-mapping requirement may provide a version of the CM model that is appropriate for neuro-AI phenomena, as has been recently proposed by Cao and Yamins (2020); Ritchie (2020).

It appears that causality is important, but that the relevant causal factors are not necessarily components of physical mechanisms. In machine learning, it is easy to imagine counterfactuals involving theoretical entities or mathematical abstractions. In some cases, as in mathematical analysis of neural networks, it is not even necessary to physically instantiate the network in order to reason about what would occur under some intervention. For example, the analysis of the training dynamics of deep linear networks enables one to exactly describe the entire learning trajectory (Saxe et al., 2015). One can then ask questions about how this trajectory might be affected by using different learning rates without ever actually instantiating a network on a computer.

While I reject the CM model, I still find the taxonomy of mechanisms discussed in Section 3.3 useful for contextualizing various scientific questions in neuroscience and AI. I agree with Craver that most contemporary references to mechanisms in neuroscience reflect a commitment to the systems approach to constitutive explanation. I think that part of the resistance to or confusion about what has been called a *deep learning framework for neuroscience* (Richards et al., 2019), which focuses on how architectures, cost functions, and learning algorithms produce intelligent behaviour, is because this framework is concerned with etiological mechanisms rather than constitutive ones. Learning and model selection procedures are structural, etiological mechanisms that establish the conditions such that a trigger, such as an image, yields the phenomenon to be explained, for example, the recognition of an object in the image. For comparison, a constitutive story about that same recognition would amount to describing the component parts of the network (the units and weights) and how they interact (likely after training) to realize the act of recognition. In the constitutive story, the recognition and learning to recognize might be two separate phenomena to be explained with their own constitutive mechanisms. In the etiological story, the learning is one of potentially many structural mechanisms that together set up the conditions such that the recognition, a triggering mechanism, can occur. While neuroscience has mostly focused on constitutive stories, machine learning research has largely focused on etiological stories. The proposal for a deep learning approach to neuroscience invites neuroscientists to consider etiological stories as well, inspired by how useful they have been in deep learning research.

My rejection of the mechanist perspective does not lead me to embrace the functional account either. I agree with Cummins that prediction and explanation can be orthogonal goals: one can predict without being able to explain and vice versa. However, I don't believe that functional analysis will be sufficient to explain neuro-AI phenomena. It is insufficient to demonstrate that an organization of decomposed subcapacities could be arranged to demonstrate the phenomenon to be explained. I tend to agree with the mechanist criticism that this provides only a how-possibly model, not a how-actually model. In deep learning, it is very easy to write down the exact compositions of functions that are computed by a network. This by itself doesn't seem to translate into much explanatory power. Some of the simplified stories about how deep networks work resemble functional explanations based on the component operations of layers. Features learned in each layer compose hierarchically: lower-level features (edges, colour patches) combine into intermediate features (shapes, textures) which in turn combine into features that discriminate between higher-level object categories at the last layer. Compositionality is surely an important concept in the theoretical motivations for deep learning, but these over-simplified stories have not up to now been sufficient to explain how neural networks work, and works like Geirhos et al. (2019, 2020) have challenged intuitions about how networks actually use such features. Individual layers and units are not in general clearly selective to human understandable features, and even when they are, the role of this selectivity is unclear (Leavitt and Morcos, 2020). Therefore, I am uncertain about the role that functional characterization of component operations, for example the system identification approach described in Wu et al. (2006), will play in a unified theory of explanation for biological and artificial intelligence.

6. Conclusion

As scientists, we ultimately want our scientific enterprises to proceed towards explanations of our phenomena of interest. To understand how our science progresses, we must consider what constitutes a scientific explanation for the specific phenomena of interest. I reviewed several theories of scientific explanation and discussed how such theories may be applied at the intersection of neuroscience and AI. I clarified the distinction between understanding (a cognitive achievement) and scientific explanation (an answer to a why-question that may be good or bad, e.g. true or false, depending at least in part on its relationship to the phenomenon to be explained). It is the character of this relationship between an explanation and the phenomenon to be explained that the theories of scientific explanation I reviewed are concerned with. I showed how different theories draw different boundaries between what is truly explanatory and what is merely descriptive. There are several themes that recur: causality, physicality, predictability, mechanism, representation, and function. Different theories of scientific explanation reflect different valuations of these themes.

What constitutes an explanation, and hence how a science progresses, may be phenomenon-specific. Therefore, when defining a new science of intelligence at the intersection of neuroscience and AI, the first step is to clearly delineate the phenomena to be explained. Articles like Lillicrap and Kording (2019) and Richards et al. (2019) do not directly address the matter of scientific explanation. However, they do propose and justify why-questions. When defining a science at the intersection of neuroscience and AI, we can literally think of an intersection on sets of observations—phenomena that exist in both biological and artificial intelligence. As such, we need a theory of explanation that holds regardless of whether the scope of the why-question includes only AI, only brains, or both. This provides useful constraints on our conceptions of scientific explanation. These constraints can help focus and direct future efforts to theorize about how a science of intelligence might progress.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., and Brain, G. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv*, pages 1407.5104 [q–bio.NC].

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv*, page 1610.01644v3.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. (2018). Systematic Generalization: What Is Required and Can It Be Learned? pages 1–16.

Barrett, D. G., Morcos, A. S., and Macke, J. H. (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55:55–64.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439).

Batterman, R. W. (2001). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence.* Oxford University Press.

Batterman, R. W. and Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, 81(3):349–376.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, 22(3):295–317.

Bechtel, W. (2004). The Epistemology of Evidence in Cognitive Neuroscience. In Jr., R. S., Allen, C., Ankeny, R. A., Craver, C. F., Darden, L., Mikkelson, G., and Richardson, R., editors, *Philosophy and the Life Sciences: A Reader*. MIT Press, Cambridge, MA.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Taylor & Francis Group, New York, NY.

Bechtel, W. and Wright, C. D. (2009). What is psychological explanation? *The Routledge Companion to Philosophy of Psychology*, pages 113–130.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method ({CompCor}) for {BOLD} and perfusion based fMRI. *NeuroImage*, 37(1):90–101.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403:309–312.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Bengio, Y., Lee, D.-h., Bornschein, J., and Lin, Z. (2014). Towards Biologically Plausible Deep Learning.

Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Front. of Neuroinformatics*.

Brette, R. (2019). Neural coding : the bureaucratic model of the brain.

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walke, E. Y., Reimer, J., Bethge, M., Tolias, A. S., and Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Real Neurons & Hidden Units NeurIPS Workshop*.

Cadieu, C., Hong, H., and Yamins, D. L. K. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12):e1003963.

Cao, R. and Yamins, D. (2020). Making sense of mechanism : How neural network models can explain brain function.

Carlson, N. L., Ming, V. L., and Deweese, M. R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS computational biology*, 8(7):e1002594.

Carr, C. E. (1993). Processing of temporal information in the brain. *Annual review of neuroscience*, 16:223–43.

Cartwright, N. (1979). Causal Laws and Effective Strategies. *Nous*, 13(4, Special Issue on Counterfactuals and Laws).

Cartwright, N. (1983). *How the Laws of Physics Lie.* Oxford University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science.* The MIT Press, Cambridge, MA.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(January).

Cichy, R. M., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., and Oliva, A. (2019). The Algonauts Project: A Platform for Communication between the Science of Biological and Artificial Intelligence. In *Computational Cognitive Neuroscience.*

Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki.

Cox, R. W. and Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, 10(4-5):171–178.

Craver, C. F. (2002). Structures of Scientific Theories. In Machamer, P. and M. Silberstein, editors, *The Blackwell Guide to the Philosophy of Science*, chapter 4. Blackwell Publishers.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience.* Clarendon Press, Oxford.

Craver, C. F. and Kaplan, D. M. (2011). Towards a Mechanistic Philosophy of Neuroscience. In French, S. and Saatsi, J., editors, *The Continuum Companion to the Philosophy of Science*, chapter 14, pages 268–292. A&C Black.

Cummins, R. (1975). Functional Analysis. *The Journal ofPhilosophy*, 72(20):741–765.

Cummins, R. (2000). "How does it work" vs. "What are the laws?" Two conceptions of psychological explanation. *Explanation and cognition*, pages 117–144.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, 9(2):179–194.

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6):2144–9.

Dayan, P. and Abbott, L. (2005). *Theoretical neuroscience: Computational and Methematical Modeling of Neural Systems.* MIT Press.

De Martino, F., Moerel, M., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E. (2014a). Spatial organization of frequency preference and selectivity in the human inferior colliculus. In *International Society for Magnetic Resonance in Medicine*, Milan.

De Martino, F., Moerel, M., Xu, J., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E. (2014b). High-Resolution Mapping of Myeloarchitecture In Vivo: Localization of Auditory Areas in the Human Brain. *Cerebral Cortex*.

De Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford Univ Press, New York.

De Valois, K. K., De Valois, R. L., and Yund, E. W. (1979). Responses of striate cortex cells to grating and checkerboard patterns. *The Journal of Physiology*, 291:483–505.

DeCharms, R. C., Blake, D. T., and Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, 85(3):1220–1234.

Devezer, B., Nardin, L. G., Baumgaertner, B., and Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE*, 14(5):1–23.

DiCarlo, J. (2018). Reverse engineering visual intelligence. In *Fifth Annual Symposium of the Stanford Neurosciences Institute: Natural/Artifical Intelligence (Oral Presentation)*.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

Diedrichsen, J. and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, 13(4).

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing research*, 157(1-2):1–42.

Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152(October 2016):184–194.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial Examples that Fool both Human and Computer Vision.

Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science*, 19(4):676–685.

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., D., K. J., Goncalves, M., DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie,

D. J., Ye, Z., Poldrack, R. A., and Gorgolewski, K. J. (2018a). fMRIPrep. *Software*.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2018b). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., and Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supple:S102.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks.

Geirhos, R., Michaelis, C., Wichmann, F. A., Rubisch, P., Bethge, M., and Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *7th International Conference on Learning Representations, ICLR 2019*.

Gelder, T. v. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioural and Brain Sciences*, 21:615–665.

Gerven, M. v. and Bohte, S. (2017). Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Front. Comput. Neurosci.*, 11(114).

Golik, P., Tuske, Z., Schluter, R., and Ney, H. (2015). Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 26–30.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, 5:13.

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., Modat, M., Dewey, B. E., Madison, C., di Oleggio Castello, M., Clark, M. G., Dayan, M., Clark, D., Keshavan, A., Pinsard, B., Gramfort, A., Berleant, S., Nielson, D. M., Bougacha, S., Varoquaux, G., Cipollini, B., Markello, R., Rokem, A., Moloney, B., Halchenko, Y. O., Demian, W., Hanke, M., Horea, C., Kaczmarzyk, J., de Hollander, G., DuPre, E., Gillman, A., Mordom, D., Buchanan, C., Tungaraza, R., Pauli, W. M., Iqbal, S., Sikka, S., Mancini, M., Schwartz, Y., Malone, I. B., Dubois, M., Frohlich, C., Welch, D., Forbes, J., Kent, J., Watanabe, A., Cumba, C., Huntenburg, J. M., Kastman, E., Nichols, B. N., Eshaghi, A., Ginsburg, D., Schaefer, A., Acland, B., Giavasis, S., Kleesiek, J., Erickson, D., Küttner, R., Haselgrove, C., Correa, C., Ghayoor, A., Liem, F., Millman, J., Haehn, D., Lai, J., Zhou, D., Blair, R., Glatard, T., Renfro, M., Liu, S., Kahn, A. E., Pérez-García, F., Triplett, W., Lampe, L., Stadler, J., Kong, X.-Z., Hallquist, M., Chetverikov, A., Salvatore, J., Park, A., Poldrack, R. A., Craddock, R. C., Inati, S., Hinds, O., Cooper, G., Perkins, L. N., Marina, A., Mattfeld, A., Noel, M., Snoek, L., Matsubara, K., Cheung, B., Rothmei, S., Urchs, S., Durnez, J., Mertz, F., Geisler, D., Floren, A., Gerhard, S., Sharp, P., Molina-Romero, M., Weinstein, A., Broderick, W., Saase, V., Andberg, S. K., Harms, R., Schlamp, K., Arias, J., Papadopoulos Orfanos, D., Tarbert, C., Tambini, A., De La Vega, A., Nickson, T., Brett, M., Falkiewicz, M., Podranski, K., Linkersdörfer, J., Flandin, G., Ort, E., Shachnev, D., McNamee, D., Davison, A., Varada, J., Schwabacher, I., Pellman, J., Perez-Guevara, M., Khanuja, R., Pannetier, N., McDermottroe, C., and Ghosh, S. (2018). Nipype. *Software*.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. (2019). Recurrent Independent Mechanisms.

Greve, D. N. and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72.

Grimm, S. (2010). Understanding. In Berneker, S. and Pritchard, D., editors, *The Routledge Companion to Epistemology*. Routledge, New York.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). Generalized Autocalibrating Partially Parallel Acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210.

Güçlü, U., Thielen, J., Hanke, M., and van Gerven, M. A. J. (2016). Brains on Beats. *arXiv*, page 1606.02627.

Güçlü, U. and van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of Neuroscience*, 35(27):10005–10014.

Güçlü, U. and van Gerven, M. A. J. (2016). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, pages 6–13.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *eLife*, 6(e22901).

Hall, W. C. (2008). The Auditory System. In Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W., Lamantia, A.-S., McNamara, J., and White, L., editors, *Neuroscience*, pages 313–342. Sinauer Associates, Sunderland, MA, 4th edition.

Hasson, U., Nastase, S. A., and Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, 105(3):416–434.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson Prentice Hall, Upper Saddle River, New Jersey, third edition.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., and Dean, J. (2013). Multilingual Acoustic Models using Distributed Deep Neural Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8619–8623.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* The Free Press, New York.

Hinton, G. E. (2011). Machine learning for neuroscience. *Neural systems & circuits*, 1(1):12.

Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Vanhoucke, V., Nguyen, P., Sainath, T. N., Kingsbury, B., and Senior, A. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, (November):1–27.

Hitchcock, C. R. (1995). Salmon on Explanatory Relevance. *Philosophy of Science*, 62(2):304–320.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4).

Hsu, A., Woolley, S. M. N., Fremouw, T. E., and Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of Neuroscience*, 24(41):9201–11.

Humphries, C., Liebenthal, E., and Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, 50(3):1202–11.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–6.

Hyde, P. S. and Knudsen, E. I. (2000). Topographic Projection From the Optic Tectum to the Auditory Space Map in the Inferior Colliculus of the Barn Owl. *The Journal of Comparative Neurology*, 421:146–160.

Imig, T. J. and Morel, A. (1985). Tonotopic Organization in Ventral Nucleus of Medial Geniculate Body in the Cat. *Journal of Neurophysiology*, 53(1):309–40.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 2146–2153.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2):825–841.

Jonas, E. and Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*, 13(1).

Kaas, J. H. and Hackett, T. a. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11793–9.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183(3).

Kaplan, D. M. and Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science*, 3(2):438–444.

Karklin, Y. and Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural computation*, 17(2):397–423.

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, 180:101–109.

Kell, A. J. and McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55:121–132.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11):e1003915.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations (ICLR)*.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., and Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, 13(2):e1005350.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. *ICLR workshop on Debugging Machine Learning Models*.

Kowalksi, N., Depireux, D. A., and Shamma, S. A. (1996). Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex : I . Characteristics of Single Unit Responses to Moving Ripple Spectra. *Journal of Neurophysiology*, 76:3503–3523.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. in Systems Neuroscience*, 2.

Krizhevsky, A. and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

Kubilius, J. (2017). Predict, then simplify. *NeuroImage*, (October):1–2.

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1):76–85.

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research*, 60(2):115–142.

Laudanski, J., Edeline, J. M., and Huetz, C. (2012). Differences between Spectro-Temporal Receptive Fields Derived from Artificial and Natural Stimuli in the Auditory Cortex. *PLoS ONE*, 7(11).

Leaver, A. M. and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22):7604–7612.

Leaver, A. M. and Rauschecker, J. P. (2016). Functional Topography of Human Auditory Cortex. *Journal of Neuroscience*, 36(4):1416–1428.

Leavitt, M. L. and Morcos, A. S. (2020). Selectivity considered harmful : evaluating the causal impact of class selectivity in DNNs.

Leavitt, M. L., Pieper, F., Sachs, A. J., and Martinez-Trujillo, J. C. (2017). Correlated variability modifies working memory fidelity in primate prefrontal neuronal ensembles. *Proceedings of the National Academy of Science*, 114(12):E2494–E2503.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009a). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8.

Lee, H., Pham, P., Largman, Y., and Ng, A. (2009b). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems*.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–63.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2014). Random feedback weights support learning in deep neural networks. page 14.

Lillicrap, T. P. and Kording, K. P. (2019). What does it mean to understand a neural network? page arXiv preprint: 1907.06374.

Lindsay, G., Moskovitz, T., Yang, G. R., and Miller, K. (2019). Do Biologically-Realistic Recurrent Architectures Produce Biologically-Realistic Models? pages 779–782.

Lindsay, G. W. (2020). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *arXiv preprint*, page 2001.07092.

Lindsay, G. W. and Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, 7:1–29.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015). Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*, volume 37.

Love, B. C. (2019). Levels of Biological Plausibility. *PsyArXiv Preprints*.

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Towards an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94.

Marcus, G. (2018). Deep Learning: A Critical Appraisal.

Marder, E. (2015). Understanding Brains: Details, Intuition, and Big Data. *PLoS Biology*, 13(5):1–6.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company, San Fransisco.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, 200(1140):269–294.

Massoudi, R., Van Wanrooij, M. M., Versnel, H., and Van Opstal, a. J. (2015). Spectrotemporal Response Properties of Core Auditory Cortex Neurons in Awake Monkey. *Plos One*, 10:e0116118.

May, B., Prell, G. L., and Sachs, M. (1998). Vowel Representations in the Ventral Cochlear Nucleus of the Cat: Effects of Level, Background Noise, and Behavioral State. *Journal of neurophysiology*, 79:1755–1767.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of Neurophysiology*, 102(6):3329–39.

Metz, C. (2018). Google Researchers Are Learning How Machines Learn.

Middlebrooks, P. (2019). Brain Inspired podcast, episode 52, Andrew Saxe: Deep Learning Theory.

Miller, L. M., Escabí, M. A., Read, H. L., and Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, 87(1):516–527.

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(41):14205–16.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8(July):1–14.

Mole, C. and Klein, C. (2010). Confirmation, Refutation, and the Evidence of fMRI. In Hanson, S. J. and Bunzl, M., editors, *Foundational Issues in Human Brain Mapping*, chapter 9, page 99–111. MIT Press, Cambridge, MA and London, England.

Morcos, A. S., Raghu, M., and Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *NeurIPS*.

Müller, M. and Ewert, S. (2010). Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662.

Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring How Deep Neural Networks Form Phonemic Categories. *Interspeech*, pages 1912–1916.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–10.

Nemenman, I. (2018). Playing Newton: Automatic Construction of Phenomenological, Data-Driven Theories and Models.

Norman-Haignere, S., Kanwisher, N. G., and Mcdermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, 88(6):1281–1296.

O'Doherty, J. P., Hampton, A., and Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104:35–53.

Ogawa, S. and Tank, D. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(July):5951–5955.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The Building Blocks of Interpretability.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13):607–609.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. In *The 9th ISCA Speech Synthesis Workshop*.

Overath, T., McDermott, J. H., Zarate, J. M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, 18(6):903–911.

Peretz, I. and Zatorre, R. J. (2005). Brain organization for music processing. *Annual review of psychology*, 56:89–114.

Piccinini, G. and Craver, C. (2011). Integrating psychology and neuroscience : functional analyses as mechanism sketches. *Synthese*, 183(3):283–311.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84(Supplement C):320–341.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement. *NeurIPS*.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *NeurIPS*.

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207):111–114.

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., and Shea-Brown, E. (2019). Dimensionality compression and expansion in Deep Neural Networks.

Reuter, M., Rosas, H. D., and Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, 53(4):1181–1196.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Ritchie, J. B. (2020). Biological Plausibility , Mechanistic Explanation , and the Promise of “Cognitive Computational Neuroscience”.

Robert, P. and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Applied Statistics*, 25(3).

Robinson, D. A. (1992). Implications of neural network for how we think about brain function. *Behavioral and Brain Sciences*, 15(4).

Rodd, J. M., Davis, M. H., and Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral cortex (New York, N.Y. : 1991)*, 15(8):1261–9.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization. *Psychological Review*, 65(6):386–408.

Ross, L. N. (2015). Dynamical Models and Explanation in Neuroscience. *Philosophy of Science*, 82(1):32–54.

Saini, A. (2019). *Superior: The Return of Race Science.* Beacon Press, Boston.

Salmon, W. C. (1971). *Statistical explanation and statistical relevance.* University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press.

Salmon, W. C. (1989). *Four Decades of Scientific Explanation.* University of Minnesota Press, Minneapolis.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):4968–4977.

Santoro, R. (2014). *The Computational Architecture of the Human Auditory Cortex.* PhD thesis, Maastricht University.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLOS Computational Biology*, 10(1):e1003412.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64(1):240–256.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2015). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference for Learning Representations (ICLR).*

Schölkopf, B. (2019). Causality for Machine Learning. pages 1–20.

Schönwiesner, M., Dechent, P., Voit, D., Petkov, C. I., and Krumbholz, K. (2014). Parcellation of Human and Monkey Core Auditory Cortex with fMRI Pattern Classification and Objective Detection of Tonotopic Gradient Reversals. *Cerebral Cortex*, pages 1–12.

Schönwiesner, M. and Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14611–6.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, page 407007.

Sercu, T., Puhrsch, C., Kingsbury, B., and LeCun, Y. (2016). Very Deep Multilingual Convolutional Neural Networks for LVCSR. In *IEEE International Conference on Acoustics, Speech and Signal Processing.*

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, 67(5):1210–1224.

Singh, N. C. and Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394.

Sitek, K. R., Faruk Gulban, O., Calabrese, E., Johnson, G. A., Ghosh, S. S., and De Martino, F. (2019). Mapping the human subcortical auditory system using histology, post mortem MRI and in vivo MRI at 7T. *eLife*, (8:e48932).

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C. M., and Van Erk, M. J. (2009). Matrix correlations for high-dimensional data: The modified RV-coefficient. *Bioinformatics*, 25(3).

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(SUPPL. 1):208–219.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23.

Stansbury, D. E. (2014). *Modeling neural representation using statistical features of natural scenes*. PhD thesis, University of California, Berkeley.

Steen Moeller, Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. (2010). Multiband Multislice GE-EPI at 7 Tesla, With 16-Fold Acceleration Using Partial Parallel Imaging With Application to High Spatial and Temporal Whole-Brain FMRI. *Magnetic Resonance in Medicine*, 63(5).

Stepp, N., Chemero, A., and Turvey, M. T. (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science*, 3(2):425–437.

Storrs, K. R. and Kriegeskorte, N. (2020). Deep Learning for Cognitive Neuroscience. In Poeppel, D., Mangun, G. R., and Gazzaniga, M. S., editors, *The Cognitive Neurosciences*. MIT Press, 6th edition.

Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-dimensional Proportional System : an Approach to Cerebral Imaging*. Stuttgart: Thieme.

Theunissen, F. E. and Elie, J. E. (2014). Neural processing of natural sounds. *Nature reviews. Neuroscience*, 15(6):355–66.

Thompson, J. A. F., Bengio, Y., Formisano, E., and Schönwiesner, M. (2016). How can deep learning advance computational modeling of sensory information processing? *NeurIPS workshop on Representation Learning in Artificial and Biological Neural Networks.*

Thompson, J. A. F., Schönwiesner, M., Bengio, Y., and Willett, D. (2019a). How transferable are features in convolutional neural network acoustic models across languages? *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP).*

Thompson, J. A. F., Yoshua Bengio, and Schönwiesner, M. (2019b). The effect of task and training on intermediate representations in convolutional neural networks revealed with modified RV similarity analysis. In *Cognitive Computational Neuroscience*.

Tuske, Z., Pinto, J., Willett, D., and Schluter, R. (2013). Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7349–7353.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Uden, C. V. (2019). *Comparing brain-like representations learned by vanilla , residual , and recurrent CNN architectures*. PhD thesis, Dartmouth College.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, 66(5):1364–1380.

Vinje, W. E. and Gallant, J. L. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, 287(5456):1273–1276.

Watanabe, S., Hori, T., and Hershey, J. R. . (2017). LANGUAGE INDEPENDENT END-TO-END ARCHITECTURE FOR JOINT LANGUAGE IDENTIFICATION AND SPEECH RECOGNITION. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 265–271.

Woodward, J. (2002). Explanation. In Machamer, P. and Silberstein, M., editors, *The Blackwell Guide to the Philosophy of Science*, chapter 3. Blackwell Publishers Ltd.

Woodward, J. (2017). Scientific Explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Fall 2017 Edition)*, page https://plato.stanford.edu/archives/fall2017/entri. Metaphysics Research Lab, Stanford University.

Woolley, S. M. N., Gill, P. R., and Theunissen, F. E. (2006). Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(9):2499–512.

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual review of neuroscience*, 29:477–505.

Wyse, L. (2017). Audio Spectrogram Representations for Processing with Convolutional Neural Networks. In *Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN*, pages 37–41.

Xu, Y. (2020). Limited correspondence in visual representation between the human brain and convolutional neural networks.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *NeurIPS*.

Young, E. D. and Oertel, D. (1999). The Cochlear Nucleus. In Shepard, G., editor, *The Synaptic Organization of the Brain*. Oxford University Press.

Yu, D. and Deng, L. (2015). *Automatic Speech Recognition: A deep learning approach*. Signals and Communication Technology. Springer.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(3770).

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nature reviews. Neuroscience*, 8(7):547–58.

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science*, 78(2):238–263.

Zeiler, M. and Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision-ECCV 2014*.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57.

Second Article.

Encoding of mixtures of dynamic ripples in human auditory cortex

by

Jessica A. F. Thompson[1], Federico De Martino[2], Elia Formisano[3], and Marc Schönwiesner[4]

([1]) Université de Montréal

([2]) Maastricht University

([3]) Maastricht University

([4]) University of Leipzig

This article was or will be submitted in Frontiers in Auditory Cognitive Neuroscience.

My contributions and the role of the coauthors

— Preparation of all stimuli and the code to execute the experiment;

— Recruited participants and conducted the experimental sessions;

— Conduct all data preprocessing and analysis;

— Preparation of the manuscript.

Prof. De Martino provided the scanner protocols and operated the scanner. He and Prof. Formisano directed the data collection and analysis. All coauthors contributed to the experimental design.

RÉSUMÉ. Il a été démontré que les fonctions de transfert de modulation calculées à partir de sons synthétiques (ondulations dynamiques) diffèrent de celles calculées à partir de sons naturels. Ici, nous avons collecté des réponses IRMf à huit ondulations simples et les 28 mélanges par paires de ces ondulations pour étudier le codage neuronal des modulations spectro-temporelles. Une analyse de codage a été réalisée pour identifier les voxels dont la réponse pourrait être modélisée en fonction linéaire des paramètres d'ondulation: fréquence fondamentale, taux de modulation temporelle et échelle spectrale. En quantifiant la performance globale du modèle de codage, une analyse d'identification de son a révélé que le modèle de codage entraîné uniquement sur de simples ondulations était capable de se généraliser à des mélanges d'ondulations. Cependant, l'activité voxel n'a pu être prédite de manière significative que chez un sujet. Ces résultats supportent le modèle de modulation spectro-temporelle spécifique à la fréquence du cortex auditif, mais le rapport signal sur bruit était insuffisant pour localiser des voxels spécifiques où ce modèle tient.
Mots clés : IRMf, codage, ondulations dynamiques, neurosciences auditives

ABSTRACT. Modulation transfer functions calculated from synthetic sounds (dynamic ripples) have been shown to differ from those calculated from natural sounds. Here, we collected fMRI responses to eight simple ripples and the 28 pairwise mixtures of those ripples to investigate the neural encoding of spectro-temporal modulations. An encoding analysis was performed to identify voxels whose response could be modelled as a linear function of the ripple parameters: fundamental frequency, temporal modulation rate, and spectral scale. Quantifying the global performance of the encoding model, a sound identification analysis revealed that the encoding model trained only on simple ripples was able to generalize to mixtures of ripples. However, voxel activity could only be significantly predicted in one participant. These results provide evidence in support of the frequency-specific spectrotemporal modulation model of auditory cortex, but the signal-to-noise ratio was insufficient to localize specific voxels where this model holds.
Keywords: fMRI, encoding, dynamic ripples, auditory neuroscience

1. Introduction

The information processing paradigm in auditory neuroscience views the auditory pathway as a sequences of processing stages that successively transform sound stimuli to support auditory perception and cognition. A popular approach is to characterize the response profiles of neurons or populations of neurons along the auditory pathway, i.e. to characterize the neural representation of sound. This approach, combined with physiological studies, have elucidated the hierarchical structure of the auditory pathway, including auditory cortex, in non-human animals. In the monkey, the AC is organized hierarchically into several primary or *core* areas (A1) that receive inputs from the thalamus and are surrounded by non-primary or *belt* and *parabelt* regions (Kaas and Hackett, 2000; Rauschecker et al., 1995). While it is generally agreed upon that the human auditory cortex is also arranged hierarchically with information passing from primary to secondary areas, the details of this model break down in the human because of considerable anatomical differences.

Auditory neurons respond preferentially to specific patterns of frequencies over time, i.e. neurons have spectro-temporal receptive fields (STRFs) (DeCharms et al., 1998).

Physiological and psychoacoustic studies suggest that the cortical representation of sound involves the explicit encoding of spectral and temporal modulations through dedicated modulation-detectors (Viemeister, 1979; Depireux et al., 2001; Miller et al., 2002). Consequently, STRFs in auditory cortex are often parameterized by modulations in both frequency (temporal modulation rate) and time (spectral scale). One can define a modulation transfer function (MTF) which describes a neuron's (or a voxel's) response as a function of temporal modulation rate and spectral scale. The two-dimension Fourier transform of an STRF gives the corresponding MTF. MTFs have been measured using the dynamic ripples as stimuli. A dynamic ripple is a complex broadband sound with a sinusoidal spectral envelope that drifts along the logarithmic frequency axis over time. The dynamic ripple is defined by the expression $S(t,x) = 1 + A\sin(2\pi(\omega t + \Omega x) + \phi)$ where t indexes time, x indexes spectral frequency, A is the modulation depth, ϕ is the phase, ω is the temporal modulation rate (in Hz), and Ω is the spectral scale (in cycles/octave). Thus, neural responses to dynamic ripples correspond to specific regions in an $f0$-specific MTF. Dynamic ripples have been used to calculate reliable MTFs in human (Schönwiesner and Zatorre, 2009) and mammalian (Kowalksi et al., 1996) auditory cortex. The working hypothesis is that auditory cortex decomposes sound into acoustic features akin to these dynamic ripples, which can support the multiresolution spectrotemporal analysis required for most complex auditory tasks (Chi et al., 2005; Massoudi et al., 2015; Leaver and Rauschecker, 2016)

Ultimately, auditory neuroscience seeks to understand how the brain represents natural, meaningful sounds like communication sounds and complex auditory scenes, not only simplified synthetic sounds. MTFs and STRFs can also be calculated using natural sounds as stimuli via reverse correlation, sometimes leading to different results than when using dynamic ripples. In the guinea pig, Laudanski et al. (2012) compared STRFs obtained from the same A1 neurons using dynamic ripples and conspecific vocalizations. They found that the best frequency, bandwidth (frequency range), latency (time to peak) and global shape of the STRFs depended on the stimulus type. Additionally, they found that neural responses to vocalizations were better predicted by STRFs calculated from vocalizations than neural responses to ripples were predicted by STRFs calculated from ripples. Stimulus-dependent tuning was also observed in the songbird, where 91% of midbrain neurons showed differences in STRFs calculated from song and acoustically matched noise (Woolley et al., 2006). Other studies have similarly shown that natural sounds elicit stronger and more reliable responses than synthetic stimuli (Theunissen and Elie, 2014; Singh and Theunissen, 2003; Hsu et al., 2004). Context-dependent auditory tuning has also been observed with synthetic sounds. David et al. (2012) showed that the task reward structure, whether to approach or avoid a target, influenced STRFs even though the sensory discrimination required was identical in both conditions. One interpretation of these results is that there are important nonlinearities in auditory cortical responses that are not captured by the simple linear model based on spectrotemporal modulation.

On the other hand, additional studies on the neural representation of natural sounds in the human brain have found spectrotemporal modulation-based acoustic features to be highly predictive of fMRI activity in auditory cortex. Leaver and Rauschecker (2010) found that voxels close to A1 responded selectively to spectral and temporal modulations, rather than to the stimulus category. Santoro et al. (2014) found that a joint frequency-specific MTF-based model predicted voxel activity better than alternative models. They proposed that posterior/dorsal auditory regions encode coarse spectral information with high temporal precision while anterior/ventral regions prefer find-grained spectral information with low temporal precision. Thus, MTFs calculated from natural sounds appear to generalize to responses to other natural sounds.

The response of any auditory neuron or voxel can be thought of consisting of a feed-forward, bottom-up component and a feedback, top-down component. Here we consider the hypothesis that there exist some voxels in auditory cortex for which the joint frequency-specific MTF model (as described in Santoro et al. (2014)) accurately captures the feed-forward component of the response. The aforementioned context and stimulus dependent responses are hypothesized to belong to the feedback component and to be a product of attentional effects related to the semantic or practical significance of the sounds. To investigate this hypothesis, we measured responses to two classes of stimuli: simple dynamic moving ripples, as previously defined, and mixtures of pairs of dynamic ripples, where two dynamic ripples are combined in a single stimulus. Our experiment is designed such that we don't expect any differences in the feedback component of responses to these sounds; there are no task-related differences between the stimuli and they have no inherent meaning. We predict that voxels nearest to A1 should show no stimulus-dependent effects; responses to simple ripples ought to generalize to responses to mixtures of ripples. Where there are nonlinearities in the voxel responses that are not captured by the simple linear model, for example, multi-peak tuning where voxels respond only to the presence of two fundamental frequencies, not to each presented separately, we expect to see stimulus-dependent responses where the responses to mixtures of ripples cannot be modelled as linear combination of the responses to simple ripples.

2. Methods and Materials

2.1. Participants

Six healthy participants (aged 26–33, 3 women) with normal hearing and no known neurological disorders were recruited to participate. All participants provided written informed consent prior to the experiment.

2.2. Stimuli

Stimuli consisted of simple dynamic ripples and simultaneous combinations (sums) of pairs of dynamic ripples. Stimuli were generated using the NSL MATLAB toolbox (available at http://www.isr.umd.edu/Labs/NSL/Software.htm) and customized MATLAB code (The MathWorks Inc.). All stimuli were one second long. Ripples were generated according to a 2 fundamental frequencies x 2 spectral scales x 2 temporal rates design, resulting in eight simple ripples and 28 unique pairwise combinations of these 8 ripples for a total of 36 distinct one-second stimuli. The parameter values of the simple ripples were selected based on several criteria. We chose fundamental frequencies $f0_1 = 132.5$ Hz and $f0_2 = 210$ Hz based on the average $f0$ of the male and female speech. Spectral scales $\Omega_1 = 0.7$ cyc/oct and $\Omega_2 = 1.7$ cyc/oct and temporal rates $\omega_1 = 2$ Hz and $\omega_2 = 6$ Hz were chosen to be within the peaks of the marginal spectral and temporal MTFs described in Santoro et al. (2014) while still being distinct enough to activate spatially distinct regions. Pair-wise ripple combinations with ratios 1:1, 1:2, and 2:1 were constructed as the weighted sum of the audio waveforms, resulting in 84 distinct mixtures. All stimuli were normalized to have equal intensity. For the subsequent analyses, the stimuli were represented by real-valued, three-dimensional feature vectors: one element for each stimulus attribute $f0$, Ω, and ω. Using $f0$ as an example, a value of 0 indicated that only $f0_1$ was present, a value of 1 indicated that only $f0_2$ was present, a value of 0.5 indicated that both $f0_1$ and $f0_2$ were present to equal degrees, a value of $\frac{1}{3}$ indicated that $f0_2$ was twice as intense as $f0_1$ in the mixture, and a value of $\frac{2}{3}$ indicated that $f0_1$ was twice as intense as $f0_2$ in the mixture. The stimulus attributes Ω and ω were similarly represented.

2.3. Magnetic Resonance Imaging Parameters

MRI data were acquired on the 7T magnetic resonance system at scannexus (www.scannexus.nl, Maastricht, The Netherlands). A Nova Medical RF head coil (single transmit, 32 receive channels) was used to acquire anatomical (T1-weighted) and functional (T2*-weighted BOLD) images. Anatomical T1-weighted images were acquired using a Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE) sequence (TR = 3100 ms; TI = 1500 ms; flip angle = 5 degrees; voxel size = 0.6 x 0.6 x 0.6 mm^3). Proton density weighted (PD-weighted) images were also acquired (TR = 1440 ms; voxel size = 0.6 x 0.6 x 0.6 mm^3). Acquisition time for anatomy was approximately 10 minutes. T2*-weighted functional data were acquired using a clustered echo planar imaging (EPI) sequence in which time gaps were placed after the acquisition of each volume (TR = 2600 ms; TA = 1200 ms; TE = 19 ms; GRAPPA acceleration X2; partial Fourier = 6/8; voxel size = 1.2 x 1.2 x 1.2 mm^3; silent gap = 1400 ms). Slices covered the brain transversally from the inferior portion of the anterior temporal pole to the superior portion of the STG bilaterally. After the anatomical acquisitions and before the beginning of the first experimental run, several reference scans with reverse phase encoding directions (posterior-anterior (PA) and anterior-posterior (AP)) were acquired to be used for distortion correction.

2.4. Protocol

The experimental procedures were approved by the ethics committee of the Faculty for Psychology and Neuroscience at Maastricht University. The experiment followed a fast event-related design where stimuli were presented during the silent gap between volume acquisitions. The experiment consisted of twelve runs: six simple runs including all eight simple ripples and six mixed runs including only the mixtures of pairs of ripples. Each simple run consisted of 48 trials and lasted 6.5 minutes. Each mixed run consisted of 52 trials and lasted 7.3 minutes. To maintain vigilance, every run included six catch trials. On catch trials, a short (0.5 s) dummy ripple was presented. The participants were asked to press a button when they heard a sound whose duration was shorter than the others. Catch trials were excluded from the analysis. Each run also included six silent trials where no sound was presented. The full set of 84 ripples mixtures was presented over two consecutive mixed runs. Thus, each mixed ripple was repeated exactly three times while each simple ripple was repeated 27 times. We chose the number of repetitions to be a multiple of three to allow us to evenly distribute repetitions across our three jitter values of two, three, and four TRs. The order of presentation in simple runs was designed to ensure that the distribution of simple ripples was equal across runs. In all runs, no two catch trials or silent trials were ever presented consecutively and no run ever began with a catch or silent trial.

2.5. MRI data pre-processing

Functional and anatomical images were preprocessed and analyzed in BrainVoyager QX 2.8.2 (Brain Innovation) and MATLAB 8.3 (R2014a) using the NeuroElf toolbox. Functional runs were 3D motion corrected and coregistered with the first AP reference volume through rigid-body transformation (3 translation and 3 rotation parameters). Preprocessing also included slice time correction with sinc interpolation and high pass temporal filtering with a cutoff of 6 cycles per run. The 'topup' method as implemented in FSL (Smith et al., 2004) was used to correct image distortions using the two reverse polarity reference scans (AP and PA) which were collected before the first run. Both anatomical and functional images were normalized to Talairach space (Talairach and Tournoux, 1988). The border between gray and white matter was segmented from anatomical volumes and used to generate cortical surface meshes of the individual participants. A broad mask of auditory cortex was drawn to include the superior temporal plane (HG, PP, PT) and the STG. Only voxels within this mask of auditory cortex were analyzed.

2.6. Response Estimation and Voxel Selection

The process for estimating voxel responses to the stimuli and selecting voxels for the subsequent analyses was the same as has been reported previously in Moerel et al. (2012). First, voxel-specific hemodynamic response function (HRF) were estimated via deconvolution with all stimuli treated as a single condition. Then, we performed a General Linear Model (GLM) analysis with one predictor per unique stimulus to compute one β per sound. This β serves as the response of an individual voxel to an individual sound. Only those voxels that showed a significant response to sound ($F > 4$) were included in the subsequent analyses.

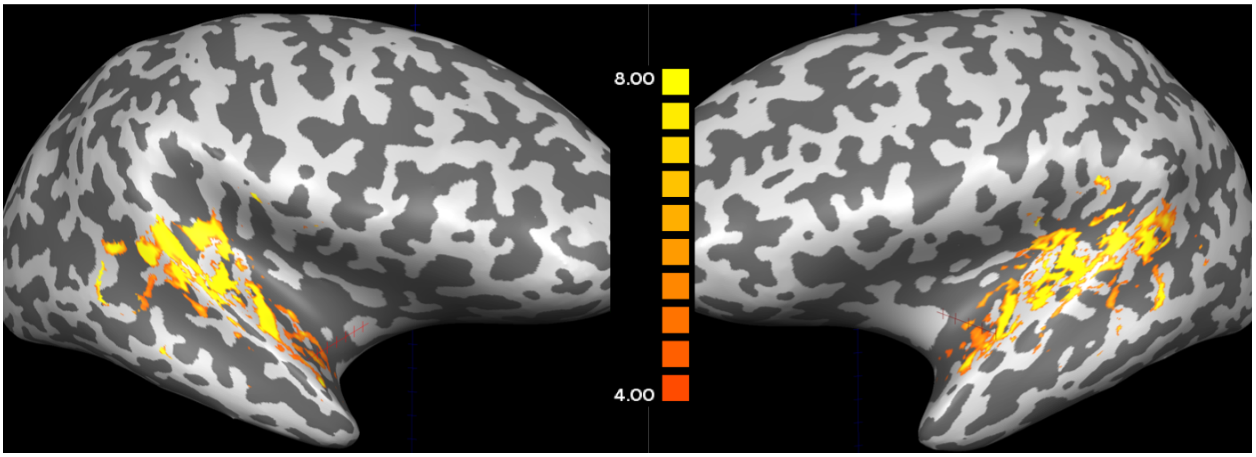


Figure 15. Response to Sound in Participant 3 Voxels within a broad anatomical mask and which showed a significant response to sound were included in the subsequent analyses. The color indicates the F-statistic of a sound vs rest contrast.

2.7. Predictive Analyses

All analyses were performed with MATLAB 8.3 (R2014a) and an in-house toolbox for fMRI encoding/decoding.

2.7.1. *Decoding Analysis*

As an omnibus tests of the stimulus-related information that could be linearly decoded from the activity of sound-responsive voxels, a multivariate decoding analysis was performed in each participant separately. Ridge regression was used to learn a linear relationship between voxel responses and stimulus attributes. Ridge regression minimizes the following cost function

$$C = \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|^2 \tag{2.1}$$

where, i indexes the stimulus and $\mathbf{w}$ are the learned weights. The target y_i and the features $\mathbf{x_i}$ correspond to a particular stimulus attribute and the vector of all voxel responses respectively. As the number of voxels may be large, the regularization term $\lambda\|\mathbf{w}\|^2$ helps to reduce overfitting. The hyperparmeter λ was selected with generalized cross validation. The weights $\mathbf{w}$, learned from the training set, were used to predict the stimulus attributes for each stimulus in the test set.

Only simple ripple runs were used for training and only mixed ripple runs were used for testing. The target variables were the stimulus attributes as described in section 2.2. Decoding performance was calculated as the Pearson correlation between the predicted feature values and the true feature values of all test sounds. Statistical significance was calculated using permutation tests. Participant-specific null distributions were calculated by permuting the labels of the training data 1000 times and recording the resulting decoding accuracy when training on permuted labels. A group null distribution was constructed as the average of the single participant null distributions. Permutation-based p-values were estimated according to the following equation

$$\hat{p} = \frac{1 + \sum_{i=1}^{M} I(r_i \geq r^*)}{M + 1} \tag{2.2}$$

where M is the number of permutations, r_i is the Pearson correlation achieved on the i^{th} permutation and r^* is the observed Pearson correlation (Ernst, 2004). Since we tested the decoding performance of every participant and the group average on every feature (21 tests), we used Bonferoni correction to correct for multiple comparisons. This omnibus test verifies the suitability of our experimental design to elicit distinctive patterns of activation in response to our stimuli.

2.7.2. *Encoding Analysis*

An encoding analysis was performed to identify voxels whose response could be modelled as a linear function of the ripple parameters: $f0$, Ω, and ω. As this encoding model involved only three orthogonal features, we used ordinary least squares (OLS) regression. The weights $\mathbf{w}$, learned from the training set, were used to predict the neural responses for each stimulus in the test set.

The 12 runs were partitioned into five different training and test sets to assess various types of generalization:

— Train Simple: Train on all simple runs, test on all mixed runs

— Train Mixed: Train on all mixed runs, test on all simple runs

— Even: Train and test on even distribution of simple and mixed runs

— Simple Only: Train on 5 simple runs, test on 1 held out simple run

— Mixed Only: Train on 3 mixed runs, test on 3 held out mixed runs

The Train Simple and Train Mixed configurations allowed us to test the hypothesis that responses to ripple mixtures are a product of the same data generating process as responses to simple ripples. If auditory cortex uses a simple linear decomposition of spectrotemporal modulation to represent both types of stimuli, then a model trained on responses to only one type of stimuli should be able to predict responses to the other type. The Even, Simple Only, and Mixed Only configurations provided control conditions to quantify the predictive performance when the training and test sets did not differ in their distribution of simple and mixed runs. In other words, the control conditions are constructed to enforce that the regression problem is one of within-distribution generalization whereas the Train Simple and Train Mixed configurations could require out-of-distribution generalization if auditory cortex doesn't represent simple and mixed ripples in the same way. Therefore, superior predictive performance for the control configurations relative to the Train Simple and Train Mixed configurations would help falsify the hypothesis that auditory cortex responds similarly to simple ripples and ripple mixtures.

Model performance per voxel was calculated as the Pearson correlation between the true and the predicted voxel to the test sounds. Permutation tests were used to assess statistical significance and correct for multiple comparisons. The analysis was repeated with 1000 permutations of the training labels to construct an empirical null distribution per voxel. Voxel-wise p-values were estimated according to the same equation as used for the decoding analysis. To correct for multiple comparisons, we calculated a cluster size threshold from the spatial maps that resulted when using permuted labels. Unlike correction for multiple comparison based on maximum statistics, this method takes into account the spatial correlation of fMRI activity. Nearby locations in cortex tend to respond similarly, therefore a large cluster (relative to what would be expected by chance) of significant voxels is less likely to be spurious.

2.7.3. *Sound Identification Analysis*

A stimulus identification analysis was used to quantify the global prediction accuracy of the Train Simple encoding model. For each test stimulus s_i, the true response patterns to all N test stimuli $\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_N$ were ranked based on their Euclidean distance to the predicted response pattern $\hat{\mathbf{y}}_i$. The rank of the true response pattern $\mathbf{y}_i$ was normalized to assign a score between 0 and 1 to each test sound. The average score over test stimuli was taken as the sound identification score for that participant. This analysis was repeated 1000 times with randomly permuted labels to calculate an empirical null distribution to assess statistical significance at the single participant level. Permutation-based p-values were estimated according to the same equation as used for the decoding and encoding analyses.

We additionally investigated whether sound identification performance (quantified as the normalized sound identification score) was better than chance at the group level.

3. Results

3.1. Decoding Analysis

The decoding results are summarized in Table 1 and Figure 16. At the single participant level, four out of six participants achieved significant decoding of $f0$ and temporal modulation rate, while spectral scale was only significantly decoded in one participant. At the group level, Only $f0$ and temporal modulation rate were significantly decoded. The highest correlation coefficient was achieved for $f0$ while spectral scale was the least well decoded. This result could reflect that our experimental design and/or stimulus design was insufficient to elicit distinctive responses to the two spectral scales. The decoding performance mirrors the perceptual salience of the stimulus attributes: differences in $f0$ are more salient than differences in temporal modulation rate and spectral scale. Larger, more salient differences in spectral scale might have elicited more decodable responses. Alternatively, these results could also reflect that the cortical representation of $f0$ is less stimulus-dependent compared to the other stimulus attributes. The cortical representation of spectral scale may depend on whether the stimulus is a simple or mixed ripple, and hence the model trained on simple ripples is unable to decode the spectral scale of mixed ripples.

Table 1. Decoding Results. Bold values indicate p-values less than $\alpha = 0.002$ which is equal to 0.05 divided by the number of tests. The smallest possible p-value that can be obtained with our procedure is $1/(M+1) = 0.001$.

Participant	f0		Temporal Rate		Spectral Scale	
	r	p	r	p	r	p
1	**0.135**	**0.001**	0.002	0.813	0.006	0.843
2	**0.405**	**0.001**	**0.254**	**0.001**	0.098	0.009
3	**0.605**	**0.001**	**0.244**	**0.001**	0.077	0.059
4	0.006	0.723	**0.138**	**0.001**	**0.205**	**0.001**
5	**0.375**	**0.001**	0.051	0.219	0.097	0.009
6	**0.348**	**0.001**	**0.396**	**0.001**	0.106	0.004
group mean	**0.312**	**0.001**	**0.180**	**0.001**	0.096	0.010

3.2. Encoding Analysis

Encoding performance was only better than chance in participant 3. Given the decoding results for participant 3, namely the superior decoding performance for $f0$, we can presume that these encoding results are largely driven by the response to $f0$. Significant voxels for all train-test configurations concentrated in the first transverse sulcus of the right hemisphere

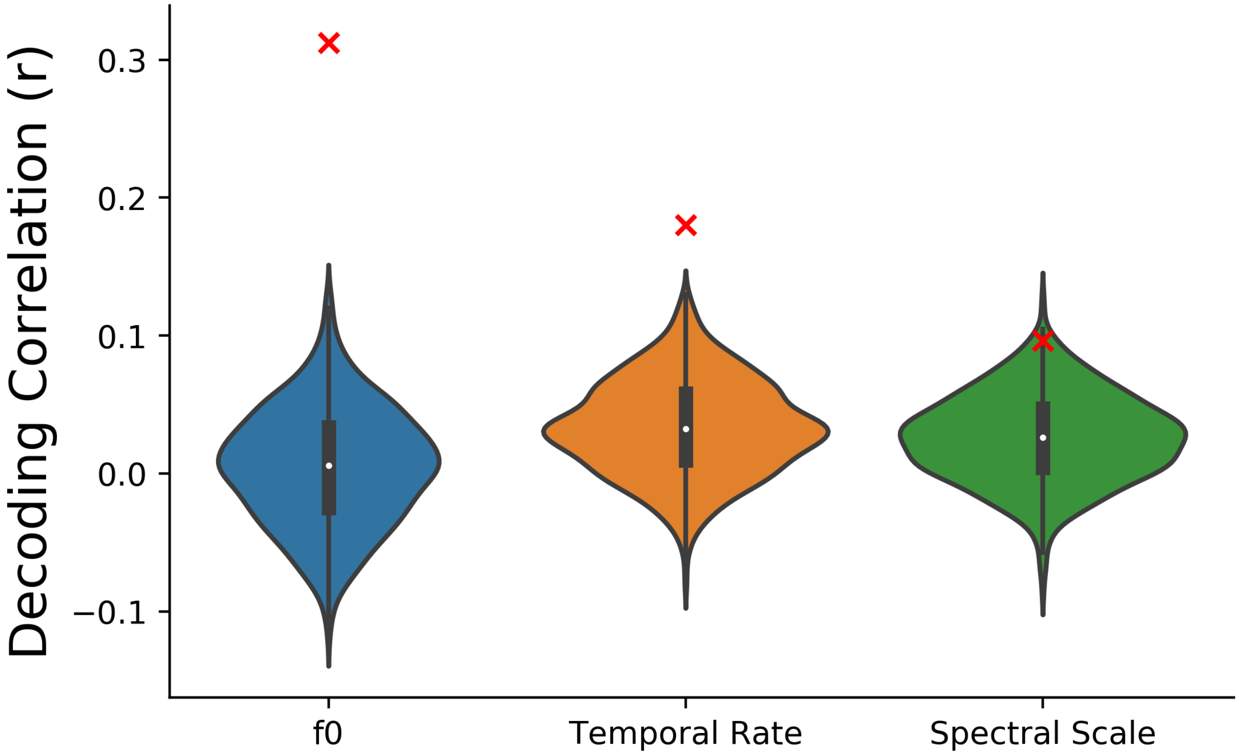


Figure 16. Average Decoding Results At the group level, significant decoding was achieved for $f0$ and temporal modulation rate but not for spectral scale. $f0$ was decoded best while spectral scale was least well decoded. Violin plots show a kernel density estimation of the average null distributions calculated by permuting the labels and retraining the decoding model 1000 times for each subject. The the white dot in the miniature boxplot indicate the median and the box shows the interquartile range. Red X's show the observed Pearson correlation for each decoding model.

(Figure 17). No voxels were significantly predicted in the Mixed Only configuration, perhaps due to the small amount of training data available (3 runs only, compared to 5 or 6 runs in the other configurations). Clusters of significant voxels across the different train-test configurations were highly overlapping, aside from a cluster of voxels in the same region of the left hemisphere that was only significantly predicted in the Train Simple configuration. Therefore, we find no evidence that simple ripples are represented differently than mixtures of ripples in auditory cortex. Since we presume that these results are largely driven by the response to $f0$, these results are consistent with the hypothesis that the response to $f0$ is not dependent on whether one or two frequencies are present in the stimulus.

3.3. Sound Identification Analysis

At the single participant level, the sound identification scores were significantly better than chance in four out of six participants (Figure 18). The magnitude of this difference

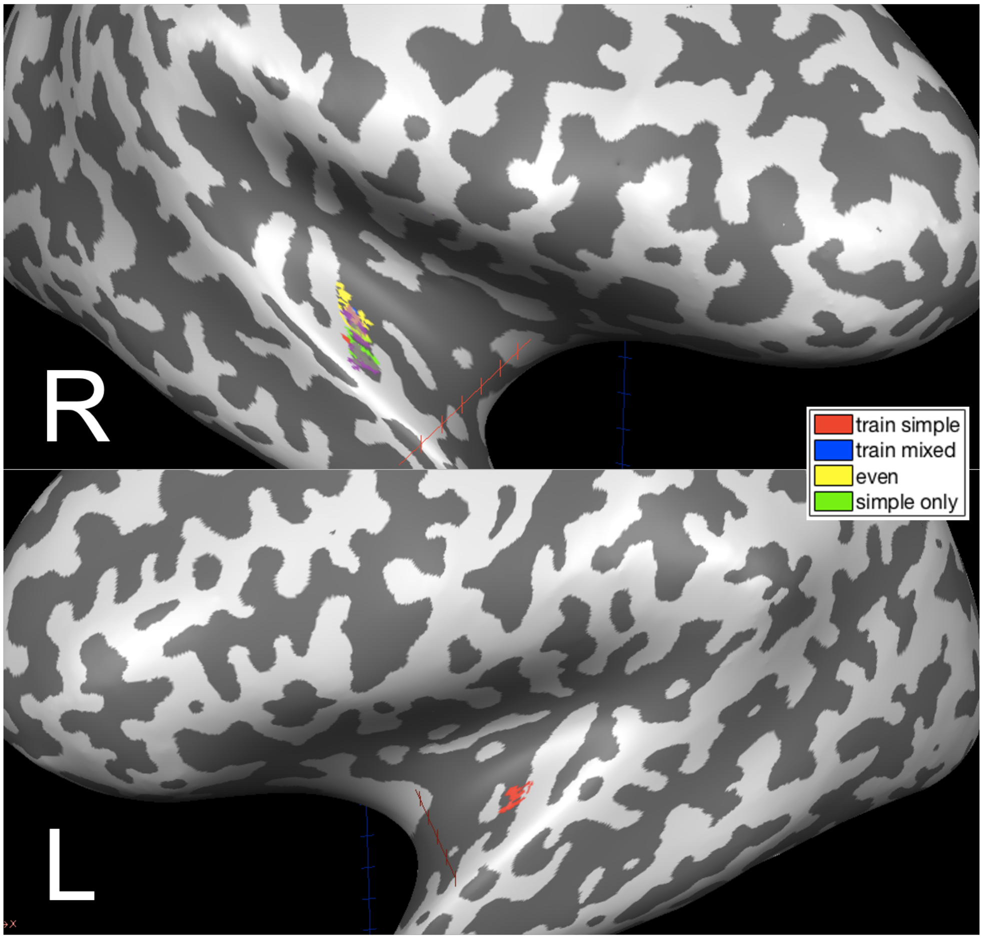


Figure 17. Encoding Results for Participant 3. Significant voxels concentrate in the first transverse sulcus of the right hemisphere. Colors indicate the partition of simple and mixed runs in to training and test sets. Train simple means the model was trained on all simple runs and tested on all mixed runs. Train mixed means all mixed runs were used for training and all simple runs were used for testing. Even means that the train and test sets both contained equal numbers of mixed and simple runs. Simple only means the model was trained and tested on only simple runs. Approximately the same cluster is found for all train-test configurations.

is small with all sound identification scores less than 0.52 and the null distributions centered around 0.5. However, this difference is significant at the group level (mean r=0.508, p=0.001). These results show that, on average, the representation of simple ripples in auditory cortex is sufficiently similar to that of mixtures of ripples that an encoding model trained only on responses to simple ripples can identify mixtures of ripples better than chance.

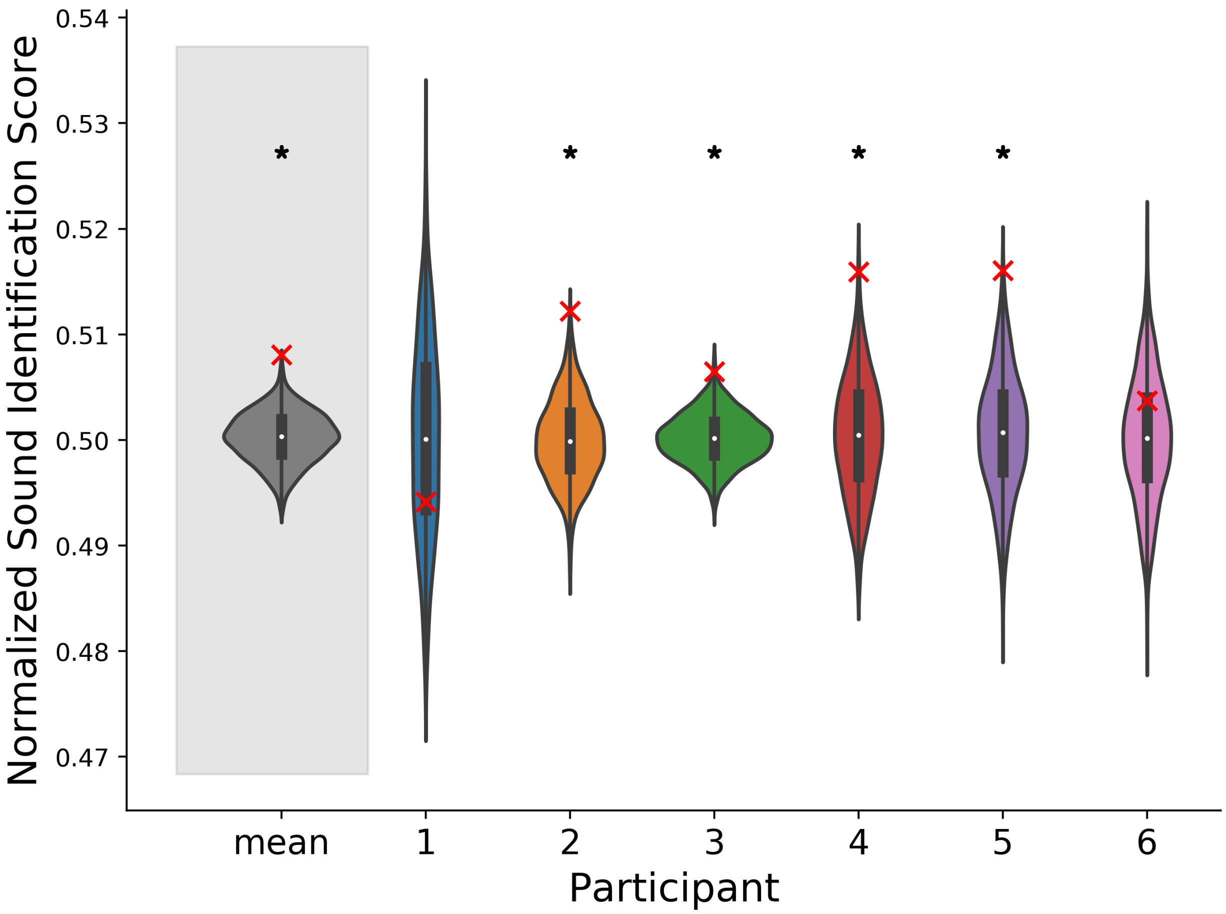


Figure 18. Permutation Tests for Sound Identification Analysis Permutation tests were conducted for each participant. Four out of six participants achieved a sound identification score significantly greater than the empirical null model, indicated by the asterisks. Red X's indicate the true sound identification score. The gray distribution shows the mean sound identification score and null distribution over participants. The violin plots features a kernel density estimation of the underlying distribution. The the white dot in the miniature box plot indicate the median and the box shows the interquartile range.

4. Discussion

Our group-level analyses show that, on average, a linear model trained only on responses to simple ripples can generalize to (make predictions about) responses to mixtures of ripples, suggesting that the neural response to simple ripples is not completely different from the response to ripple mixtures. The relationship between cortical activity and our three ripple parameters ($f0$, ω, and Ω) did not appear to be completely stimulus-dependent. However, the poor encoding performance at the single participant level and poor decoding of spectral scale complicates the interpretation. Focusing on the group level sound identification results, one may conclude that we've provided evidence in support of our hypothesis that the joint frequency-specific MTF model is a good model of the feed-forward component of the A1 response because any stimulus-dependency did not impede significant sound identification. In the encoding analysis, voxel activity was only significantly predicted in one participant. Since the significant voxels were largely overlapping for the different train-test configurations, we failed to identify separate regions of AC where responses to ripple mixtures could be modelled as a linear combination of the response to other ripple mixtures but not to simple ripples. Therefore, we found no evidence of voxels whose response contained non-linearities not captured by the joint frequency-specific MTF model.

On the other hand, we did not find strong evidence in support of the joint frequency specific MTF model either. Our sound identification scores are relatively low compared to those reported in Santoro et al. (2014) and the decoding results suggest that our predictions may be primarily driven by responses to $f0$. The lower sound identification scores likely reflect that our stimuli were significantly less distinct and elicited smaller magnitude responses than the natural sound clips used in Santoro et al. (2014). Although Santoro et al. (2014) do compare to a frequency-only model (tonotopy), they do not assess temporal modulation rate and spectral scale separately. Therefore, we cannot know how much spectral scale contributed to their sound identification performance.

The poor encoding performance, even in the Even train-test configuration, suggests that the SNR was too low to localize the voxels that facilitated the generalization we observed. To accomplish our localization goals, ideally we would have found significant encoding performance in all participants (except for participant 1 for whom only two thirds of the data was collected). Since the number of brain volumes that can be acquired in a single fMRI session is low, in predictive analyses, usually the average performance over cross validation folds is reported so that all of the data can be used during training. In the train-test configurations used here, the amount of training data varies and is typically only half of the recorded data. This severely limits the ability to train a good predictive model. Additionally, varying the amount of training data confounds the results as we expect models trained on more data to perform better than models trained on less data. Participant 1, for whom we collected only 8 runs instead of 12, achieves the poorest predictive performance. Comparison of participant 1 to the other participants can help us to understand how much data is needed for such analyses.

That synthetic stimuli like dynamic ripples elicit smaller magnitude responses make such analyses more difficult. Especially to a naive listener, our stimuli were not very perceptually distinguishable. More perceptually distinct stimuli may elicit more distinguishable patterns of activity. While our ripple parameters were well-motivated by the marginal transfer functions reported in Santoro (2014), these marginal MTFs may differ when using synthetic sounds.

Recent work has shown that fMRI responses to natural sounds are better predicted from features extracted from a task-optimized neural network than by an MTF-based model (Kell et al., 2018). In fact, they found that features extracted from a randomly initialized network predicted activity nearly as well as their MTF model. This line of work provides additional evidence that the response in AC cannot be easily captured as a linear function of the spectrogram. However, it remains unclear where to attribute these non-linearities. Are they a product of the feedforward, bottom-up computation, or are they a product of the top-down, feedback activity which contains information about the meaning of the stimulus?

The goal of the present study, to focus on the feed-forward processing stream, absent any stimulus-dependent top down influences, is one that may require synthetic stimuli, composed to be as meaningless as possible. However, escaping top-down influences may be impossible without some method to interrupt the feedback component. For example, it may be possible with TMS to disrupt the top-down activity, allowing researchers to focus on the bottom-up pathway. An alternative approach would be to build and analyze recurrent, non-linear models to better understand the possible interplay between bottom-up and top-town components.

5. Conclusion

This study provided an opportunity to falsify the popular MTF model of auditory cortex at particular voxels in auditory cortex. We had hoped to localize where this model is predictive and where it is not, but the SNR appears to have been insufficient for this analysis. At a global (AC-wide) scale, had we found that models trained on simple ripples cannot generalize to mixtures of ripples, this would have provided general evidence against the MTF model. Instead, we found that the relationship between cortical activity and each of our ripple parameters is not completely stimulus dependent, which provides general evidence in support of the MTF model. It seems clear from the literature that AC responses contain non-linearities that are not captured by the joint frequency-specific MTF model. However it remains unclear how much of this non-linearity stems from bottom-up or top-down influences.

Bibliography

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887.

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6):2144–9.

DeCharms, R. C., Blake, D. T., and Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, 85(3):1220–1234.

Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science*, 19(4):676–685.

Hsu, A., Woolley, S. M. N., Fremouw, T. E., and Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of Neuroscience*, 24(41):9201–11.

Kaas, J. H. and Hackett, T. a. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11793–9.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.

Kowalksi, N., Depireux, D. A., and Shamma, S. A. (1996). Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex : I . Characteristics of Single Unit Responses to Moving Ripple Spectra. *Journal of Neurophysiology*, 76:3503–3523.

Laudanski, J., Edeline, J. M., and Huetz, C. (2012). Differences between Spectro-Temporal Receptive Fields Derived from Artificial and Natural Stimuli in the Auditory Cortex. *PLoS ONE*, 7(11).

Leaver, A. M. and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22):7604–7612.

Leaver, A. M. and Rauschecker, J. P. (2016). Functional Topography of Human Auditory Cortex. *Journal of Neuroscience*, 36(4):1416–1428.

Massoudi, R., Van Wanrooij, M. M., Versnel, H., and Van Opstal, a. J. (2015). Spectrotemporal Response Properties of Core Auditory Cortex Neurons in Awake Monkey. *Plos One*, 10:e0116118.

Miller, L. M., Escabí, M. A., Read, H. L., and Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, 87(1):516–527.

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(41):14205–16.

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207):111–114.

Santoro, R. (2014). *The Computational Architecture of the Human Auditory Cortex*. PhD thesis, Maastricht University.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLOS Computational Biology*, 10(1):e1003412.

Schönwiesner, M. and Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14611–6.

Singh, N. C. and Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(SUPPL. 1):208–219.

Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-dimensional Proportional System : an Approach to Cerebral Imaging*. Stuttgart: Thieme.

Theunissen, F. E. and Elie, J. E. (2014). Neural processing of natural sounds. *Nature reviews. Neuroscience*, 15(6):355–66.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, 66(5):1364–1380.

Woolley, S. M. N., Gill, P. R., and Theunissen, F. E. (2006). Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(9):2499–512.

Third Article.

How transferable are features in convolutional neural network acoustic models across languages?

by

Jessica A. F. Thompson[1], Marc Schönwiesner[2], Yoshua Bengio[3], and Daniel Willett[4]

([1]) Université de Montréal
([2]) Universität Leipzig
([3]) Université de Montréal
([4]) Nuance Communications

This article was or will be submitted in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019).

My contributions and the role of the coauthors

— Conception of all the ideas;
— Conduct all experiments;
— Preparation of the manuscript.

Dr. Willett coordinated access to the data and computational resources. Prof. Bengio and Prof. Schönwiesner assisted with the interpretation of the results. All coauthors helped revise the manuscript.

Résumé. La caractérisation des représentations apprises dans les couches intermédiaires des réseaux profonds peut fournir des informations précieuses sur la nature d'une tâche et peut guider le développement de stratégies d'apprentissage bien adaptées. Nous étudions ici des modèles acoustiques basés sur un réseau neuronal convolutif (CNN) dans le contexte de la reconnaissance automatique de la parole. En adaptant une méthode proposée par Yosinski et al. (2014), nous mesurons la transférabilité de chaque couche entre l'anglais, le néerlandais et l'allemand pour évaluer leur spécificité linguistique. Nous avons observé trois régions distinctes de transférabilité: (1) les deux premières couches étaient entièrement transférables entre les langues, (2) les couches 2 à 8 étaient également hautement transférables mais nous avons trouvé des preuves de spécificité linguistique, (3) les couches suivantes entièrement connectées étaient plus spécifiques à la langue mais pouvaient être adaptées avec succès sur la langue cible. Pour étudier l'effet de l'immobilisation du poids, nous avons effectué des expériences de suivi en utilisant l'entraînement figé (Raghu et al., 2017). Nos résultats sont cohérents avec l'observation selon laquelle les CNN convergent «de bas en haut» pendant l'entraînement et démontrent les avantages de l'entraînement figé, en particulier pour l'apprentissage par transfert.
Mots clés : CNNs, modélisation acoustique, interprétabilité, transférer l'apprentissage, spécificité linguistique, entraînement figé

Abstract. Characterization of the representations learned in intermediate layers of deep networks can provide valuable insight into the nature of a task and can guide the development of well-tailored learning strategies. Here we study convolutional neural network (CNN)-based acoustic models in the context of automatic speech recognition. Adapting a method proposed by Yosinski et al. (2014), we measure the transferability of each layer between English, Dutch and German to assess their language-specificity. We observed three distinct regions of transferability: (1) the first two layers were entirely transferable between languages, (2) layers 2–8 were also highly transferable but we found some evidence of language specificity, (3) the subsequent fully connected layers were more language specific but could be successfully finetuned to the target language. To further probe the effect of weight freezing, we performed follow-up experiments using freeze-training (Raghu et al., 2017). Our results are consistent with the observation that CNNs converge 'bottom up' during training and demonstrate the benefit of freeze training, especially for transfer learning.
Keywords: CNNs, acoustic modeling, interpretability, transfer learning, language-specificity, freeze training

1. Introduction

The acoustic properties of speech vary across languages. This is evidenced by the fact that monolingual acoustic models (AMs) are the de facto standard in automatic speech recognition (ASR), while multi-lingual AMs are an active area of development (Heigold et al., 2013; Tuske et al., 2013; Sercu et al., 2016; Watanabe et al., 2017). Requiring large amounts of training data to build separate AMs for every language is a barrier to successful ASR systems for low-resource languages. Ideally, AMs would be designed to strategically leverage off-task data as much as possible. AMs often take the form of a deep network which learns to map from acoustic features to context-dependent phones in a language-specific phone set. It is not clear how exactly this transformation is performed or what is represented in the intermediate layers of such networks. Better characterization of the intermediate representations of AMs may help to guide data-efficient training procedures.

Similar characterizations of networks trained on visual tasks have inspired new transfer learning procedures. For example, Yosinski et al. (2014) characterized the task specificity at each layer of a network trained on ImageNet using transferability as a proxy for task-specificity. This characterization motivated Adaptive Transfer Networks (Long et al., 2015) where parts of a network are trained on the source domain while other parts of the network are finetuned or adapted to the target domain, preserving the limited target data for learning highly task-specific parameters. Similar adaptive transfer learning procedures may also prove to be useful for building AMs for low-resource (data-poor) languages. A necessary first step is to characterize the shape of the transition from task-general to task-specific representations through the layers of deep network-based AMs.

Much of the previous work on characterizing intermediate layers of deep networks has focused on relatively solvable tasks in the visual domain (e.g. hand written digit recognition, visual object recognition) (Zeiler and Fergus, 2014). Few studies have characterized the intermediate representations of networks trained on acoustic tasks (Lee et al., 2009; Golik et al., 2015; Nagamine et al., 2015), which, in practice, are not always trained long enough to converge completely (test error still slowly decreasing at the end of training) due to the long training time required. It is not clear to what extent existing methods developed to probe networks trained on visual tasks will be applicable and useful to study networks that may be underfitting on difficult acoustic tasks.

Here we studied convolutional neural networks (CNNs) used for ASR systems. We characterized the language-specificity of each layer across languages using an approach inspired by Yosinski et al. (2014). Subsets of a network previously trained on one language were 'implanted' into another network which was subsequently trained on a second language. The effect of the implant on performance indicated the language-specificity of the features in the implant. Our main contribution is the characterization of the language-specificity of intermediate layers of CNN-based acoustic models. We also demonstrate the adaptation of an analysis method originally designed to probe visual networks to study networks in an underfitting regime on a phone classification task. Additionally, follow up experiments explore the role of weight freezing in transfer learning.

Table 2. Speech Data English, German and Dutch speech datasets consisted of utterances read by several speakers in a quiet room.

	English	German	Dutch
Hours	82h:44m	67h:42m	63h:46m
Utterances	87906	62294	95350
Phoneset size	54	49	48

2. Experiments

The datasets for this experiment consisted of German, Dutch and American English speech, recorded in similar environments, with corresponding text transcriptions. We chose these languages because we expected a large degree of transferability based on their phonetic similarity. Logarithmic Mel filter bank features were calculated, creating a 45-dimensional feature vector for every 10ms of audio (spectrograms). Each observation was associated with one of 9000 context-dependent phone classes (language-specific). A summary of the speech data can be found in Table 2.

2.1. Baseline models

For each language, a CNN consisting of nine convolutional layers followed by three fully connected layers was trained to recognize context-dependent phones. The architecture was as follows, where triplets specify the filter size and number of feature maps in each convolutional layer and the singletons specify how many units in each fully connected layer: (7, 7, 1024), (3, 3, 256), (3, 3, 256), (3, 3, 128), (3, 3, 128), (3, 3, 128), (3, 3, 64), (3, 3, 64), (3, 3, 64), (600), (190), (9000). This resulted in a total of approximately 7.2 million parameters. All networks were trained using the ADAM optimizer (Kingma and Ba, 2015) as implemented in Tensorflow (Abadi et al., 2016) with a minibatch size of 256, a starting learning rate of $10e^{-5}$ and rectified linear units. Approximately 98% of the data was used for training and the remaining 2% for testing. All model parameters were replicated on four GPUs. Different minibatches were given to each GPU and their gradients were averaged to calculate updates. As a balance between training time and accuracy, each network was trained for a fixed period of 100 epochs (which took approximately two weeks).

2.2. Experimental setup

The subsequent experimental setup was similar to that described in Yosinski et al. (2014). Several ‘network surgeries‘ were performed. The first n layers of a network trained on Language A were implanted into a new network of identical architecture where the layers after layer n were randomly initialized. This ‘chimera’ network was further trained in four different ways. It was either trained on Language A (self-transfer or ‘selfer‘ network) or

Language B (transfer network) and the implanted parameters were either fixed or allowed to be finetuned during training. This process was repeated $\forall\ 1 \leq n \leq 11$ and for all pairs of languages resulting in a total of 198 networks (see Figure 1 in Yosinski et al. (2014) for a graphical depiction of a similar experimental setup). The selfer networks served as a control to capture any changes in performance associated with the surgery but unrelated to the transfer. As in Yosinski et al. (2014), we also measured the effect of leaving the first n layers untrained, i.e. fixed at their random initialization, while training subsequent layers normally. All networks were trained for 100 epochs. Training parameters were identical to those of the baseline models.

3. Results

We found representations throughout the networks to be highly transferable between all three languages. Top-1 test phone classification accuracy for each network is plotted as a function of the layer at which the surgery was performed in Figure 19. Phone classification accuracy is measured with respect to per frame phone-labels established in a forced alignment.

3.1. Transfer networks

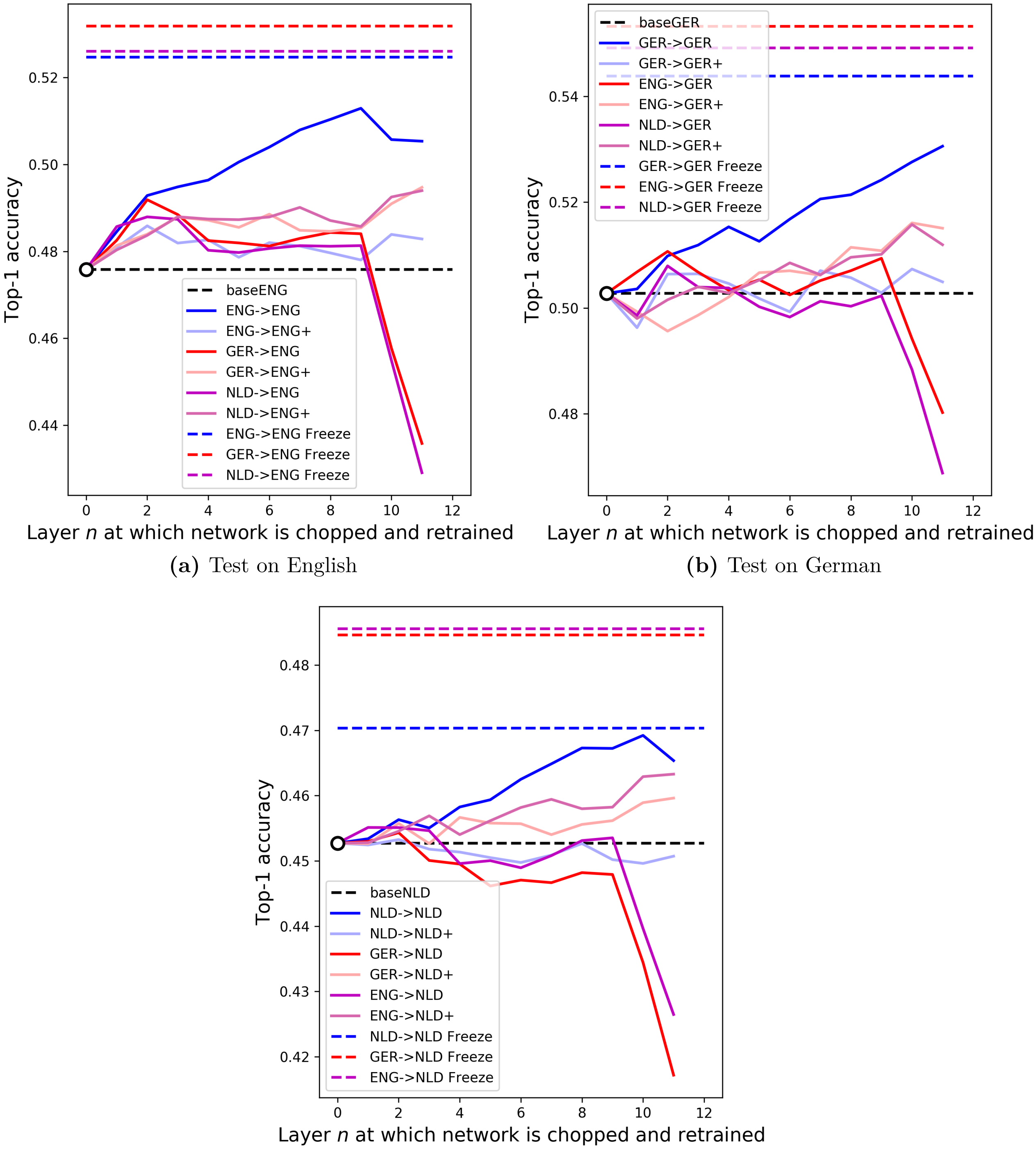
The only models that performed considerably worse than the monolingual baseline models were the transfer networks without finetuning whose surgery occurred at one of the fully connected layers (the penultimate two layers). Transfer networks cut at any of the convolutional layers performed as well as the monolingual baseline model, regardless of whether the implanted layers were finetuned or not. We observed a slight improvement over the monolingual baseline (1.3 percentage points (pp)) for transfer networks with finetuning chopped at one of the fully connected layers.

3.2. Selfer networks

All selfer networks with finetuning performed at the same level as the mono-lingual baseline. Somewhat unexpectedly, the selfer networks without finetuning performed best overall among the chimera networks. Selfer networks chopped at late layers whose implants were not finetuned showed an improvement of 2.7 pp.

3.3. Random features

Previous work has shown that random, untrained weights can often perform remarkably well in certain scenarios (Jarrett et al., 2009; Rahimi and Recht, 2007). Figure 19 shows accuracy as a function of the layer at which training began, meaning that layers below layer n were randomly initialized and never updated. We observed a gradual drop in performance



(a) Test on English

(b) Test on German

(c) Test on Dutch

Figure 19. Test accuracy as a function of depth after 100 epochs. The plus sign indicates that the implanted pretrained layers were finetuned. Th dashed black line indicates the performance of the monolingual baseline models. Up to the ninth layer, layers trained on one language could be copied directly (without finetuning) in a network whose subsequent layers were trained on another language with little to no loss in performance compared to baseline. Selfer networks without finetuning show an improvement compared to baseline. Freeze trained transfer networks yielded the best overall performance. The pattern is similar for all three languages.

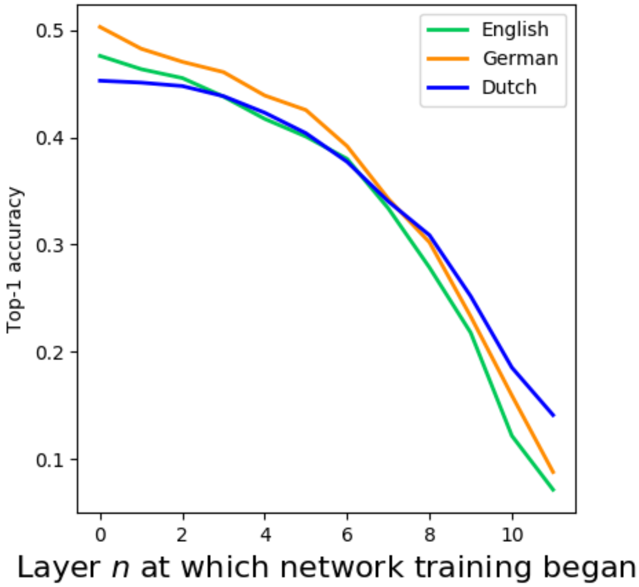


Figure 20. Test accuracy with random weights up to layer n. The leftmost points represent the baseline models. Performance decays gradually as more layers are left untrained, only reaching near-chance performance when nearly all layers are random.

as a function of the depth at which training began. Random weights in early layers did not have a large impact on performance. Using random weights for all but the last layer resulted in near-chance performance. This verifies the non-triviality of the success of our transfer networks without finetuning.

3.4. Freeze Training

The training of our selfer networks without finetuning somewhat resembles the *freeze training* procedure proposed by Raghu et al. (2017). According to this procedure, layers are successively frozen over the course of training, gradually reducing the number of parameters to be updated until, by the end of training, only the last layer is being updated. We hypothesized that weight freezing partly explained the success of our selfer networks without finetuning, so we created freeze trained versions of both our selfer and transfer networks. Starting with a pre-trained network, layers 1–11 (excluding layer 0) were trained for 5 epochs. Then, for the next 5 epochs, only layers 2–11 were trained. From then on, another layer was removed from the trainable parameters every 10 epochs for a total of 100 training epochs. The freeze trained models are represented by the coloured dashed lines in Figure 19. All freeze trained networks outperformed all other networks. The freeze trained transfer networks performed best overall, achieving 4.5 pp above baseline on average.

4. Discussion

Our results suggest that, despite a large degree of transferability of intermediate acoustic features between languages, naive approaches to transfer (e.g. initializing with parameters from another language) are not the most effective nor the most efficient. In particular, early layers need not be finetuned on the target language at all. Subsequent layers benefit greatly from freeze training on the target language. These freeze trained transfer networks outperform all networks trained solely on the target language, which demonstrates the improved generalization that can be achieved when incorporating data from multiple sources.

There are many differences between the current experiments and those presented in Yosinski et al. (2014) (task, domain, architecture). While comparison between these studies is not straightforward, it may still aid interpretation of our results. To what extent do these characterizations apply to all convolutional architectures and tasks, in which case we expect alignment of our results, and to what extent can the deviations that we observe be explained by the particulars of our task or setup?

The performance of the networks with finetuning is largely consistent with Yosinski et al. (2014). However, the performance of networks without finetuning deviates considerably. The transfer networks without finetuning in Yosinski et al. (2014) show a gradual drop in performance, starting at the 4th convolutional layer and eventually dropping nearly 8 pp by the penultimate layer (see Figure 2 from Yosinski et al. (2014)). Our transfer networks without finetuning, on the other hand, show a sharp drop in performance that starts only at the first fully connected layer (layer 9). For the selfer networks without finetuning, we did not observe a performance drop when networks were chopped at middle layers, as was reported in Yosinski et al. (2014). Instead, our selfer networks without finetuning outperformed all other 'chimera' networks, with accuracy increasing nearly monotonically with the depth at which the surgery was performed. Finally, Yosinski et al. (2014)'s experiments with random weights quickly drop to near-chance performance by layer 3, whereas our networks with random weights decline gradually with depth, only approaching near-chance performance when all but the last layer are random.

The success of our selfer networks without finetuning is at least partly explained by the fact that we are in an underfitting regime. Unlike in Yosinski et al. (2014), our baseline model has not converged completely and we would expect continued training to improve performance. However, if that were the only factor at play, we would also expect our selfer networks with finetuning to show improvement over baseline, but they do not. This difference between selfer networks with and without finetuning may be explained by weight freezing and the fact that smaller networks train faster (Saxe et al., 2015). However, we don't see a benefit of weight freezing in the transfer networks without finetuning. Something about freezing all but the last layer(s) facilitates a 2.7 pp improvement over baseline in the selfer but not the transfer networks. This suggests that there is some important language-specific information in the layers that show a difference between the selfer and transfer networks without finetuning (layer 3 and above). Layers 10 and 11 show worse than baseline performance for the transfer network without finetuning, indicating a larger degree of language-specificity in these representations.

Our freeze training results corroborate the interpretation that weight freezing is responsible for the success of our selfer networks without finetuning. Furthermore, our freeze-trained transfer networks performed best overall, demonstrating that freeze training can actually recover the language-specific information lacking in our transfer networks without finetuning, yielding improved generalization. This likely reflects the observation from Raghu et al. (2017) that CNNs converge 'bottom-up' during training, with early layers stabilizing earlier in training. Relatedly, Alain and Bengio (2016) state the proposition that no intermediate layer of a multi-layer neural network will contain more target-related information than the raw input, which requires a 'bottom-up' flow of information; intermediate layers cannot pass on target-related information that they do not receive. Thus, we conclude that freezing the weights of a given layer can only improve performance if that layer already passes on the target-related information in a representation that can be disentangled by subsequent layers. This was not generally the case in our transfer chimera networks because important language-specific information was not being conveyed. The progressive freeze training regime, proposed by Raghu et al. (2017), allowed this important language-specific information to be learned, whereas generic fine-tuning did not. In this way, making fewer parameter updates actually led to significant performance gains.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., and Brain, G. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation.*

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv*, page 1610.01644v3.

Golik, P., Tuske, Z., Schluter, R., and Ney, H. (2015). Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 26–30.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., and Dean, J. (2013). Multilingual Acoustic Models using Distributed Deep Neural Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8619–8623.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 2146–2153.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations (ICLR)*.

Lee, H., Pham, P., Largman, Y., and Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems*.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015). Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*, volume 37.

Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring How Deep Neural Networks Form Phonemic Categories. *Interspeech*, pages 1912–1916.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement. *NeurIPS*.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *NeurIPS*.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2015). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference for Learning Representations (ICLR)*.

Sercu, T., Puhrsch, C., Kingsbury, B., and LeCun, Y. (2016). Very Deep Multilingual Convolutional Neural Networks for LVCSR. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Tuske, Z., Pinto, J., Willett, D., and Schluter, R. (2013). Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7349–7353.

Watanabe, S., Hori, T., and Hershey, J. R. . (2017). LANGUAGE INDEPENDENT END-TO-END ARCHITECTURE FOR JOINT LANGUAGE IDENTIFICATION AND SPEECH RECOGNITION. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 265–271.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *NeurIPS*.

Zeiler, M. and Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision-ECCV 2014*.

Fourth Article.

The effect of task and training on intermediate representations in convolutional neural networks revealed with modified RV similarity analysis

by

Jessica A. F. Thompson[1], Yoshua Bengio[2], and Marc Schönwiesner[3]

([1]) Université de Montréal

([2]) Université de Montréal

([3]) Universität Leipzig

This article was or will be submitted in Proceedings of the Conference on Cognitive Computational Neuroscience (CCN2019).

My contributions and the role of the coauthors

— Conception of all ideas and analyses;

— Run all analyses;

— Preparation of the manuscript.

Prof. Bengio and Prof. Schönwiesner helped with the interpretation of the results. All coauthors helped to revise the manuscript.

RÉSUMÉ. L'alignement centré du kernel (CKA) a récemment été proposé comme métrique de similarité pour comparer les modèles d'activation dans les réseaux profonds. Ici, nous expérimentons avec le coefficient RV modifié (RV2), qui a des propriétés similaires à CKA tout en étant moins sensible à la taille de l'ensemble de données. Nous comparons les représentations de réseaux qui ont reçu des quantités variables d'entraînement sur différentes couches: un réseau entraîné de manière standard (tous les paramètres sont mis à jour à chaque étape), un réseau «freeze trained» (couches figé progressivement pendant l'entraînement), des réseaux aléatoires (seulement quelques couches entraînées) et un réseau complètement non entraîné. Nous avons constaté que RV2 était capable de récupérer les motifs de similarité attendus et de fournir des matrices de similarité interprétables qui suggéraient des hypothèses sur la façon dont les représentations sont affectées par différentes recettes de formation. Nous proposons que la performance supérieure obtenue par l'entraînement figé peut être attribuée aux différences de représentation dans l'avant-dernière couche. Les comparaisons avec des réseaux aléatoires suggèrent que les entrées de données et les cibles servent d'ancrage aux représentations dans les couches les plus basses et les plus hautes.
Mots clés : analyse de similarité, caractéristiques aléatoires, CNNs, entraînement figé, coefficient RV

ABSTRACT. Centered Kernel Alignment (CKA) was recently proposed as a similarity metric for comparing activation patterns in deep networks. Here we experiment with the modified RV-coefficient (RV2), which has similar properties to CKA while being less sensitive to dataset size. We compare the representations of networks that received varying amounts of training on different layers: a standard trained network (all parameters updated at every step), a freeze-trained network (layers gradually frozen during training), random networks (only some layers trained), and a completely untrained network. We found that RV2 was able to recover expected similarity patterns and provide interpretable similarity matrices that suggested hypotheses about how representations are affected by different training recipes. We propose that the superior performance achieved by freeze training can be attributed to representational differences in the penultimate layer. Comparisons to random networks suggest that the inputs and targets serve as anchors on the representations in the lowest and highest layers.
Keywords: similarity analysis, random features, CNNs, freeze training, RV coefficient

1. Introduction

The study of artificial and biological neural networks often requires quantification of the similarity of activation patterns between two networks. Common approaches to this problem are variants of canonical correlation analysis (CCA) (Hotelling, 1936). For example, Singular

Vector CCA and Projection-Weighted CCA have recently been used to uncover insights about training dynamics and generalization in deep networks Raghu et al. (2017); Morcos et al. (2018). Regularized CCA is often used in neuroscience to find relationships between neural and behavioural or clinical variables (Bilenko and Gallant, 2016). However, these variants of CCA can require large amounts of data and so are often impractical for analyzing neural activations where the number of observations may be small and the dimensionality may be large.

When comparing two sets of variables $\mathbf{X}$ and $\mathbf{Y}$, CCA will find the linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ which maximizes their correlation. This means that CCA is invariant to any invertible linear transformation. There are several reasons why one might want a similarity metric with different invariance properties. For example, in a deep network, it is not just the linear information content of a representation that is meaningful but also the specific configuration of that information. For example, the insertion of an invertible linear transformation between two layers of a deep network can alter the network's behaviour (e.g. in batch normalization). Therefore, when comparing representations in deep neural networks, one may wish to use a similarity metric that is not invariant to invertible linear transformation so as to be sensitive to meaningful differences between representations Kornblith et al. (2019); Thompson et al. (2016).

Kornblith et al. (2019) propose the use of Centered Kernel Alignment (CKA) based on the fact that CKA is only invariant to orthogonal transformations and isomorphic scaling (not arbitrary linear invertible transformations) and that it demonstrates intuitive notions of similarity, namely that corresponding layers are most similar to themselves in networks of identical architecture trained from different random initializations. They state that CKA with a linear kernel is equivalent to the RV coefficient. The RV coefficient is a matrix correlation method for comparing paired observations $\mathbf{X}$ and $\mathbf{Y}$ with different numbers of columns (Robert and Escoufier, 1976).

$$RV(\mathbf{X},\mathbf{Y}) = \frac{tr(\mathbf{XX'YY'})}{\sqrt{tr[(\mathbf{XX'})^2]tr[(\mathbf{YY'})^2]}} \tag{1.1}$$

The RV coefficient is still sensitive to dataset size. When the number of observations is too small relative to the number of dimensions, the RV coefficient will tend to 1, even for random, unrelated matrices. The modified RV coefficient (RV2) addresses this problem by ignoring the diagonal elements of $\mathbf{XX'}$ and $\mathbf{YY'}$, which pushes the numerator to zero when $\mathbf{X}$ and $\mathbf{Y}$ are random matrices, even for small sample sizes (Smilde et al., 2009).

$$RV_2(\mathbf{X},\mathbf{Y}) = \frac{Vec(\widetilde{\mathbf{XX'}})'Vec(\widetilde{\mathbf{YY'}})}{\sqrt{Vec(\widetilde{\mathbf{XX'}})'Vec(\widetilde{\mathbf{XX'}}) \times Vec(\widetilde{\mathbf{YY'}})'Vec(\widetilde{\mathbf{YY'}})}} \tag{1.2}$$

Where $\widetilde{\mathbf{XX'}} = \mathbf{XX'} - diag(\mathbf{XX'})$ and similarly for $\widetilde{\mathbf{YY'}}$. Thus RV2 provides a similarity metric with the same invariance properties as CKA while being less sensitive to dataset size, making it a good candidate for comparing neural activities of large artificial and biological neural networks.

Here we explore the use of RV2 to characterize intermediate representations of simple convolutional neural networks. Our main contributions are (a) extending Kornblith et al.'s validation of CKA-flavored similarity metrics by using RV2 to recover expected similarity patterns in simple networks, and (b) showing that RV2 can generate interpretable patterns that can suggest hypotheses about the nature of intermediate representations in deep neural networks.

2. Experiments

Trained networks in the following analyses were previously reported in Thompson et al.. All networks were of identical architecture consisting of nine convolutional layers and three fully connected layers. Networks were trained to recognize context-dependent English or Dutch phones for 100 epochs (except for the untrained network). Networks differed in the training that they received. The standard networks were randomly initialized and all parameters were updated on every mini-batch. The untrained network was randomly initialized and never trained. The procedures for the freeze-trained and random networks are described below. Please refer to the original text for details about the datasets, architecture and training.

Activations to one hour of English speech from 60 speakers (1-minute each) were measured from all networks. We used the hoggorm python package to calculate RV2 for all pairs of layers. To make the experiments feasible, we performed average-pooling on all feature maps and downsampled the resulting activation vectors by a factor of 40, leading to activation vectors of length 23,582 per 'unit'.

2.1. Untrained vs Trained

We replicated Figure F.4 from Kornblith et al. to verify that a slightly different metric, RV2, applied to activations from a different model trained on a different task generates similar patterns of similarity between trained and untrained networks. Figure 21 (bottom row) shows the self-similarity of an untrained network and the similarity between the untrained network and two different trained networks: standard training and transfer freeze-training (described in the next section). We observe approximately the same patterns as are reported in Kornblith et al..

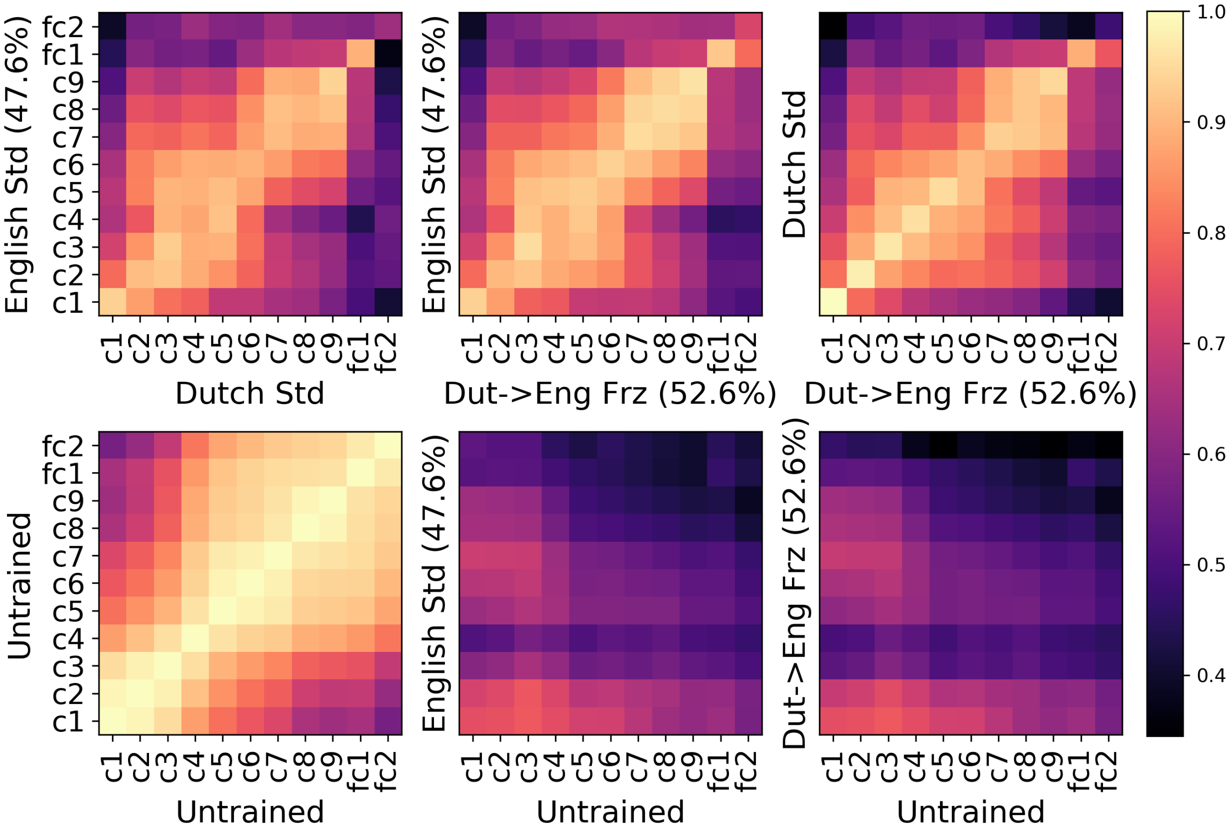


Figure 21. Representational Comparison of Training Recipes. (Top row) Similarity between English and Dutch standard networks and the Dutch-to-English transfer freeze-trained network. The largest differences are in fc2. Lower layers in the transfer freeze-trained network are most similar to their corresponding layer in the Dutch standard model. (Bottom row) Network self-similarity at initialization (left) and the similarity between untrained and trained networks, either standard net (middle) or the transfer freeze-trained net (right). The parenthetical percentages indicates the top-1 accuracy.

2.2. Freeze Training

It has been suggested that convolutional neural networks converge 'bottom-up', with early layers converging to their final form earlier in training (Raghu et al., 2017; Alain and Bengio, 2016). Based on this observation, Raghu et al. proposed *freeze training*. During freeze training, at regular intervals, the parameters of an additional layer are frozen (i.e. removed from the set of trainable variables). Layers are frozen in order by depth such that, by the end of training, only the final layer is being updated. The freeze-trained transfer networks from Thompson et al., which were initialized with parameters from a network previously trained on one language and then freeze-trained on another, outperformed all other freeze-trained networks (no transfer) and other transfer networks (no freeze training).

Here, we compare the activations of the English standard, Dutch standard and Dutch-to-English freeze-trained networks from Thompson et al.. We predict with high confidence that the early layers of the Dutch-to-English freeze-trained network will be more similar to the Dutch than the English standard model since they were initialized with the parameters from the Dutch standard network and received relatively little training afterwards. This provides a good test of whether RV2 is able to recover this expected pattern. Additionally, we were interested to see if the superior performance of the transfer freeze-trained network could be attributed to any representational differences between the compared networks.

For all comparisons between the standard and transfer freeze-trained networks (Figure 21, top row), the highest similarity values were near the diagonal. This pattern provides further validation that, like CKA, RV2 finds the most similar layer in one network to be near the corresponding layer in another network of identical architecture. As predicted, early layers in the Dutch-to-English freeze-trained network were most similar to the corresponding layer in the Dutch standard model and less similar to the English standard model. Near corresponding layers in the English and Dutch standard models were considerably similar to one another, despite being trained on different languages. The largest differences in all comparisons occured in layer fc2. Thus, the superior performance of the transfer freeze-trained network may be primarily attributable to differences in representation at fc2.

2.3. Random Features

Yosinski et al. investigated the effect on performance of leaving progressively more layers untrained in convolutional neural networks trained to recognize objects in images. Performance dropped sharply to zero when the first three layers were left at their random initialization and only subsequent layers were trained. Thompson et al. replicated this experiment with networks trained on speech and found a different pattern (see Figure 22). Performance gradually declined as more layers were left untrained, only reaching near-zero performance when all but the last layer were left untrained (Thompson et al., 2019a).

Random features have a long history of success in kernel machines (Rahimi and Recht, 2007). However, the effect of several consecutive random layers is less well understood. In particular, how do intermediate representations reconfigure as more layers are left untrained?

We presume that the effect of several consecutive random layers is the same as the effect of one random layer: a random projection of the input. None of the work of disentangling the relevant factors of variation has been performed by these random layers and so the remaining trainable layers have the same job to do as was done by the full set of layers in the standard network. According to this hypothesis, the representational transformations originally performed by all 12 layers in the standard network must be somehow compressed into the remaining trained layers of the random networks. The hypothesis that these representational transformations will be evenly distributed across the remaining trainable layers is depicted in Figure 23. The performance of the random network would only be dependent on whether the structure and capacity of the remaining layers is sufficient to learn and represent the necessary transformations. Under this interpretation, a gradual degradation in performance as more layers are left untrained seems more likely and the sharp drop in performance observed in Yosinski et al. is unexplained. To test this hypothesis, we calculated RV2 similarity matrices comparing each random network to the standard English network.

The comparisons between the standard model and the random networks are shown in Figure 24. In the following, 'random net *n*' refers to the network with random layers up to layer *n*; only layers above layer *n* were trained. Layers are named c1, c2, ..., c9, fc1, fc2 to distinguish the convolutional and fully connected layers. In contrast with our hypothesis, late layers remain most similar to their corresponding layer in the standard network, even as more early layers are left untrained. This pattern is especially clear in the similarity matrix for random net 4. The first trained layer of random net 4, layer c5, is diffusely similar to layers c2–c6 in the standard network, while the remaining layers show maximum similarity near the diagonal. When a network is mostly composed of random layers and only the fully connected layers are trained (e.g. random nets 9-10), the trained layers are not similar to any layer in the standard network. While these networks are still able to perform the task to some extent, they clearly do so in a way that does not mimic the standard network.

3. Discussion

Kornblith et al. validated the CKA method by showing that it can identify corresponding layers in two networks trained from different random initializations. Our comparisons of freeze-trained networks, standard networks and untrained networks extend this validation by showing that a related similarity metric, RV2, applied to networks trained on speech, can recover expected and interpretable patterns of similarity.

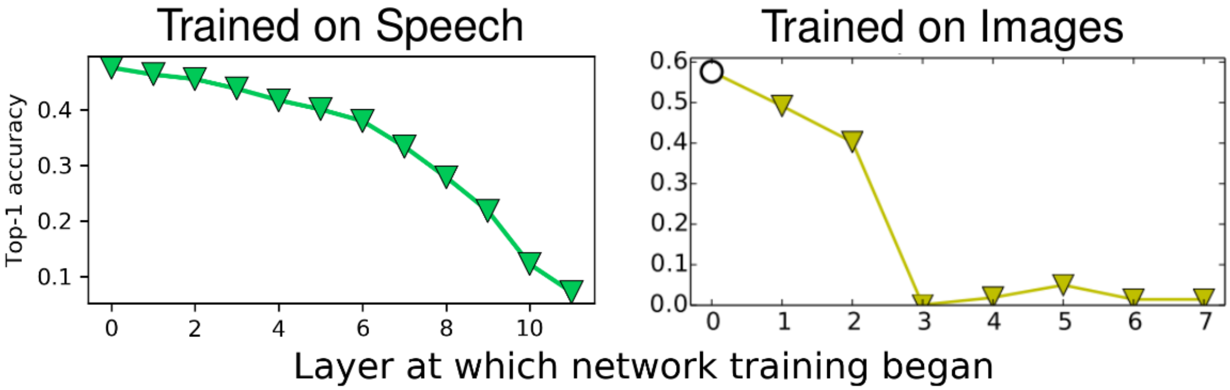


Figure 22. Performance of Random Networks. Test accuracy as a function of layer at which training began as reported in Thompson et al. (left) and Yosinski et al. (right).

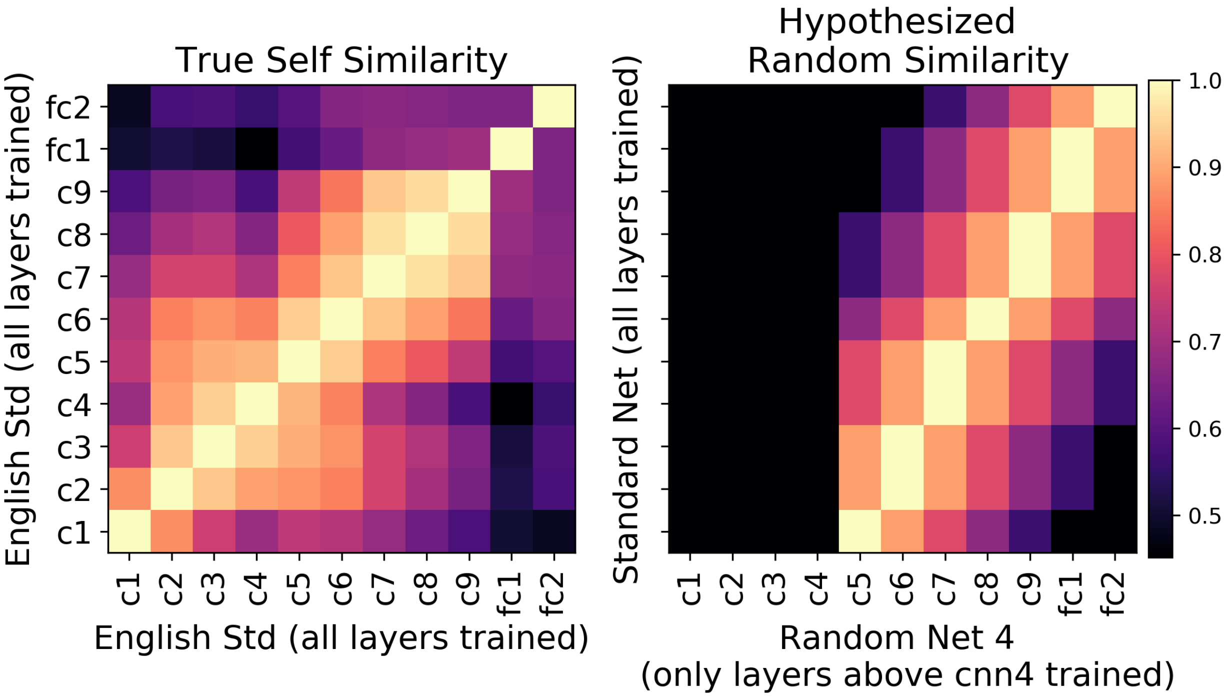


Figure 23. Hypothesized Similarity Matrix of Random Network 4.(Left) Self-similarity of the English standard network. (Right) Idealized diagram of the hypothesis that the representational transformations of the standard network will be evenly distributed across the trained layers of a random net.

Our random networks do not show an even distribution of the needed representational transformations across all trained layers. Instead, early trained layers compensate more for the reduced number of trained layers, such that the representations in late trained layers are less affected. This may reflect architectural constraints on representation. For example, fully connected layers may tend to be more similar to other fully connected layers than to convolutional layers and the fully connected layers may require a particular representation in the preceding convolutional layers. This top-down influence on representations in late layers may also be attributable to the targets serving as an anchor in the same way that the inputs anchor the representations in early layers. While there may be many computational solutions to the classification problem at hand, the form of the inputs and targets themselves are fixed, which may constrain the form of representations near the input and targets.

4. Acknowledgments

Thanks to Nuance and Mitacs for their support and to João Felipe Santos and Guillaume Alain for helpful comments.

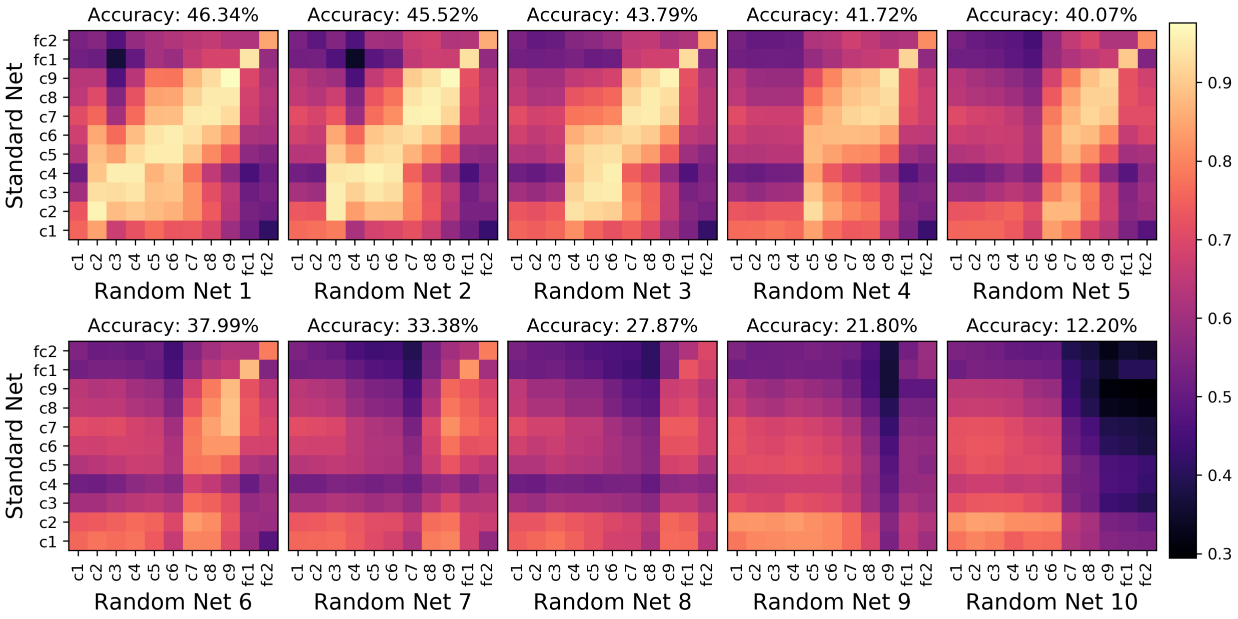


Figure 24. Random Network Similarity. The similarity matrices indicate the RV2 similarity between the baseline model (all layers trained) and networks of identical architecture with only layers above layer n trained, $\forall n \in$ [c1, fc1]. From left to right, progressively more layers are left untrained.

Bibliography

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv*, page 1610.01644v3.

Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Front. of Neuroinformatics.*

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4).

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. *ICLR workshop on Debugging Machine Learning Models.*

Morcos, A. S., Raghu, M., and Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *NeurIPS.*

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement. *NeurIPS.*

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *NeurIPS.*

Robert, P. and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Applied Statistics*, 25(3).

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C. M., and Van Erk, M. J. (2009). Matrix correlations for high-dimensional data: The modified RV-coefficient. *Bioinformatics*, 25(3).

Thompson, J. A. F., Bengio, Y., Formisano, E., and Schönwiesner, M. (2016). How can deep learning advance computational modeling of sensory information processing? *NeurIPS* *workshop on Representation Learning in Artificial and Biological Neural Networks*.

Thompson, J. A. F., Schönwiesner, M., Bengio, Y., and Willett, D. (2019). How transferable are features in convolutional neural network acoustic models across languages? *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *NeurIPS*.

Fifth Article.

No evidence of shared hierarchy between convolutional neural networks and the human auditory pathway

by

Jessica A. F. Thompson[1], Federico De Martino[2], Yoshua Bengio[3], Elia Formisano[4], and Marc Schönwiesner[5]

([1]) Université de Montréal

([2]) Maastricht University

([3]) Université de Montréal

([4]) Maastricht University

([5]) Universität Leipzig

This article was or will be submitted in NeuroImage.

My contributions and the role of the coauthors

— Preparation of all stimuli and the code to execute the experiment;

— Recruited participants and conducted the experimental sessions;

— Design and conduct all data preprocessing and analysis;

— Preparation of the manuscript.

Prof. De Martino provided the scanner protocols and operated the scanner. He and Prof. Formisano directed the data collection. Prof. De Martino, Prof. Formisano, and Prof. Schönwiesner assisted with the experimental design. Prof. Schönwiesner, Prof. Bengio, and Prof. De Martino provided advice on the analysis. Prof. Schönwiesner and Prof. Bengio helped with the interpretation of the results.

RÉSUMÉ. La correspondance entre l'activité des neurones artificiels dans les réseaux de neurones convolutionnels (CNN) entraînés pour reconnaître les objets dans les images et l'activité neuronale collectée à travers le système visuel des primates a été bien documentée. Les couches inférieures de CNN sont plus similaires aux zones corticales visuelles de bas niveau et les couches plus profondes ont tendance à être plus similaires aux zones visuelles de haut niveau, fournissant l'évidence d'une hiérarchie représentationnelle partagée. Ce phénomène n'a pas été rigoureusement étudié dans le domaine auditif. Nous avons comparé les représentations des CNN entraînés pour reconnaître la parole à l'activité IRMf 7-Tesla collectée tout au long de la voie auditive humaine (y compris les régions sous-corticales et corticales) pendant que les participants écoutaient la parole. Nous n'avons trouvé aucune preuve d'une hiérarchie représentationnelle partagée. Au lieu de cela, toutes nos régions auditives d'intérêt étaient les plus similaires à une seule couche des CNN: la première couche entièrement connectée. Cela suggère que des conceptions architecturales ou des objectifs d'entraînement alternatifs peuvent être nécessaires pour obtenir une correspondance par couche avec la voie auditive humaine.

Mots clés : apprentissage profond, cortex auditif, IRMf 7T

ABSTRACT. The correspondence between the activity of artificial neurons in convolutional neural networks (CNNs) trained to recognize objects in images and neural activity collected throughout the primate visual system has been well documented. Shallower layers of CNNs are typically more similar to early visual areas and deeper layers tend to be more similar to later visual areas, providing evidence for a shared representational hierarchy. This phenomenon has not been thoroughly studied in the auditory domain. Here, we compared the representations of CNNs trained to recognize speech (triphone recognition) to 7-Tesla fMRI activity collected throughout the human auditory pathway (including subcortical and cortical regions) while participants listened to speech. We found no evidence for a shared representational hierarchy. Instead, nearly all of our auditory regions of interest were most similar to a single layer of the CNNs: the first fully-connected layer, which has previously been shown to be located at the boundary between the relatively task-general intermediate layers and the highly task-specific final layers. This suggests that alternative architectural designs and/or training objectives may be needed to achieve layer-wise correspondence with the human auditory pathway.

Keywords: deep learning, auditory cortex, 7T fMRI

1. Introduction

The use of deep neural networks (DNNs) as models of biological neural networks has been discussed as an opportunity for synergy between neuroscience and artificial intelligence (Barrett et al., 2019; Marblestone et al., 2016; Richards et al., 2019). Modeling animal cognition and neural function with artificial neural networks, which were themselves inspired by biological neural networks, is not new. The merits of connectionist models of cognition and the implications of neural networks for how we think about brain function were hotly debated in the 1980s and 1990s (Smolensky, 1988; Robinson, 1992). However, recent developments in this area are novel in that state-of-the-art (SOTA) machine learning systems, trained only to maximize their performance on a specific task, without any explicit goal to mimic neural activity, appear to learn representations that are similar to those found in the brains of animals engaged in a similar task (Kriegeskorte, 2015). Critics of this approach claim that DNNs are not interpretable, and hence their comparison to biological neural networks contributes little to the project of explaining neural function and behaviour (Kay, 2018). Proponents have offered multiple alternative perspectives. One view is that this approach shifts the focus from the nature of neural representations themselves to the processes that generate them (what cost functions, architectures and learning algorithms produce brain-like representations?). Additionally, the fact that DNNs are entirely scrutable—their inner workings can be probed much more easily that those of biological systems—defends against the criticism of uninterpretability (see Thompson et al., 2016; Yamins and DiCarlo, 2016; Kubilius, 2017; Kietzmann et al., 2019; Kell and McDermott, 2019; Hasson et al., 2020; Lindsay, 2020).

The paradigm of comparing DNN activity to neural activity has been most thoroughly explored in the study of the primate visual system. The seminal work by DiCarlo and Cox (2007) proposed that visual object recognition is accomplished via successive layers of nonlinear transformations that effectively *untangle* visual inputs, linearizing the boundaries between object categories. Similar language is used to describe how DNNs work (Bengio et al., 2013). Di Carlo's group went on to pioneer the earliest comparisons of representations learned in SOTA convolutional neural networks (convnets) trained to recognize objects in images to multi-unit electrophysiology recordings from visual cortex in the macaque monkey (Cadieu et al., 2014; Yamins et al., 2014). They found the output layer of Alexnet (Krizhevsky and Hinton, 2012) to be highly predictive of IT spiking responses to natural images and intermediate layers to be highly predictive of V4 responses. Similar comparisons have been made between modern convnets and the human visual system as recorded with functional magnetic resonance imaging (fMRI) (Khaligh-Razavi and Kriegeskorte, 2014; Agrawal et al., 2014; Eickenberg et al., 2017; Güçlü and van Gerven, 2016). The most convincing demonstration that modern convnets learn representations that are meaningful to neurons in the primate visual system is work from Bashivan et. al. showing that DNNs can be used to control the activity of macaque V4 neurons. They found that stimuli synthesized to maximally activate specific units in the DNN also drove activity of matched sites in V4 well beyond their maximum firing rate in response to natural images (2019).

Comparisons of DNNs to biological sensory pathways often come with claims of shared hierarchy. Regions of interest (ROIs) along some pathway are mapped to layers of a DNN based on their similarity. Early layers in the network tend to be more similar to early ROIs in the pathway and late layers to late ROIs (Cichy et al., 2016; Güçlü and van Gerven, 2015). These results suggest that DNNs are not just learning representations that are similar to specific regions, but rather that they constitute models of an entire hierarchy of sensory processing. However, not all results have shown evidence of shared hierarchy. Cadena et al. (2019) compared representations at several layers of a convnet trained on ImageNet to neural activation in the mouse visual cortex. While they found their network outperformed classical predictive models, they found no evidence for a shared hierarchy and no benefit over a random network whose weights had never been trained. The authors suggest that networks trained on more ethologically valid tasks may be required to capture the functional organization of the rodent visual cortex.

Relatively few experiments have compared DNNs trained on acoustic tasks to biological auditory systems. The human auditory cortex is generally less well understood than its visual counterpart. In particular, the hierarchical structure of auditory cortex, to the extent that it exists, has not been clearly identified. Classical approaches have modeled early auditory processing as decomposition into spectrotemporal modulation-based basis functions (Chi et al., 2005; Schönwiesner and Zatorre, 2009). Regions that respond selectively to specific auditory stimuli like vocal sounds (Belin et al., 2000) and music (Norman-Haignere et al., 2015) have been identified, but the functional organization of secondary auditory cortex remains underspecified. Kell et al. (2018) trained convnets on speech and music and compared their learned representations to fMRI responses in human auditory cortex. They found that intermediate representations learned in their DNN explained more variance in auditory cortex responses than the spectrotemporal modulation-based baseline model. To assess the existence of a shared hierarchy, they looked only at voxels that showed a reliable response to sound and layers of their network which were predictive of voxel activity across auditory cortex. They found that the most predictive layers of primary "core" auditory cortex were intermediate layers, while the most predictive layers of secondary auditory cortex were later layers. From this, they conclude that the hierarchical distinction between primary and secondary auditory cortex is mirrored in their convnet. Güçlü et al. (2016) also reported evidence for a shared hierarchy in auditory cortex, but they only looked at the superior temporal gyrus (STG). They used representational similarity analysis (RSA) to compare representations learned in a DNN trained to predict tags from excerpts of musical audio.[8] They found a gradient of complexity across STG where anterior voxel clusters were more similar to early layers while posterior voxel clusters were more similar to late layers. While both of the above studies report evidence for a shared hierarchy between human auditory cortex and DNNs trained on sound, they report largely orthogonal spatial patterns of similarity gradients.

Several different analysis tools are used in the comparison of representations. The ultimate goal of these analyses is to quantify the similarity of two representations. Similarity is an ambiguous term that must be defined by the experimenter. In many of the aforementioned studies, some sort of encoding analysis is performed where firing rates or voxel activity are/is predicted by a regularized linear model of the neural network representations. The definition of similarity in this case is the predictive accuracy on a held out set (or the average over cross validation folds), i.e. a representation is similar to another to the extent that one can be linearly predicted from the other. There are other notions of representational similarity that have been explored to study deep neural networks. Singular value canonical correlation analysis (SVCCA) and projection-weighted canonical correlation analysis (pwCCA) have both been used to characterize how network representations change over training, to compare representations in different architectures, and to understand the difference between networks that memorize and networks that generalize (Raghu et al., 2017; Morcos et al., 2018). Kornblith et al. (2019) recently proposed that a meaningful notion of similarity should find corresponding layers in two networks of identical architecture and training, differing only in their random initialization, to be most similar to each other. Of the tested metrics, which included SVCCA, pwCCA and linear regression, Centered Kernel Alignment (CKA) was the only method which found that corresponding layers were most similar to each other, achieving an accuracy of 99.3% on the layer identification task. The next best metric, linear regression, achieved only 45.4%. This result may be related to the fact that CKA is only invariant to orthogonal transformations and isomorphic scaling, unlike canonical correlation analysis (CCA), which is invariant to any linear invertible transformation and linear regression, which is invariant to any invertible transformation of the predicted variables[9]. Representational similarity analysis (RSA) (Kriegeskorte et al., 2008), commonly employed in fMRI analysis, is similar to CKA in that it takes all the pairwise distances between all examples in two domains. However, CKA provides a more general framework with interpretable units, proven convergence rates, and the option to explore different kernels.

[8] Tags are descriptive text annotations like genre or instrumentation labels.

[9] To further justify why linear regression may not be an ideal tool to assess similarity, consider the success of random features in machine learning. Let $\hat{X}$ be a random projection of X into a higher dimensional space. $\hat{X}$ is likely to be more predictive than X of some non-linearly related target y by virtue of its higher dimensionality. However, it seems counter intuitive to say that $\hat{X}$ is more *similar* to y.

Here, we use CKA to quantify the similarity between representations learned in convnets trained on speech and activity throughout the human auditory pathway during speech listening as measured with 7-Tesla (7T) fMRI. The increased spatial resolution of 7T fMRI allows us simultaneously measure activity from auditory cortex as well as subcortical auditory regions, which are normally left out of auditory fMRI analyses due to their small size. Since significant auditory processing occurs in brainstem and midbrain regions, this provides us with several distinct regions with a relatively known connectivity structure with which to compare to the convnet representations. If there exists a shared hierarchy between the convnets and the human auditory pathway, the pattern of similarity should at least distinguish between cortical and subcortical regions. We visualize the results of the similarity analysis as similarity matrices with network layers as the rows and auditory ROIs as the columns. Evidence of a shared hierarchy would manifest as a diagonal pattern in one such similarity matrix, where early layers are more similar to early regions and later layers more similar to later regions. While we do observe similarity values that exceed the similarity of a random network, we find no such diagonal pattern. Instead we find that, on average, nearly all ROIs are most similar to the first fully-connected layer.

2. Methods and Materials

2.1. Participants

Six healthy subjects (aged 28–31, three women) with normal hearing and no known neurological disorders were recruited to participate. All subjects provided written informed consent prior to the first session. All subjects also consented to their data being made publicly available. The native languages of the subjects were English (one subject), German (three subjects) and Dutch (two subjects).

2.2. Stimuli

To facilitate comparison with the convnets, the experimental stimuli should be similar to the sounds that the networks were trained on. Thus, we selected utterances from the same corpus that the networks were trained on. The comparison is complicated by the fact that, although the networks were only trained on phonetic labels, the human participants also understand the meaning and high-level structure of the speech. Therefore, as described below, we transformed the natural speech to remove higher-level structure while preserving phonemes.

The audio datasets from which the stimuli were generated were the same datasets that were used in Thompson et al. (2019a) and Thompson et al. (2019b) which were provided by Nuance Communications. Each of the three datasets, one for English, Dutch and German, contained 64–83 hours of spoken text read by several native speakers in a quiet room. The datasets also included phonetic transcriptions established in a forced alignment with the text transcriptions.

The quilting procedure, adapted from (Overath et al., 2015) and depicted in Figure 25, chops a sound file into small segments and reorders the segments according to a heuristic designed to hide the *seams* of the quilt (the segment boundaries) [10]. A random segment is chosen as the first segment in the quilt. Subsequent segments are chosen to best match the segment-to-segment boundaries in the chochleogram of the original audio. In this way, temporal patterns longer than the segment length are destroyed while minimizing the artefacts introduced by reordering the segments.

Instead of using fixed segment lengths, as in Overath et al. (2015), we used the provided phonetic boundaries to divide the speech into variable length segments containing single phonemes. The resulting quilts are out-of-order sequences of phonemes, preserving phonetic information while destroying the words and semantic content of the speech. Since the quilting procedure will be more effective the larger the input corpus relative to the desired quilt length, we selected the 60 speakers (30 women and 30 men) with the longest set of utterances in each language. Given all the utterances from a single speaker as input, the quilting procedure generated a one-minute quilt. The experimental stimuli consisted of 180 one-minute speech quilts (60 per English, Dutch and German). The final stimuli were filtered to account for the frequency response profile of the foam-tip earphones over which the stimuli were presented in the scanner.

2.3. Experimental Protocol

The experimental procedures were approved by the ethics committee of the Faculty for Psychology and Neuroscience at Maastricht University. MR images were collected over two sessions, each consisting of 10 functional runs. Nine quilts were presented in each run, grouped into blocks of three quilts from the same language. Within a block, the quilts were presented one after another with no interruption. Blocks were separated by short periods of rest which were sometimes followed by a question about the preceding block. To ensure that participants were awake and paying attention to the stimuli, we asked them to identify the language of the speech presented in the last block. Participants used a button box to indicate their response. To save time, we didn't ask this question after every block. However, the design was such that the participants could not easily predict whether or not they would be questioned and so had to pay attention during every block. Each run contained one block for each language. Each quilt was presented only once. The stimuli order was randomized for each subject separately.

[10] Original sound quilting code can be found here: http://mcdermottlab.mit.edu/downloads.html

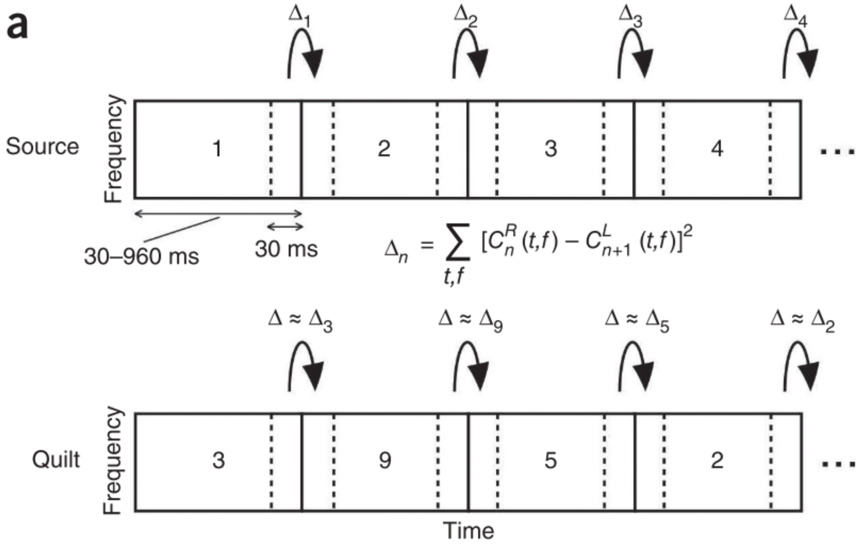

a
Δ_1
Δ_2
Δ_3
Δ_4
Source
Frequency
1
2
3
4
...
30–960 ms
30 ms
$\Delta_n = \sum_{t,f} [C_n^R(t,f) - C_{n+1}^L(t,f)]^2$
$\Delta \approx \Delta_3$
$\Delta \approx \Delta_9$
$\Delta \approx \Delta_5$
$\Delta \approx \Delta_2$
Quilt
Frequency
3
9
5
2
...
Time

Figure 25. Sound Quilting Algorithm Segment 3 is randomly chosen as the first segment in the quilt. Δ_3 is a vector summarizing the change in spectral power that occurs at the transition from segment 3 to segment 4. The next segment in the quilt is selected as the segment that leads to the segment-to-segment change nearest to Δ_3 (excluding segment 4). This process repeats, selecting segments without replacement, until the desired quilt length has been achieved or all segments have been selected. Figure copied from Overath et al. (2015).

2.4. fMRI Acquisition Parameters

Images were acquired at Maastricht University, Maastricht, Netherlands on a 7T Siemens MAGNETOM scanner (Siemens Medical Solutions, Erlangen, Germany), with 70 mT/m gradients and a head RF coil (Nova Medical, Wilmington, MA, USA; single transmit, 32 receive channels). Foam pads were used to minimize head motion.

2.4.1. *Anatomical*

At the start of each session, a T1-weighted (T1w) image and a proton density weighted (PDw) image were acquired using a 3D MPRAGE sequence [voxel size=1.0mm isotropic; repetition time (TR)=2370 ms; echo time (TE)=2.31 ms; flip angle=5°; generalized autocalibrating partially parallel acquisitions (GRAPPA)=3 (Griswold et al., 2002); field of view (FOV)=256 mm; 256 slices, phase encoding direction: anterior to posterior, inversion time (TI) for T1w only=1500 ms].

2.4.2. *Functional*

Functional MRI data were acquired with a 2-D Multi-Band Echo Planar Imaging (2D-MBEPI) sequence (Steen Moeller et al., 2010; Setsompop et al., 2012). In order to include the entire brainstem and thalamus as well as primary and secondary auditory cortex, slices

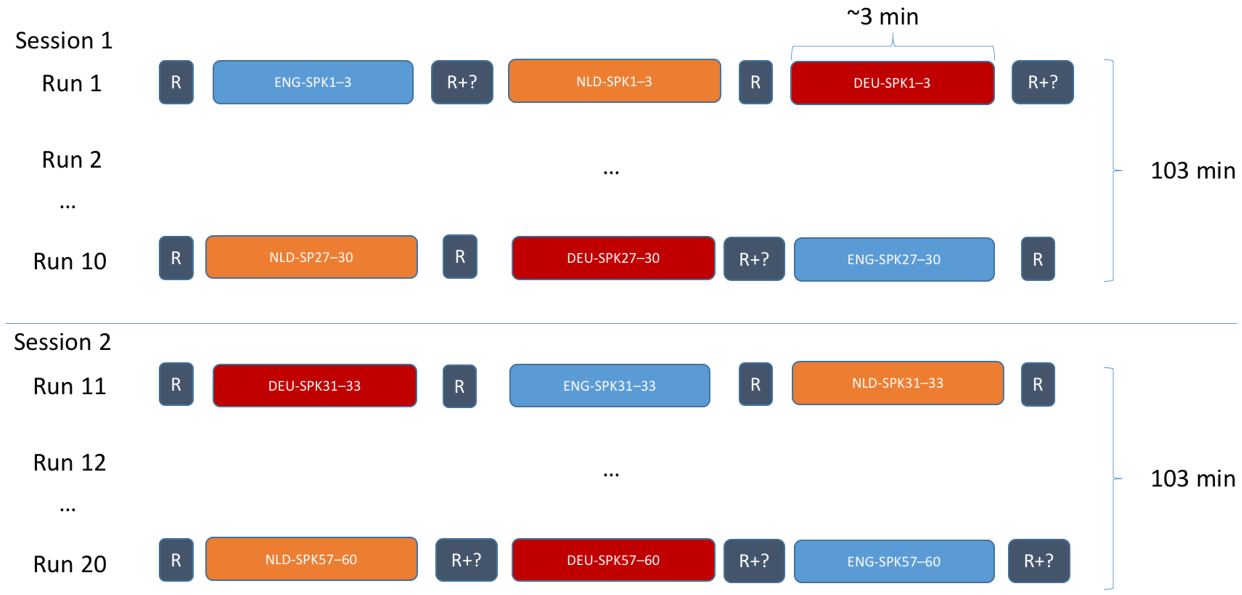


Figure 26. Experimental Protocol. The experiment consisted of two sessions over which the stimuli were presented only once. Each session consisted of 10 runs and each run consisted of three blocks. The blocks consisted of three speech quilts generated from three separate speakers of the same language. The order of language blocks and speech quilts within blocks was randomly assigned in a unique way for each subject. In total, the experimental stimuli amount to one hour of speech quilts in each of the three languages.

were arranged in a coronal oblique orientation (TR=1700 ms; TE=20 ms; flip angle=70°; GRAPPA=3; Multi-Band factor=2; FOV=206 mm; 1.7 mm isotropic voxels; phase encode direction inferior to superior).

2.5. Preprocessing

The MRI preprocessing was performed using *fMRIPrep* 1.4.1 (Esteban et al. (2018b); Esteban et al. (2018a); RRID:SCR_016216), which is based on *Nipype* 1.2.0 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502). The following description was prepared by *fMRIPrep*.

2.5.1. *Anatomical data preprocessing*

A total of 2 T1-weighted (T1w) images were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757). The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted

T1w using `fast` (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). A T1w-reference map was computed after registration of 2 T1w images (after INU-correction) using `mri_robust_template` (FreeSurfer 6.0.1, Reuter et al., 2010). Brain surfaces were reconstructed using `recon-all` (FreeSurfer 6.0.1, RRID:SCR_001847, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: *ICBM 152 Nonlinear Asymmetrical template version 2009c* [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym].

2.5.2. *Functional data preprocessing*

For each of the 20 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). BOLD runs were slice-time corrected using `3dTshift` from AFNI 20160207 (Cox and Hyde, 1997, RRID:SCR_005927). The BOLD time-series, were resampled to surfaces on the following spaces: *fsaverage5*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in ['MNI152NLin2009cAsym'] space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

2.6. Regions of Interest

We extracted BOLD responses at specific regions of interest (ROIs) along the auditory pathway: cochlear nucleus (CN), superior olivary complex (SOC), inferior colliculus (IC), medial geniculate nucleus (MGN), Heschl's gyrus (HG), planum temporale (PT), planum polare (PP), superior temporal gyrus anterior portion (STGa), superior temporal gyrus posterior portion (STGp). We used the subcortical region definitions from the atlas recently published by Sitek et al. (2019) [11] Cortical regions were defined using the Harvard-Oxford parcellation included in FSL 5.0 and accessed through nilearn 0.5.2 (Abraham et al., 2014). ROI definitions included both left and right hemispheres. Nilearn's `NiftiMasker` was used to extract activity from each of the ROIs. The masks for the cortical regions took the

[11] Due to the small size of CN and SOC and the difficulty of inter-subject alignment of the brainstem, we cannot be completely certain that the activity we extract truly corresponds to activity in these small brainstem regions. However, the participants in the present study were also participants in the auditory fMRI sessions reported in (Sitek et al., 2019), providing some assurance that these region definitions are reasonable.

Table 3. Regions of Interest Subcortical definitions came from (Sitek et al., 2019) while cortical region definitions came from the Harvard-Oxford atlas as access via nilearn. The number of voxels refer to the functional data collected in this experiment.

Anatomical Region	Number of Voxels	Source
Cochlear Nucleus (CN)	14	Sitek & Gulban
Superior Olivary Complex (SOC)	41	Sitek & Gulban
Inferior Colliculus (IC)	271	Sitek & Gulban
Medial Geniculate Nucleus	240	Sitek & Gulban
Heschl's Gyrus (HG)	1040	Harvard-Oxford
Planum Polare (PP)	1684	Harvard-Oxford
Planum Temporale (PT)	1213	Harvard-Oxford
Superior Temporal Gyrus anterior portion (STGa)	794	Harvard-Oxford
Superior Temporal Gyrus posterior portion (STGp)	3180	Harvard-Oxford

intersection with the subject's brain mask, as prepared by *fMRIPrep*. To improve the signal-to-noise-ratio (SNR), the `NiftiMasker` detrended, standardized and removed confounding variables from the masked fMRI signals. The confounds were those calculated by *fMRIPrep* described in the previous section, including global signal, CSF, white matter, motion correction parameters and their derivatives as well as physiological component regressors. This resulted in preprocessed and denoised bold data for each ROI, subject and run.

2.7. Convolutional Neural Network Activations

The trained neural networks analyzed here are a subset of those analyzed in Thompson et al. (2019a). All networks were trained to perform context-dependent phone (triphone) classification. Here we look only at the 9 freeze trained networks, which outperformed all other models. These 9 networks consist of 3 monolingual networks for each of the three languages (English, Dutch and German) and 6 transfer networks which were first trained on one language and then freeze trained on another. In all cases, the networks were first trained normally for 100 epochs and then freeze trained for an additional 100 epochs. Freeze training (Raghu et al., 2017) refers to the procedure by which layers are gradually removed from the set of trainable variables over the course of training and in order of depth. All networks are of identical architecture and consist of 9 convolutional layers followed by 3 fully connected layers. The layers were as follows, where triplets specify the filter size and number of feature maps in each convolutional layer and the singletons specify how many units in each fully connected layer: (7, 7, 1024), (3, 3, 256), (3, 3, 256), (3, 3, 128), (3, 3, 128), (3, 3, 128), (3, 3, 64), (3, 3, 64), (3, 3, 64), (600), (190), (9000). The input data were 45-dimensional filterbank features calculated at a rate of one frame every 10 ms.

For every network, the activation when the original (unquilted) speech stimuli was presented as input was recorded. For convolutional layers, the average activation within each feature map was recorded. For fully connected layers, the activation at each unit was recorded. Only the activation to every second frame of the audio features was saved. Subsequently, the network activations were segmented according to the same phonetic boundaries and were quilted according to the same segment order used when generating the experimental stimuli. This produced a sequence of network activations corresponding to each of the 180 speech quilts presented in the scanner.

2.8. CKA Similarity Analysis

CKA takes two matrices as input: in this case one for the BOLD responses and one for the neural network responses to the same stimuli. These matrices must have the same number of rows, corresponding to time points or observations, but can differ in the number of columns, corresponding to voxels or units. Since the temporal rate of fMRI is much slower than that of our acoustic features, temporal rescaling and alignment is required. The preprocessed BOLD timeseries from each ROI and each run were put into pandas `DataFrames` (McKinney, 2010, 2011) with `TimeDelta` indices, which enables indexing into the time series at specific time values. These `DataFrames` were then upsampled to match the rate of the network activations (one frame every 20 ms) using the pandas resample function with the padding strategy—values were simply repeated to achieve the desired frame rate. This strategy allows us to preserve the temporal precision of the network activations without need for summary or binning.

To store the network activations, new `DataFrames` were initialized for each run using the corresponding index from the upsampled BOLD `DataFrame`. The quilted activations were then inserted at the timepoints at which they were presented. Timepoints when no stimulus was presented were set to zero. Since the timing of the experimental runs and the stimuli presentation order was different for each subject, this resulted in one `DataFrame` per subject per run for each layer of each convnet. The Glover model of the hemodynamic response function (HRF) (kernel length=32 seconds), as implemented in *nistats* 0.0.1b0, was convolved with the network activations.

Finally, we extracted and concatenated only the time segments corresponding to the blocks of continuous auditory stimulation from both the fMRI and network activity. The first six seconds of each block was excluded from the analysis to allow for the HRF to ramp up. Thus, the to-be-analyzed fMRI activity does not include the on/off response at the onset of the blocks. Responses to each block were trimmed to exactly 8599 frames, which, when concatenated, resulted in matrices with 515940 rows for both the fMRI and neural network activity. CKA similarity was then calculated for all ROI-layer pairs using code from the Google colab that was released with Kornblith et al. (2019). We used CKA with a linear kernel, in feature space and with a unbiased estimator of the dot product similarity. The CKA similarity value in this case is equivalent to the modified RV-coefficient (RV2) (Robert and Escoufier, 1976; Smilde et al., 2009). The resulting similarity value is still biased, but less so than standard CKA. Calculating CKA in feature space is equivalent but faster than calculating in the space of examples when the number of examples is greater than the number of features.

2.8.1. *Neural similarity score*

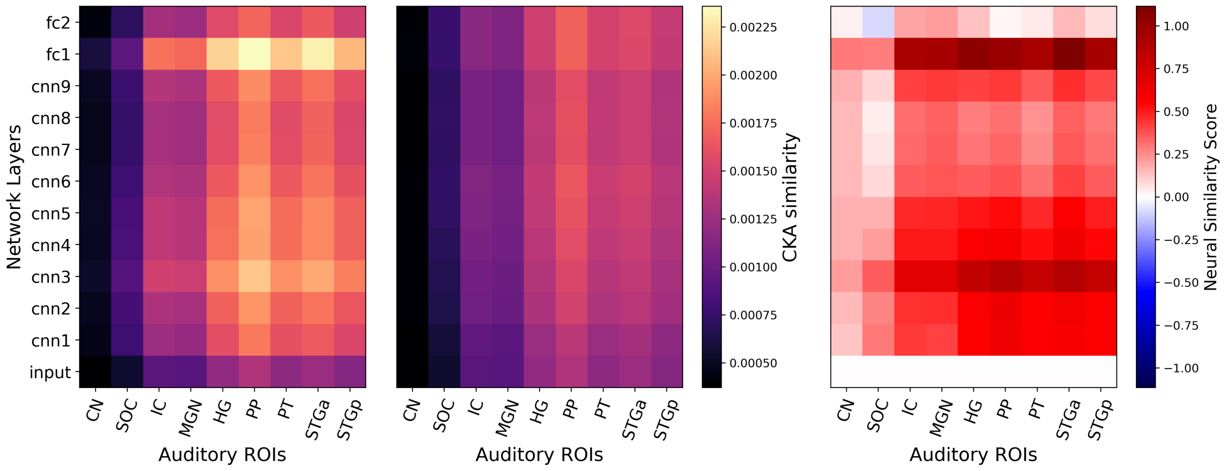
We define the *neural similarity score* as the difference of standardized CKA scores between the trained network of interest and the untrained network. This untrained network has the same architecture as the trained models, but its parameters have been randomly initialized and never updated. If the optimization procedure has increased the correspondence to the brain, the CKA scores for a trained network should be greater than that of the untrained network. Within each subject, the CKA scores are standardized using the mean μ_s and standard deviation σ_s^2 calculated over all models and ROI-layer pairs. The CKA scores of the untrained network are standardized using the same mean and standard deviation. The neural similarity score ϕ_m^s is a difference of z-scores which reflects the similarity achieved by model m in subject s relative to the untrained model.

$$\phi_m^s = \frac{cka_m - \mu_s}{\sigma_s^2} - \frac{cka_{untrained} - \mu_s}{\sigma_s^2} \tag{2.1}$$

Thus a neural similarity score of 1 indicates that the similarity achieved by the trained model is 1 standard deviation greater than that achieved by the untrained network.

3. Results

We calculated the CKA similarity for each of the nine trained networks, for each subject, and for each ROI-layer pair. The results of these analyses can be summarized in similarity matrices whose rows correspond to layers of a network and whose columns correspond to the auditory ROIs. Figure 27 shows the grand average similarity matrices, averaged over subjects and models. Training the networks increased their correspondence with the auditory ROIs, as evidenced by the fact the the neural similarity score matrix is largely red, indicating positive values (Figure 27c). However, we find little evidence of a shared hierarchy, which would manifest itself as a diagonal pattern of high neural similarity scores. This hypothesized diagonal pattern also does not occur in the raw CKA similarity scores, neither for the trained or untrained networks (Figure 27a–b). For all ROIs, the first fully connected layer (fc1) achieves the highest CKA similarity. This pattern does not occur in the similarity matrix for the untrained network, suggesting that it was introduced by optimization procedure and not by the architecture.



(a) Trained **(b)** Untrained **(c)** Neural Similarity Score

Figure 27. Grand Average Similarity. No shared hierarchy is observed. **(Left)** Raw CKA similarity averaged over subjects and models. **(Middle)** Raw CKA similarity for the untrained network, averaged over subjects. **(Right)** Neural similarity score averaged over subjects and models. The similarity matrix is largely red indicating that training increased correspondence but there is no diagonal pattern to indicate a shared hierarchy. The first fully connected layer (fc1) achieves the highest mean neural similarity score for nearly all ROIs.

We calculated the average neural similarity score matrix for each model to investigate how the different training data would affect the correspondence. Figure 35 displays nine similarity matrices arranged in a grid. The monolingual models, which were only ever trained on one language, are along the diagonal of the grid. The off-diagonal matrices correspond to the transfer networks which were first trained on one language and subsequently freeze trained on another. The patterns observed in the grand average, are largely replicated in the model-specific similarity matrices. Layer fc1 generally achieves high neural similarity scores and none of the models show any clear evidence for a shared hierarchy. In some models, most strikingly in the English and Dutch monolingual models, the neural similarity scores for layer fc2 are negative, indicating that training actually decreased their correspondence to the brain.

We hypothesized that the differences between models observed in Figure 35 may be related to the models' accuracy on the phone classification task on which they were trained. In Figure 29 we plot the peak neural similarity score as a function of triphone classification accuracy. For Dutch and English models, all slopes are positive, indicating a positive relationship between model accuracy on the speech recognition task and the similarity to the human brain. However, this pattern is largely driven by the low neural similarity scores achieved by layer fc2 in the English and Dutch monolingual networks. The pattern does not persist if layer fc2 is removed from the analysis. For the German models, we found a strong

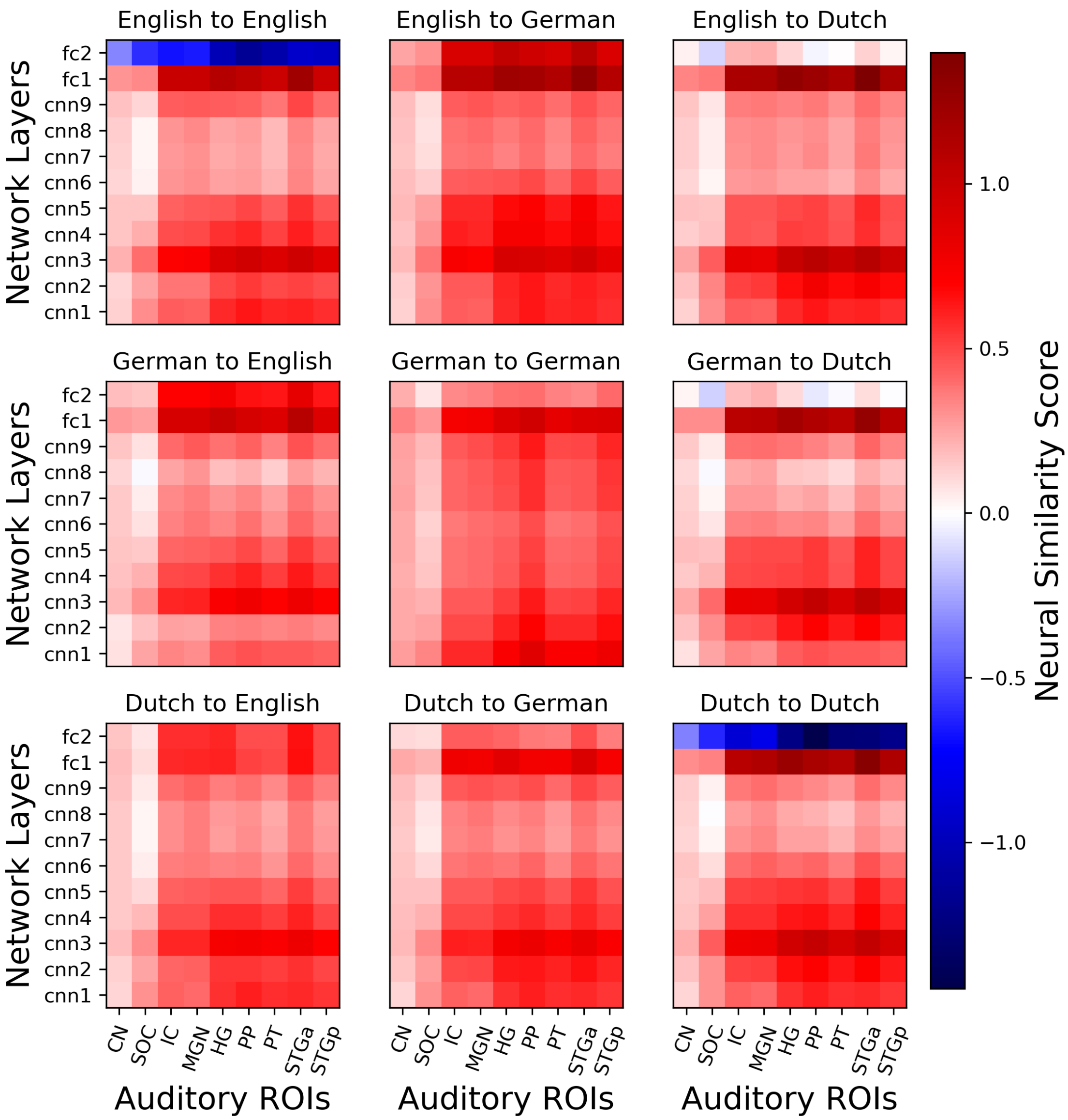


Figure 28. Average Neural Similarity Score. Each similarity matrix shows the effect of training on CKA similarity averaged over the six subjects. The subtitles of the form "Language 1 to Language 2" indicate that the model was first trained on Language 1 and then freeze trained on Language 2. Training generally increased the correspondence between brain and networks. Layer fc1 shows the highest neural similarity score and there is little evidence for shared hierarchy (no diagonal pattern). In some models, training actually reduced the similarity between layer fc2 and the ROIs (shown in blue).

positive relationship only for German native speakers. The regression statistics are reported in Table 4.

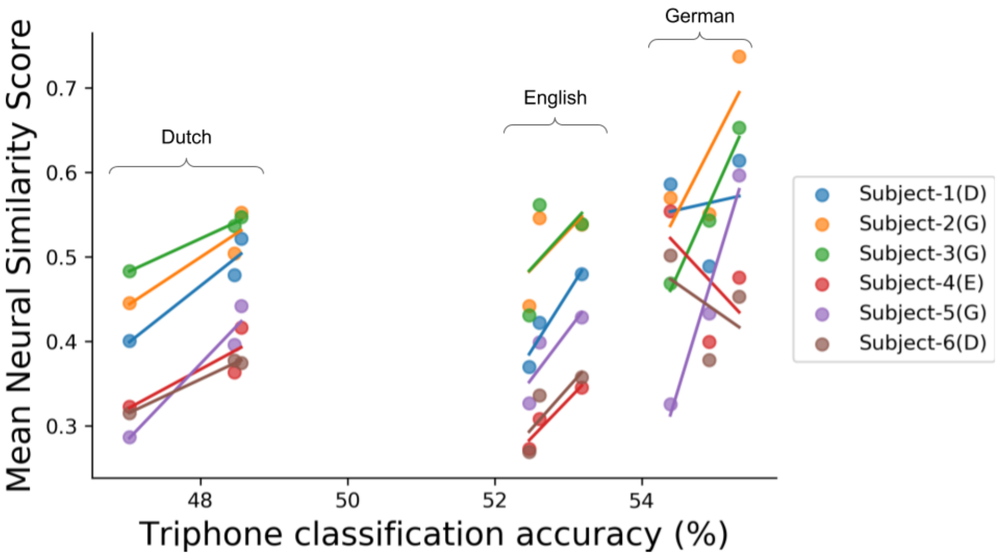


Figure 29. Mean Neural Similarity Score vs Model Accuracy. There are nine points per subject for the nine different network models. Lines show the linear regression fit to the three models (one monolingual and two transfer) for each language and subject. Triphone classification accuracy indicates the top-1 test accuracy achieved by each model. For Dutch and English models, there is a positive relationship between model accuracy and the correspondence to the human brain. Paretheticals in the legend indicate the native language of each subject.

Table 4. Network Accuracy vs. Neural Similarity Score Summary of the relationship between network accuracy on the triphone recognition task and peak neural similarity score, as depicted in Figure 29, averaged over subjects. All mean slopes are positive, but the pattern is less consistent for networks tested on German.

Language model was tested on	Mean slope	Standard deviation of slope
English	0.10	0.02
German	0.09	0.14
Dutch	0.06	0.02

4. Discussion

Our primary aim was to characterize the degree to which convnets trained on speech tasks learn hierarchical representations that parallel the hierarchical structure of the human auditory pathway. To the best of our knowledge, this is the first study to compare DNN representations to activity throughout the human subcortical and cortical auditory pathway as measured with 7T fMRI. Unlike the previous results of Kell et al. (2018) and Güçlü et al. (2016), we find no evidence of a shared hierarchy. Instead, the first fully-connected layer, fc1, achieved the highest similarity score across nearly all ROIs. This suggests that the sequence of representational transformations learned in our convnets does not mirror that performed by the human brain. However, the two solutions found in the convnets and the human brain appear to intersect most at layer fc1.

Kell et al. (2018) similarly found that the median variance explained across auditory cortex was maximal at layers near but not at the end of the network. This common observation may be related to the notion of dimensionality expansion and compression in DNNs. Recanatesi et al. (2019) and Ansuini et al. (2019) describe a two-stage process by which trained DNNs perform a task. The first stage, which we might call 'feature extraction', is characterized by increasing intrinsic dimensionality (dimensionality expansion) in the early layers of the network. The second, dimensionality compression, is characterized by decreasing intrinsic dimensionality in the last layers of the network, as the network projects the data to a low-dimensional manifold from which the target can be linearly decoded. Our layer fc1 may be the last 'expansion' layer before the 'compression' of the final layers. From Thompson et al. (2019a), we know that layer fc1 is at the barrier between the intermediate layers which are largely transferable between languages, and the final layers which are highly task specific. In Thompson et al. (2019b), layer fc1 was the deepest layer to show a high degree a similarity in networks trained on different languages (Figure 21, top row). The last layers of networks trained on narrowly defined tasks such as triphone recognition may simply learn representations that are more task-specific than any representations employed by the human brain, whose ultimate goal during speech listening is typically natural language understanding, not phoneme recognition.

There are a number of methodological inconsistencies in the previous work on comparing activations in artificial and biological neural networks. Most analyses have employed some type of regularized linear regression or RSA, rather than the CKA used here, to quantify similarity. Uden (2019) compared CKA score and ridge regression accuracy as similarity metrics for the comparison of fMRI activity in human visual cortex and convnet activations of networks trained to recognize objects in images. Like us, Uden also found no diagonal pattern in her similarity matrices, neither with CKA nor with ridge regression. Based on the analysis presented in Kornblith et al. (2019), we think that the use of CKA is well justified and may not yield appreciably different results than linear regression in many cases.

It is common in fMRI encoding analyses to first select a subset of to-be-analyzed voxels based on their response profiles, for example, based on their selective response to sound or specific sound categories (as in Kell et al. (2018)). Such selection procedures are typically employed to increase the signal-to-noise-ratio, and assume that the voxels of interest—the voxels that contain cells involved in the task-of-interest—are those that show a selective response to the stimulus or condition. However, research is accumulating in both neuroscience and machine learning that casts doubt on that assumption (c.f. Leavitt et al. (2017); Morcos et al. (2018)). In all presented results, we selected voxels based only on anatomical ROIs and included all voxels within a given ROI. We explored the effect of further voxel selection within cortical ROIs based on selective response to sound, but found that the pattern of similarity was unaffected by this additional selection.

Unlike much of the previous work in this area, our neural similarity score quantifies the similarity between layers and ROIs relative to the similarity value achieved by a random, untrained network. This is an important contrast to understand the effect of training above and beyond architecture design. In Kell et al. (2018), the median variance explained in auditory cortex by the layers of a random network follows a similar pattern to that of the trained models: variance explained increases until the last pooling layer, after which it declines. At the last pooling layer, the variance explained by the random network is nearly equal to that explained by the spectrotemporal baseline model. Thus random networks do not capture the similarity expected 'by chance', but rather include several important aspects of the model and experiment, minus the effect of network training.

The networks analyzed here were trained to recognize triphones and we specifically designed our fMRI experiment to focus on the acoustics of speech, omitting any syntactic or semantic information. This differs from previous work that used various natural sounds as stimuli and networks that were trained to classify words or music. In particular, Kell et al. (2018) trained the speech branch of their dual head network to classify words embedded in noisy backgrounds. Perhaps triphone classification is too low-level to reveal a shared hierarchy. Similarly, differences in experimental design and analysis may impact the results. Kell et al. (2018) presented 165 unique two-second sounds (5.5 minutes) from a variety of natural sound categories, repeated several times in a block design to measure reliable responses. We, on the other hand, presented 3 hours of unique speech in a continuous design with no repetitions, to maximize the number of unique samples rather than the reliability of measurements. Further work is needed to understand how these elements of experimental design influence the similarity between artificial and biological neural network activations.

Future work may want to explore non-convolutional architectures as there are a number of reasons why convnets may not be ideal architectures to use with audio spectrogram or chochleogram features. Auditory objects display differently in spectrograms than visual objects in images. In particular, auditory objects tend to be less local than visual objects. The part of the spectogram corresponding to a particular sound object may be distributed across several frequencies and time points. Additionally, auditory objects do not occlude each other as visual objects in images do. Instead, overlapping auditory objects in a spectrogram will combine additively. In this way, the inductive bias of convolutional filters is less appropriate for traditional spectrogram-like features (Wyse, 2017) and thus perhaps less likely to be employed by the brain. Alternative recurrent or autoregressive architectures, which have been very successful in audio synthesis (Oord et al., 2016), may be ideal candidates to investigate in future work.

5. Conclusion

In general, we conclude that DNNs can be good models of animal sensory systems, not because they learn representations that perfectly match our observations of biological neural networks, but because they provide a framework to investigate how architectures, tasks, and learning procedures influence the correspondence between patterns of activity in artificial networks and animal brains. Many more studies will be needed to explore the space of model hyperparameters, tasks and experimental designs.

Bibliography

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv*, pages 1407.5104 [q–bio.NC].

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Barrett, D. G., Morcos, A. S., and Macke, J. H. (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55:55–64.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439).

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method ({CompCor}) for {BOLD} and perfusion based fMRI. *NeuroImage*, 37(1):90–101.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403:309–312.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walke, E. Y., Reimer, J., Bethge, M., Tolias, A. S., and Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Real Neurons & Hidden Units NeurIPS Workshop*.

Cadieu, C., Hong, H., and Yamins, D. L. K. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12):e1003963.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(January).

Cox, R. W. and Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, 10(4-5):171–178.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, 9(2):179–194.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152(October 2016):184–194.

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., D., K. J., Goncalves, M., DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., and Gorgolewski, K. J. (2018a). fMRIPrep. *Software*.

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2018b). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., and Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supple:S102.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, 5:13.

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., Modat, M., Dewey, B. E., Madison, C., di Oleggio Castello, M., Clark, M. G., Dayan, M., Clark, D., Keshavan, A., Pinsard, B., Gramfort, A., Berleant, S., Nielson, D. M., Bougacha, S., Varoquaux, G., Cipollini, B., Markello, R., Rokem, A., Moloney, B., Halchenko, Y. O., Demian, W., Hanke, M., Horea, C., Kaczmarzyk, J., de Hollander, G., DuPre, E., Gillman, A., Mordom, D., Buchanan, C., Tungaraza, R., Pauli, W. M., Iqbal, S., Sikka, S., Mancini, M., Schwartz, Y., Malone, I. B., Dubois, M., Frohlich, C., Welch, D., Forbes, J., Kent, J., Watanabe, A., Cumba, C., Huntenburg, J. M., Kastman, E., Nichols, B. N., Eshaghi, A., Ginsburg, D., Schaefer, A., Acland, B., Giavasis, S., Kleesiek, J., Erickson, D., Küttner, R., Haselgrove, C., Correa, C., Ghayoor, A., Liem, F., Millman, J., Haehn, D., Lai, J., Zhou, D., Blair, R., Glatard, T., Renfro, M., Liu, S., Kahn, A. E., Pérez-García, F., Triplett, W., Lampe, L., Stadler, J., Kong, X.-Z., Hallquist, M., Chetverikov, A., Salvatore, J., Park, A., Poldrack, R. A., Craddock, R. C., Inati, S., Hinds, O., Cooper, G., Perkins, L. N., Marina, A., Mattfeld, A., Noel, M., Snoek, L., Matsubara, K., Cheung, B., Rothmei, S., Urchs, S., Durnez, J., Mertz, F., Geisler, D., Floren, A., Gerhard, S., Sharp, P., Molina-Romero, M., Weinstein, A., Broderick, W., Saase, V., Andberg, S. K., Harms, R., Schlamp, K., Arias, J., Papadopoulos Orfanos, D., Tarbert, C., Tambini, A., De La Vega, A., Nickson, T., Brett, M., Falkiewicz, M., Podranski, K., Linkersdörfer, J., Flandin, G., Ort, E., Shachnev, D., McNamee, D., Davison, A., Varada, J., Schwabacher, I., Pellman, J., Perez-Guevara, M., Khanuja, R., Pannetier, N., McDermottroe, C., and Ghosh, S. (2018). Nipype. *Software*.

Greve, D. N. and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). Generalized Autocalibrating Partially Parallel Acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210.

Güçlü, U., Thielen, J., Hanke, M., and van Gerven, M. A. J. (2016). Brains on Beats. *arXiv*, page 1606.02627.

Güçlü, U. and van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of Neuroscience*, 35(27):10005–10014.

Güçlü, U. and van Gerven, M. A. J. (2016). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, pages 6–13.

Hasson, U., Nastase, S. A., and Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, 105(3):416–434.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2):825–841.

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, 180:101–109.

Kell, A. J. and McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55:121–132.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts

Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11):e1003915.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., and Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, 13(2):e1005350.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. *ICLR workshop on Debugging Machine Learning Models.*

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. in Systems Neuroscience*, 2.

Krizhevsky, A. and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems.*

Kubilius, J. (2017). Predict, then simplify. *NeuroImage*, (October):1–2.

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1):76–85.

Leavitt, M. L., Pieper, F., Sachs, A. J., and Martinez-Trujillo, J. C. (2017). Correlated variability modifies working memory fidelity in primate prefrontal neuronal ensembles. *Proceedings of the National Academy of Science*, 114(12):E2494–E2503.

Lindsay, G. W. (2020). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *arXiv preprint*, page 2001.07092.

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Towards an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14.

Morcos, A. S., Raghu, M., and Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *NeurIPS*.

Norman-Haignere, S., Kanwisher, N. G., and Mcdermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, 88(6):1281–1296.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw

Audio. In *The 9th ISCA Speech Synthesis Workshop.*

Overath, T., McDermott, J. H., Zarate, J. M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, 18(6):903–911.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84(Supplement C):320–341.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement. *NeurIPS*.

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., and Shea-Brown, E. (2019). Dimensionality compression and expansion in Deep Neural Networks.

Reuter, M., Rosas, H. D., and Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, 53(4):1181–1196.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Robert, P. and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Applied Statistics*, 25(3).

Robinson, D. A. (1992). Implications of neural network for how we think about brain function. *Behavioral and Brain Sciences*, 15(4).

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64(1):240–256.

Schönwiesner, M. and Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14611–6.

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, 67(5):1210–1224.

Sitek, K. R., Faruk Gulban, O., Calabrese, E., Johnson, G. A., Ghosh, S. S., and De Martino, F. (2019). Mapping the human subcortical auditory system using histology, post mortem MRI and in vivo MRI at 7T. *eLife*, (8:e48932).

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C. M., and Van Erk, M. J. (2009). Matrix correlations for high-dimensional data: The modified RV-coefficient. *Bioinformatics*, 25(3).

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23.

Steen Moeller, Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. (2010). Multiband Multislice GE-EPI at 7 Tesla, With 16-Fold Acceleration Using Partial Parallel Imaging With Application to High Spatial and Temporal Whole-Brain FMRI. *Magnetic Resonance in Medicine*, 63(5).

Thompson, J. A. F., Bengio, Y., Formisano, E., and Schönwiesner, M. (2016). How can deep learning advance computational modeling of sensory information processing? *NeurIPS workshop on Representation Learning in Artificial and Biological Neural Networks.*

Thompson, J. A. F., Schönwiesner, M., Bengio, Y., and Willett, D. (2019a). How transferable are features in convolutional neural network acoustic models across languages? *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP).*

Thompson, J. A. F., Yoshua Bengio, and Schönwiesner, M. (2019b). The effect of task and training on intermediate representations in convolutional neural networks revealed with modified RV similarity analysis. In *Cognitive Computational Neuroscience*.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Uden, C. V. (2019). *Comparing brain-like representations learned by vanilla , residual , and recurrent CNN architectures*. PhD thesis, Dartmouth College.

Wyse, L. (2017). Audio Spectrogram Representations for Processing with Convolutional Neural Networks. In *Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN*, pages 37–41.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57.

Appendix

5.1. Similarity matrices for each subject

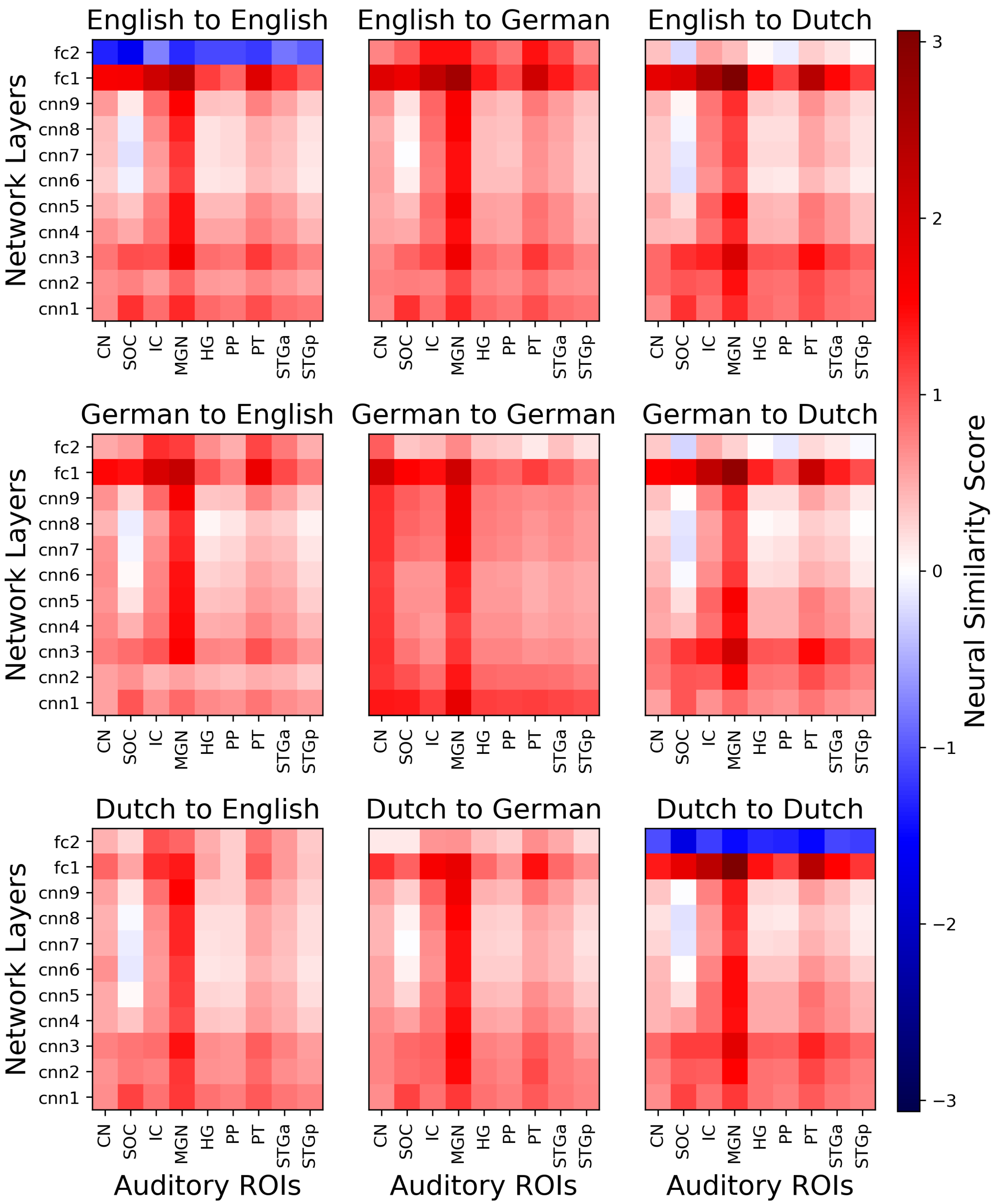


Figure 30. Subject-1 Neural Similarity Matrices. Dutch speaker.

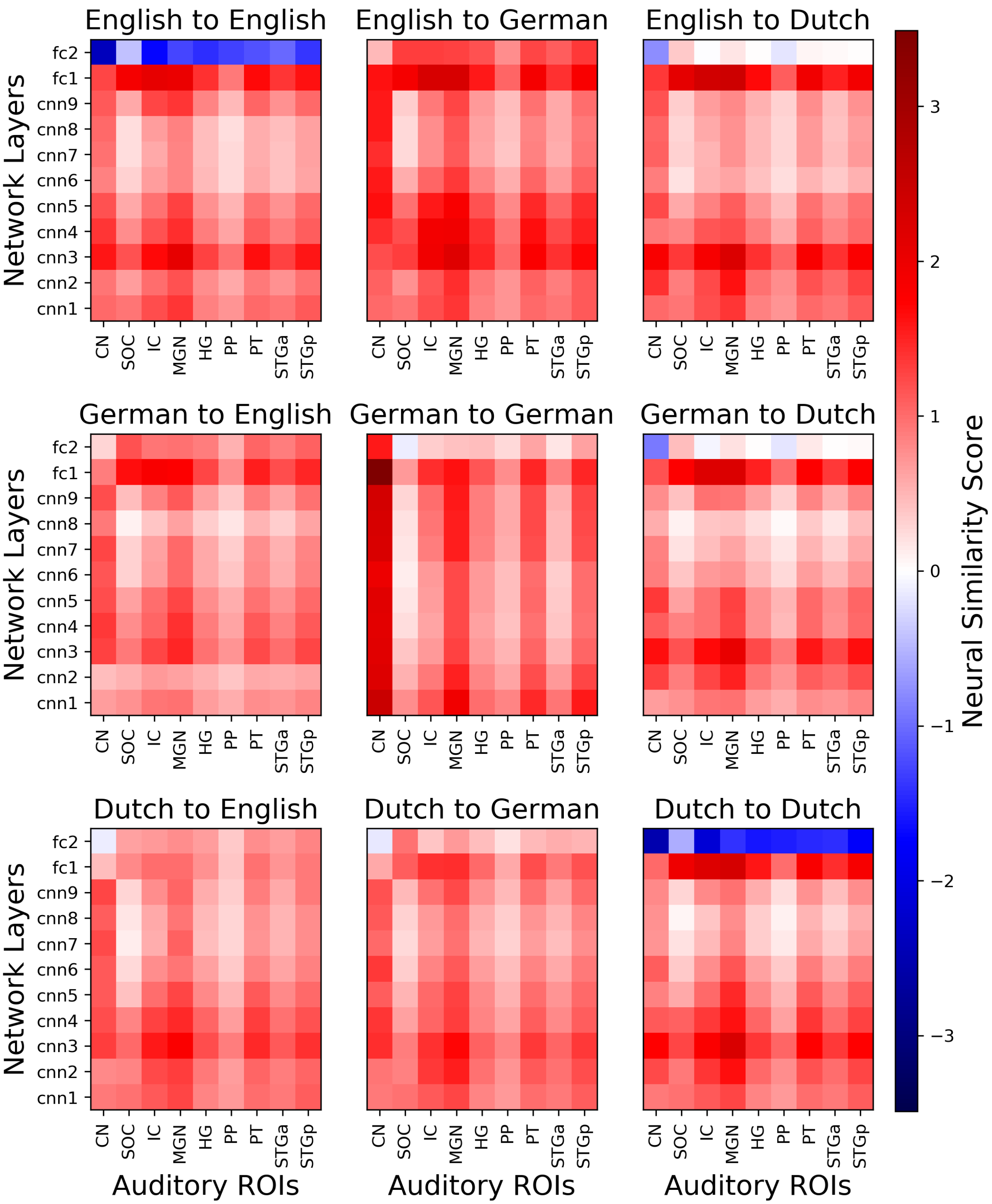


Figure 31. Subject-2 Neural Similarity Matrices. German speaker.

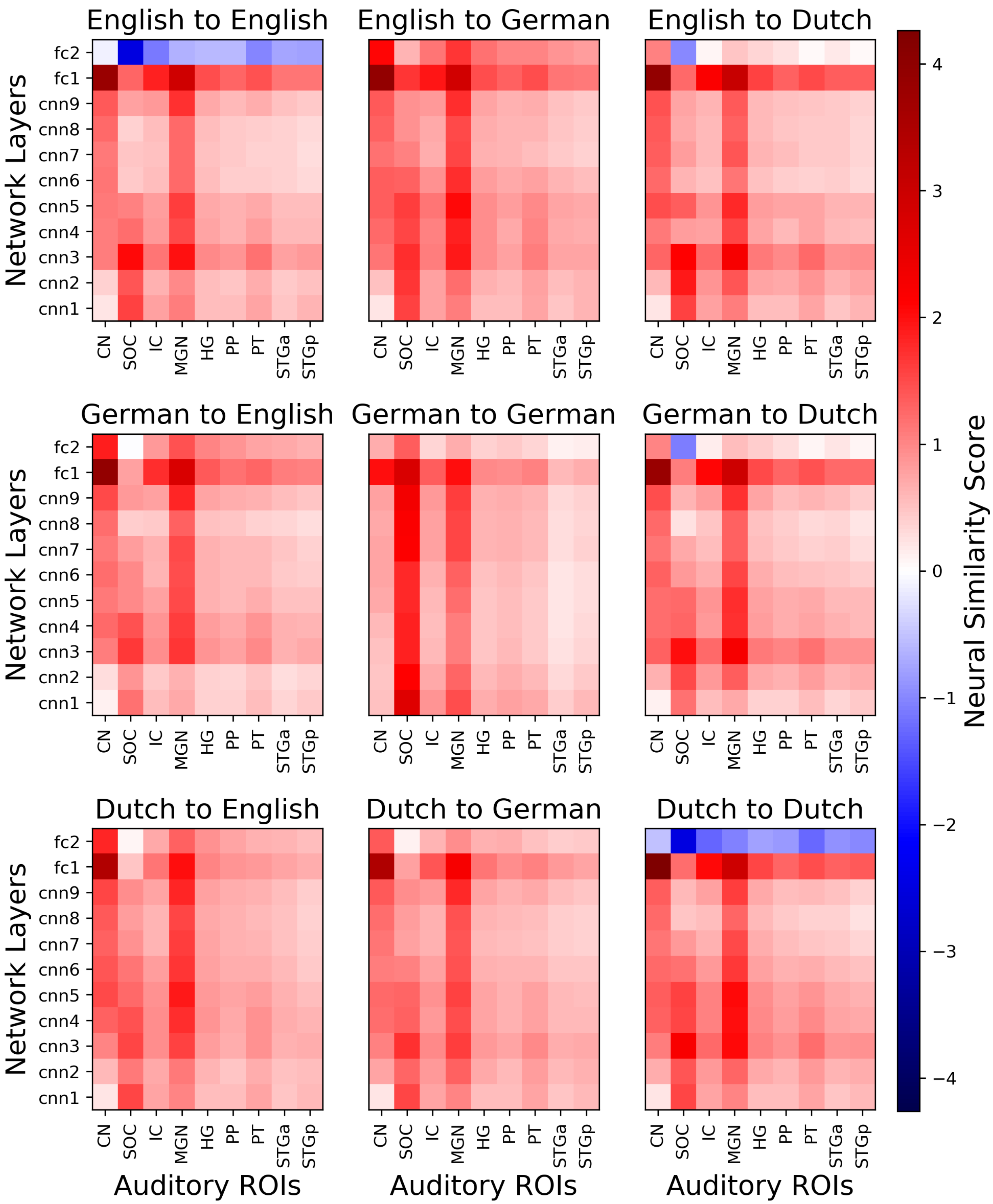


Figure 32. Subject-3 Neural Similarity Matrices. German speaker.

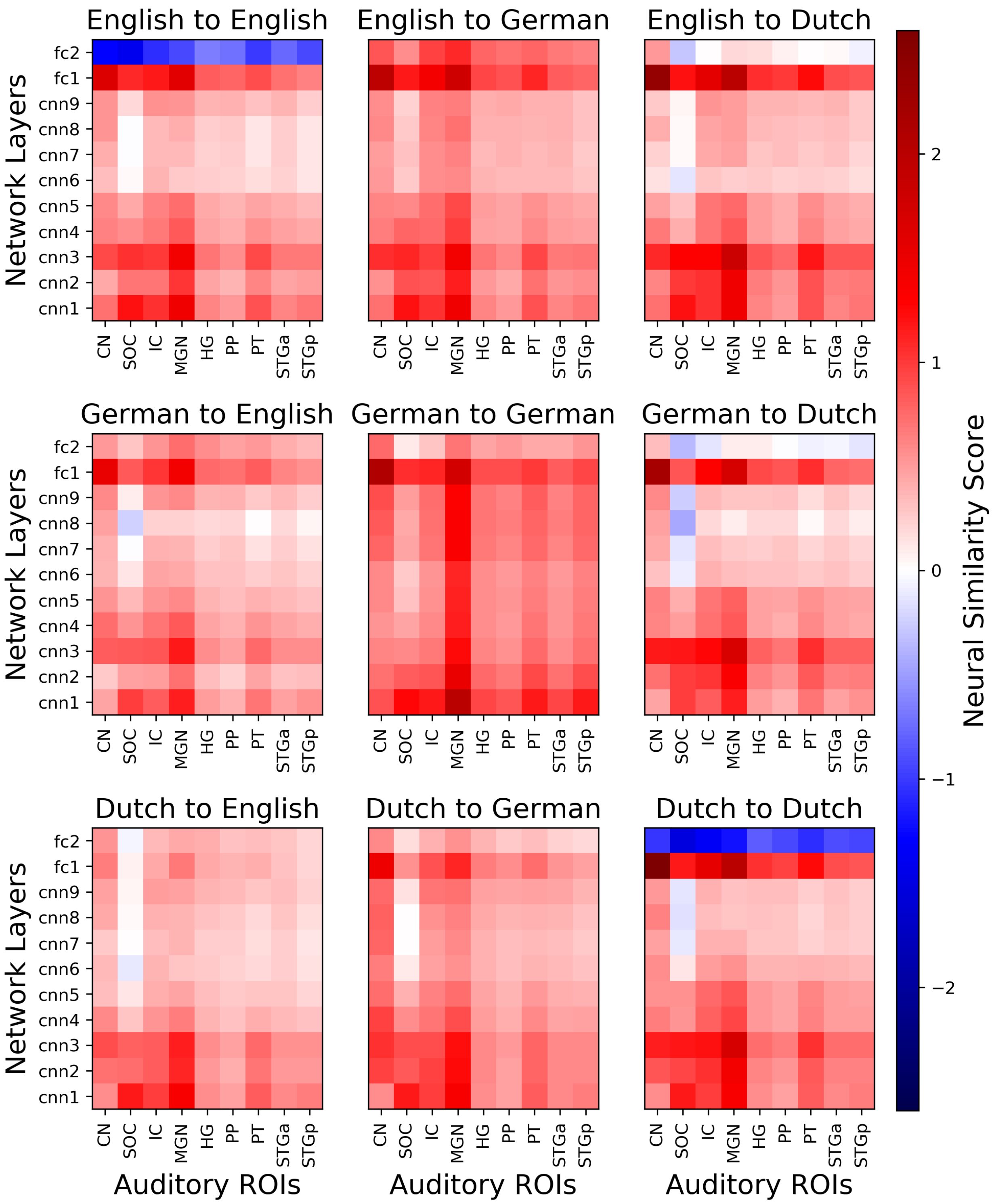


Figure 33. Subject-4 Neural Similarity Matrices. English speaker.

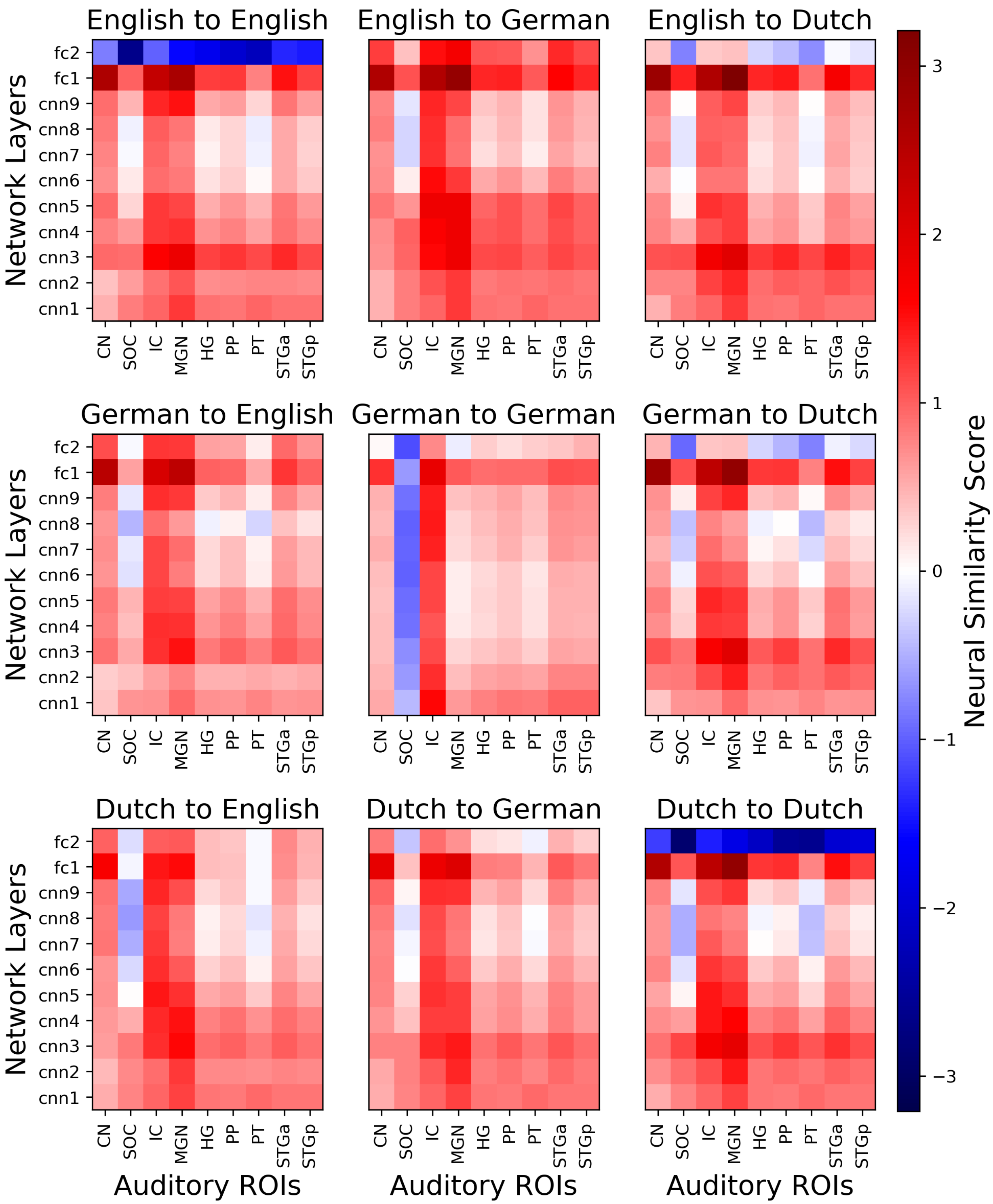


Figure 34. Subject-5 Neural Similarity Matrices. German speaker.

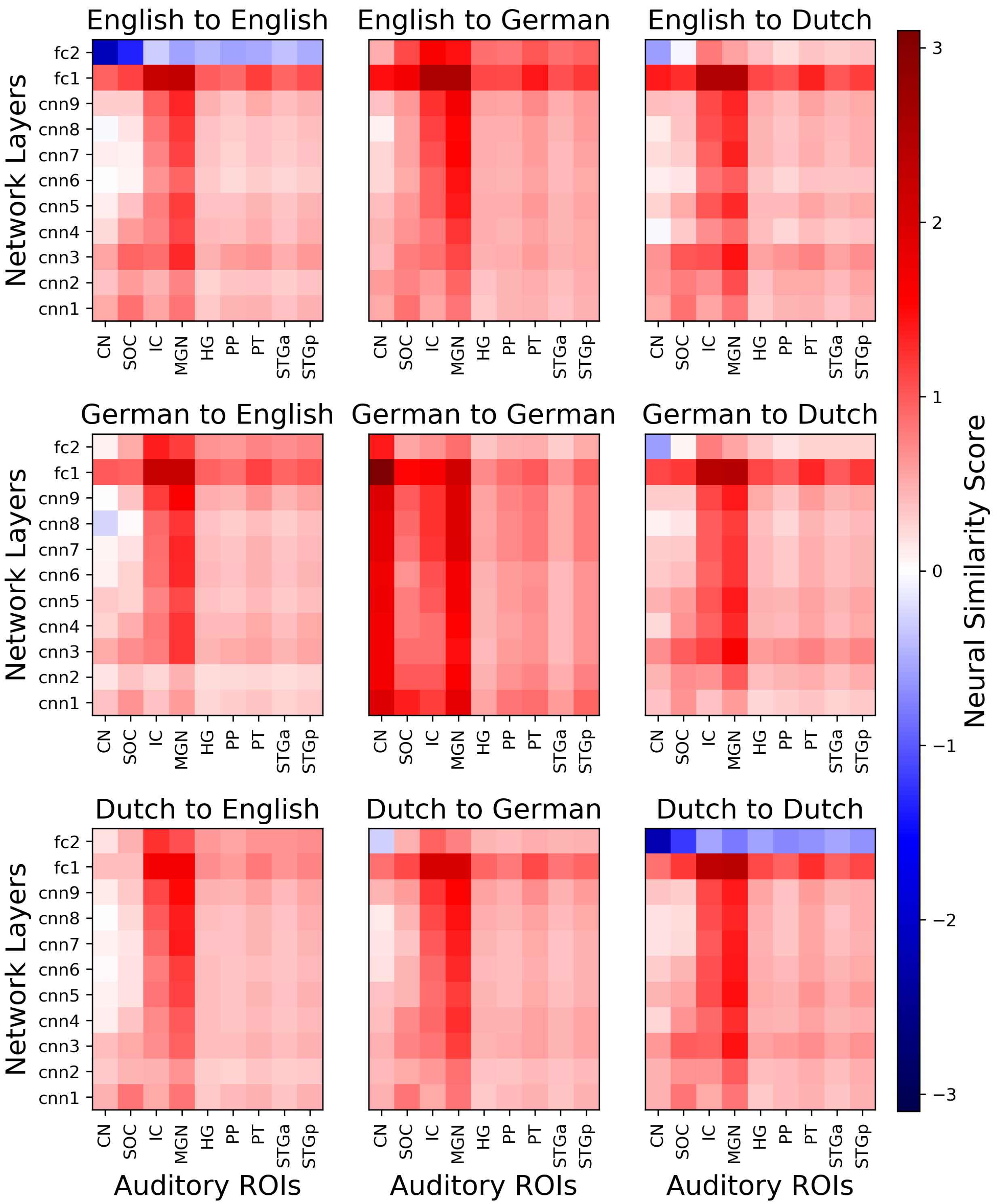


Figure 35. Subject-6 Neural Similarity Matrices. Dutch speaker.

Discussion

6.1. Summary of main findings

The scientific goals of this thesis were to characterize and compare auditory representations in DNNs and in the human auditory pathway. The related meta-scientific goals were to unify computational neuroscience and deep learning science approaches to studying neural computation and to address the various methodological and philosophical issues that challenge this unification. Towards these goals, this thesis has explored several approaches to modeling auditory neural processing and considered the philosophical underpinnings of these various approaches. Article 2 reflects a functional modeling approach and a functional few of explanation in neuroscience. Articles 3 and 4 analyse representations in DNNs, which can be viewed as 'model organisms' in the same way that scientists study non-human animals as models of specific phenomena of interest. Article 5 reflects a normative approach, whereby instead of attempting to characterize the specific mathematical function computed by some neural population (as in Article 2), the goal is to characterize the necessary and sufficient conditions under which such a function can be learned. This approach can be (but is not necessarily) ambivalent to the nature of the specific function being computed, and rather focuses on the optimization procedures and network architectures that yield brain-like representations. Article 1 provides the necessary philosophical vocabulary to reason about the merits of theses various approaches and to begin to envision a unified science of intelligence concerned with phenomena at the intersection of neuroscience and artificial intelligence.

The functional approach employed in Article 2 sought to characterize how well cortical responses to sound can be modeled as a linear function of spectrotemporal modulations. Our stimuli were highly similar, such that the encoding and decoding analyses were not only testing our hypothesis but also the limits of the acuity and sensitivity of 7T fMRI. Despite the small differences between our stimuli, we found that, on average, linear predictive models trained only on simple ripples can generalize to mixtures of ripples, supporting the hypothesis that cortical responses to sound can be modeled as a linear function of spectrotemporal modulations. This result is consistent with the well established model of cortical auditory responses being selective to patterns of spectrotemporal modulation. However, this model is a simplification that we know to be false (like all models). The primary aim of this project was to localize where (for which voxels) this model was least wrong. We were unable to achieve such maps due to low SNR. The demands of our experimental design exceeded the sensitivity limits of 7T fMRI. Thus, the primary implications of our findings concern experimental design for 7T auditory fMRI experiments.

If using synthetic stimuli instead of natural stimuli, one must ensure that they elicit sufficiently distinct patterns of neural activity. When possible, it is likely preferable to use natural sounds. In some cases it may be possible to construct modified versions of natural sounds that possess the desired attributes, while still eliciting large magnitude neural responses. However, recent results have shown that synthetic stimuli do not necessarily elicit smaller magnitude responses. Ponce et al. (2019); Bashivan et al. (2019) synthesized images with neural networks specifically to maximize the response of individual neurons. In both cases, the authors found synthetic images that did not resemble natural images but that none the less drove visual neurons exceptionally well, often far exceeding the neuron's firing in response to a set of natural stimuli. Thus, there may not be anything intrinsic about synthetic stimuli that elicit smaller responses. One could attempt to use similar approaches to maximize the response of individual voxels, but the efforts would likely only be successful to the extent that neurons within a voxel share the same selectivity.

Much of basic neuroscience research focuses on documenting the selective responses of individual neurons or populations of neurons. Such approaches have also been employed to characterize representations in deep neural networks. For example, visualization techniques can be used to characterize what types of input maximally activate a unit and attribution techniques can trace how the activation of a given neuron affects the output of the network (Olah et al., 2018). However, some have questioned the usefulness and limits of such characterizations (Richards et al., 2019; Lillicrap and Kording, 2019; Jonas and Kording, 2017). In machine learning, the problem of interpretable and explainable AI is typically viewed as problem of human computer interaction rather than an integral part of a science of machine learning. Characterization of the selectivity of individual units in a network or how they contribute to a decision of classification, may indeed be critical for certain applications or may aid in debugging why a network is making mistakes. However, it does not generally appear to provide actionable insight that could be leveraged to design new algorithms or architectures. Especially for intermediate representations that reflect neither physical properties of the world nor semantic categories of objects, it is entirely possible that whatever individual units or neurons respond to is not easily describable in human language and may not map on to existing human concepts. Humans may eventually develop language and concepts for these *in between* representations, but until then, it may be worth exploring what other types of characterizations are possible.

Articles 3 and 4 explore alternative characterizations at the level of single layers that are not based on selectivity. In Article 3 we quantified the language specificity of layers of acoustic models trained on different languages and explored the role of weight freezing in transfer learning. Article 4 focused on neural similarity analysis as a tool to characterize representations. We evaluated the influence of several modifications to the training procedure on the learned representations by comparing the resulting activations to those of the standard-trained monolingual networks from Article 3 as a reference. Overall, we found that the largest representational changes occurred in the last two layers, which we also found to be most language-specific. The representations of models trained on different languages were largely similar up until the last two layers, which was consistent with the results of Article 3 which found the final two layers to be most language-specific. Our analysis of the representational similarity of networks with random networks led us to hypothesize that both the input and target serve as anchors on the representational form, such that the intermediate layers have the most flexibility in what form they can take without negatively affecting performance.

These results are consistent with the observation made by Raghu et al. (2017) that neural networks converge *bottom-up* with shallow layers converging to their final representation earlier in training. The plots from that paper seem to suggest a top-down component as well, with intermediate layers converging after the deepest layers. However, this story is inconsistent with recent results from Mehrer et al. (2020) which showed that the variability in representation due to different random initializations increases as a function of depth. Our hypothesis would have predicted an inverted U-shaped relationship where intermediate representations are most variable. Some of this discrepancy can be attributed to the choice of similarity metric. Raghu et al. (2017) use SVCCA which they show to be invariant to different random initializations as a demonstration of its appropriateness for neural similarity analysis. Similarly, CKA-based similarity has been validated by showing that it can recover corresponding layers of networks trained from different random initializations. Mehrer et al. (2020) use RSA with correlation distance-based DSMs and Kandall's Tau to compare DSMs. Since correlation distance is not invariant to orthogonal linear transforms (as SVCCA and CKA is), it may be more sensitive to the types of representational variation that occur at deeper layers of a network. For example, RSA with correlation distance would not be invariant to a permutation of the output units, whereas SVCCA and CKA would be. Whether such rotations are relevant likely depends on the research question.

It is a limitation of this thesis that variance across several random initializations is not reported. This is primarily due to the computational costs of training. We analyzed 198 networks which each took approximately two-weeks to train. While we were able to train several networks in parallel, it would not have been feasible to train all 198 models several times from different random initializations. In Article 3, we repeat the analysis for all 3 language pairs and find a similar pattern for each pair. Thus, it is improbable that our results are attributable to different random initializations. However, smaller networks that train faster could have been used. Such networks would have performed less well on the phoneme recognition task. We prioritized performance because we wanted to make sure the networks learned as much as possible about each language. However, since most of our analyses are based on relative differences in accuracy rather than on absolute performance, we may have found similar results with smaller, quicker training networks.

The approaches explored in Article 3 and Article 4 cannot easily be directly applied to characterize biological neural representations because one does not have the same control over biological brains. However, one can compare DNN activity to neural activity. This is the analysis we employ in Article 5 where the subset of the networks that were the object of study in Articles 3 and 4 are compared to activity in ROIs throughout the human auditory pathway. In particular, we sought to assess whether the human auditory pathway possess a similar representational hierarchy as our DNNs, as has been previously reported in the auditory and visual systems. While we did observe considerable similarity between our networks and our ROIs, we did not find any evidence for a shared representational hierarchy. This work is the first to compare DNNs trained on auditory tasks to subcortical and cortical regions of the auditory pathway, and as such, represents a powerful test of the hypothesis of shared hierarchy. That we did not observe different patterns of similarity for our subcortical and cortical ROIs is a clear indicator that the human auditory pathway is employing a different representational trajectory than our DNNs.

Article 5 demonstrates that it is non-trivial to train DNNs that possess a similar representational hierarchy as the human auditory pathway. Similar results in the visual domain have been reported (Cadena et al., 2019). However, we do not conclude, as Xu (2020) do, that DNNs thus may not serve as sound working models of sensory systems. DNNs are promising models of sensory processing not because they always perfectly mimic neural measurements, but because they provide a framework to investigate the necessary and sufficient conditions to learn brain-like representations and to bridge knowledge about intelligence in artificial and biological systems. In fact, if all DNNs learned representations that mimicked the hierarchical structure of sensory systems, they would be less useful for studying the conditions that yield such hierarchies. To paraphrase Eve Marder giving a keynote lecture at the 2019 Bernstein Conference on computational neuroscience, a good model is not one that ends up being perfectly correct, but one that makes you realize that you should be doing a different experiment that you would not have thought of otherwise (Marder, 2015). In our case, we observed that all ROIs were most similar to the same layer: the first fully connected layer. This motivates followup experiments to understand why that layer in particular appears to be the most brain like. Characterizing the nature of that layer in particular may inspire alternative objective functions that yield representational hierarchies that are a better match to the human auditory system. In other words, the ultimate goal of this line of research is not simply to find a model that achieves high similarity or that explains as much variance as possible, but to identify the specific attributes of network design—what specifically about what architectures and what learning procedures—yield the types of representations we observe in biological sensory systems.

6.2. On using deep neural networks as models of sensory systems

The use of DNNs as models of animal sensory systems is largely in the context of a model comparison approach to scientific discovery. Through evaluation, comparison, and iterative refinement, models hopefully get closer to some truth about the phenomena under study. [12] Within this view, models that are more constrained are likely to be closer to the truth since they occupy smaller region of the search space known to include the true model (Fig 36). In practice, however, we don't usually know which constraints are necessary to answer a specific scientific question.

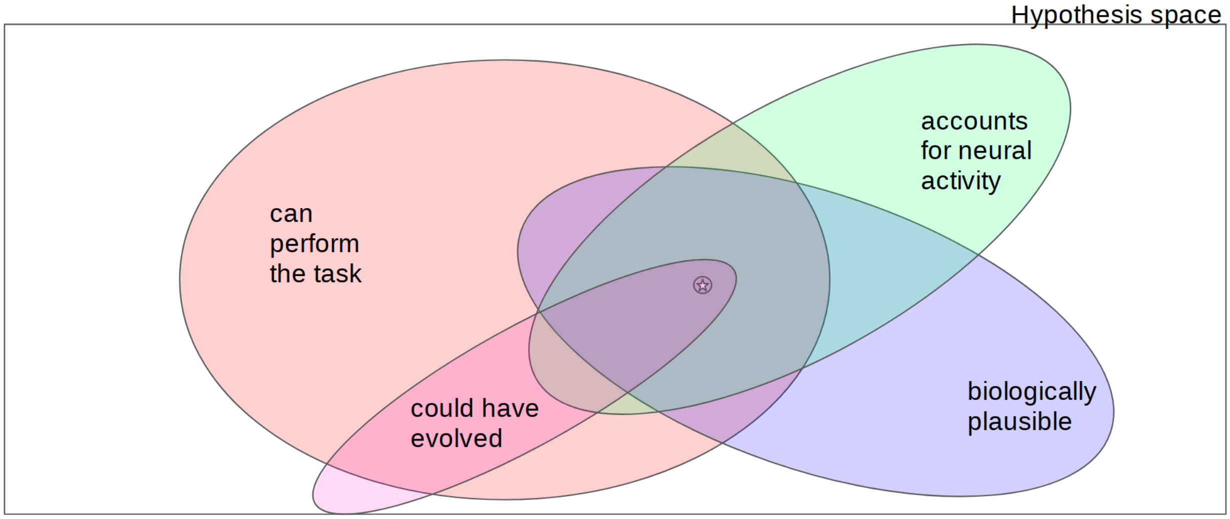


Figure 36. Models lie at the intersection of one or more constraints The rectangle indicates the space of all possible models where each point in the space represents a different model of some phenomenon. Regions within the colored ovals correspond to models that satisfy specific specific model constraints (where satisfaction could be defined as meeting some threshold on a continuous value). If a constraint is well justified, it implies that the true model is contained within the set of models that satisfy that constraint. Models that meet more constraints, then, are more likely to live within a smaller region of the hypothesis space and hence will be closer to the truth, indicated by the star in this diagram.

The use of DNNs as models of sensory systems emphasizes a different subset of constraints than alternative modeling approaches. For example, Kell and McDermott (2019) discuss the importance of task constraints and of models that exhibit the phenomenon to be explained,

[12] Or at least the models become more useful, if one prefers a more pragmatic, less realist account

e.g. if one wants to study face recognition, a reasonable possible model constraint is that the model be capable of recognizing faces. Emphasizing task-performance and accounting for animal behaviour may come at the expense of other possible model constraints since it is usually impossible to satisfy all model constraints at once.

The various differences between DNNs and biological brains are often repeated to refute their usefulness as models. In particular, the biological (im)plausibility of DNN models and their limited ability to replicate high-level cognition are often cited. Marcus (2018) describes the limitations of current DNNs. They are not capable of relational reasoning, cannot accommodate non-stationarity, do not extrapolate well, and cannot separate correlation from causation, among other limitations that human brains have managed to overcome. Zador (2019) questions the relevance of models that required large datasets to learn when most animal behaviour is the result of eons of evolution and encoded in the genome rather than learned over the course of a lifetime. DNNs and animals fail in different ways. DNNs are susceptible to adversarial examples—examples that have been only slightly modified such that the differences are not noticeable by humans, but can severely affect the performance of a network (Goodfellow et al., 2015).[13] Since DNNs are only loosely inspired by biological neural networks, there are enumerable physiological details that are missing. Biological plausibility has been presented as a requirement for models to be useful for studying biological neural computation (Gerven and Bohte, 2017) and the biological plausibility of learning in DNNs has been questioned. There are many differences between DNNs and biological brains, but what implication they have on how we think about DNNs as models? Why are these differences meaningful?

Criticisms of the use of DNNs as models of sensory systems often amount to claims that a different subset of constraints should be privileged. The proposed requirement that models must be biologically plausible in order to have bearing on neuroscience prioritizes the purple region of Figure 36 which contains only models that are deemed biologically plausible. According to the view put forth by Love (2019), positions of this type reflect value judgements about which datasets are most important. On what basis are such value judgements made? Within the model comparison framework, constraints (or datasets) are selected to narrow the search space. Constraints could be privileged based on how much they narrow the search space. However, when comparing two constraints like biological plausibility and task performance, it is not obvious that one will narrow the search space more than the other. It is entirely possible that exploring the space of possible models that can perform some task of interest will lead to truth faster than exploring the set of models that are biologically plausible. These known unknowns can inform how we think about optimizing scientific progress in a model comparison framework.

[13] Although there has been some suggestion that there exist adversarial examples that fool both DNNs and humans c.f. Elsayed et al. (2018)

We can try to reason about which constraints are more limiting and it may be more or less possible for different research questions. In Article 1, I emphasized the importance of specifying the phenomenon to be explained. Similarly, Love (2019) emphases a similar need to identify the datasets to be accounted for. Different researchers, even researchers who are concerned with the same natural phenomenon, may still choose to privilege different model constraints and this is a feature, not a bug. Due to our uncertainty about the nature of the hypothesis space to be explored, we need different researchers to come at the same problem from as many different angles as possible. This has been studied using simulations of scientific discovery in a model-centric framework to identify the relationship between several attributes of scientific communities and the success of their research program. Devezer et al. (2019) found that innovative research speeds up the discovery of scientific truth by facilitating the exploration of model space and that epistemic diversity, the use of several research strategies, optimizes scientific discovery by protecting against ineffective research strategies. The authors compare epistemic diversity to diversifying an investment portfolio to reduce risk while trying to optimize returns. If one knew how the market was going to change, one wouldn't need a diverse investment portfolio. Similarly, uncertainty about scientific truth and how to search for it should lead us to embrace epistemic diversity.

The long list of differences between DNNs and brains has no general implication for the suitability of DNNs as models of biological intelligence and learning. Specific differences may be relevant to specific research questions. Nevertheless, researchers are currently working on addressing several of these differences to further narrow the model search space. Machine learning researchers are currently working on biologically plausible learning algorithms (Bengio et al., 2014; Lillicrap et al., 2014; Guerguiev et al., 2017), relational reasoning (Bahdanau et al., 2018; Santoro et al., 2017), and causal inference (Schölkopf, 2019; Goyal et al., 2019). Neuro-AI researchers have been exploring the effects of adding elements of biological realism to DNNs to see how they affect representational correspondence (Lindsay and Miller, 2018; Lindsay et al., 2019). Storrs and Kriegeskorte (2020) hypothesize that, as the field of deep learning continues to progress, neural network models will only become more relevant and useful for cognitive neuroscience. They discuss how the study of relational reasoning in artificial systems helps to identify the necessary and sufficient conditions for such abilities to develop and how artificial systems trained in simulated environments can be used as a tool for studying embodied cognition. The use of DNNs as models of biological neural system is one of several well-justified modeling approaches. DNN models focus on different regions of model space than alternative approaches, and thus constitute an innovative strategy that increases the epistemic diversity of computational neuroscience.

6.3. Unifying Neuroscience and AI: Disambiguating prediction, representation and explanation

Many terms are used in different ways at the intersection of neuroscience, AI and philosophy of science. An integration of neuro and AI will require a consistent language. Here, I try to map between related concepts in cognitive science, statistics and machine learning.

6.3.1. Representation and encoding

Much effort has been directed at representations and their role in cognition and explanation. Marr and Nishihara (1978) defines a representation as “a formal system for making explicit certain entities or types of information, together with a specification of how the system does this.” He denotes a specific instance of an entity in a given representational system as a *description*. For example, the Arabic numeral 37 is a description of the number 37 that makes explicit its decomposition into powers of ten. A binary representation of the same number would make explicit its decomposition into powers of two. A representation will often be a useful abstraction. For example, we can represent strands of DNA as sequences of nucleotides, represented by the letters A, T, C and G. Similarly, the information processing approach to cognitive neuroscience presumes that the brain is likely to use various representations of sensory information at different stages along some pathway to facilitate certain computations. The terms *encoding* and *decoding* refer to representational transformations from the sensory input (encoding) and to perception or behaviour (decoding). According to Diedrichsen and Kriegeskorte (2017), information-based analyses of neural measurements (encoding analysis, decoding analysis, representational similarity analysis, etc.) test representational models, which describe how patterns of activity relate to sensory stimuli, motor actions, or cognitive processes. Their definition of representation within this framework is that a represented variable can be linearly decoded from a down-stream area. This paradigm, sometimes referred to as neural coding, has led researchers to make statements about what is ’encoded’ in neural signals based on the results of encoding and decoding analyses.

This paradigm has received criticisms on several fronts. Brette (2019) points out that the language of the neural coding framework implies causal relationships for which the analysis typically does not provide evidence. That the activity of a population of neurons can be well predicted by a particular representational model does not in itself imply that the hypothesized representation is in fact used by the neural system to accomplish the task of interest. Many candidate representational models may predict the relevant neural activity equally well. Using predictive performance as the only arbiter of model fit does not establish the causal relevance of the hypothesized representation. This debate reflects tensions between functional and causal mechanical theories of explanation. The neural coding paradigm entails the functional analysis of a neural system: decomposition of the component operations of a phenomenon. According to the functional theory of explanation, the causal mechanical implementation of those component operations are not needed. Although not stated explicitly, in essence, Brette's warning regarding the interpretation of results in the neural coding paradigm reflect a warning against a functionalist view of explanation in neuroscience.

The neural coding paradigm has also received criticism from the dynamical camp. The dynamical hypothesis, is that 'cognitive agents are dynamical systems' (Gelder, 1998). The antirepresentational stance adopted by some dynamicists and radical embodied cognitive scientists claims that cognition is not inherently representational (Chemero, 2009): "Unlike digital computers, dynamical systems are not inherently representational. A small but influential contingent of dynamicists have found the notion of representation to be dispensable or even a hindrance for their particular purposes. Dynamics forms a powerful framework for developing models of cognition that sidestep representation altogether" (Gelder, 1998, 622). A dynamical explanation may make no reference to representation and instead describe the details of a particular neural circuit, for example.

The definition of representation in cognitive science and neuroscience is distinct from the notion of representation in machine learning. The field of representation learning is concerned with procedures for automatically learning useful transformations of data. The input data, say a set of images, are originally represented by a set of three-dimensional (RBG) pixel values. This pixel space is one representational space. Learned representations will consist of one or more transformations of this original form. In this sense, machine learning representations are representations of some signal whereas in cognitive science literature, a representation is a representation of some variable. In a DNN classifer, the target could be seen as a variable of interest. From the data processing inequality, we know that the mutual information with the target will be maximal at the input layer. All the information related to the target class is present at the input. The subsequent representational transformations change the form of that information, gradually linearizing the decision boundaries, such that the target class can be linearly read out at the output layer. One can add linear classifier probes at each layer of a deep network to see how well the target class can be decoded from each layer. For a trained network, one should see that the performance of these linear probes will increase with depth, but the decoding performance could be above chance at all depths (Alain and Bengio, 2016). In this case, where would the cognitive neuroscientist say the target is represented? At every layer? Or maybe at the input since that is where the mutual information is greatest? Or at the final layer since the decoding accuracy is highest there? From a machine learning perspective, what can be linearly decoded from a layer's activity only provides a snapshot of its representational form.

6.3.2. Prediction, explanation, and generalization

In machine learning, the output of a model is a prediction. In classification, the prediction takes the form of a categorical label which represents the model's best guess of the category of the input example. Traditionally, the goal of supervised machine learning is to discover statistical regularities and invariances in the training data that enable accurate predictions for a given task. The data are typically assumed to be independently and identically distributed (i.i.d); all observations are sampled independently from the same data generating process. The goal is a model with good generalization performance, which means that the predictions are accurate for any other sample from that data generating process. A model that overfits to the training data will not generalize well. For some models, there are analytic bounds on the generalization gap. In practice, this is typically verified empirically by separating datasets into training and testing sets. The performance on the test set estimates how well the model would predict any random sample from the same data generating process; this is refered to as within-distribution generalization. Some efforts in machine learning are focused instead on out-of-distribution generalization, which refers to the setting where the training and test sets are not i.i.d.

One example of out-of-distribution generalization is systematic generalization in language, which refers to the ability to rationalize about logical rather than purely statistical relationships between tokens. For example, (Bahdanau et al., 2018) investigated the ability 'to reason about all possible object combinations despite being trained on a very small subset of them':

> Clearly, given known objects X, Y and a known relation R, a human can easily verify whether or not the objects X and Y are in relation R. Some instances of such queries are common in daily life (is there a cup on the table), some are extremely rare (is there a violin under the car), and some are unlikely but have similar, more likely counter-parts (is there grass on the frisbee vs is there a frisbee on the grass). Still, a person can easily answer these questions by understanding them as just the composition of the three separate concepts. Such compositional reasoning skills are clearly required for language understanding models.

Systematic generalization is something that is relatively easy for humans but difficult for artificial natural language understanding systems. Out-of-distribtion generalization also shows up in other applications. For example, one may wish to train a robotic arm first in a simulated environment controlled by a physics engine and want it to generalize to the real-world. Out-of-distribution generalization is one of the frontiers of AI research at the moment and will be required for AI systems to mimic human cognitive abilities. In this way, not all predictions are equal. Different predictions will test different generalizations.

In statistical hypothesis testing, commonly employed in the analysis of neural data, the word predict is employed in a different way. One variable is said to *predict* another if a significant statistical relationship has been found between the two. This use of the term is more akin to what philosophers call *accommodation*: how well a scientific theory accommodate the data that was already known at the time the scientific theory was constructed. When regression is used for statistical hypothesis testing, one variable (or set of variables or intersection of variables) is said to predict another based on an assessment of the experimental data. When using regression in machine learning, the model as a whole is predicting the target. The model is evaluated by how well the model predicts held out data (data that was not used during the training of the model). However, neither of these uses seems to parallel the use of *predict* in philosophy of science where the emphasis is on the prediction being novel, i.e., something that hasn't been observed yet.

Confusingly, the word *explain* is also used in the context of statistical hypothesis testing. The statistical measure R-squared (R^2) is the proportion of variance in one variable that is *explained* by another in a linear regression. This use of the word *explain* in statistical hypothesis testing is distinct from scientific explanation, but the two are sometimes not clearly distinguished in scientific writing. For example, consider this motivating statement for the Algonauts project, whose 2019 edition is dedicated to "Explaining the Human Visual Brain":

> Currently, particular deep neural networks trained with the engineering goal to recognize objects in images do best in accounting for brain activity during visual object recognition (Schrimpf et al., 2018; Bashivan, Kar, & DiCarlo, 2019). However, a large portion of the signal measured in the brain remains unexplained. This is so because we do not have models that capture the mechanisms of the human brain well enough. Thus, what is needed are advances in computational modelling to better explain brain activity (Cichy et al., 2019).

The authors allude to statistical explanation when discussing unexplained signals, while talk of capturing neural mechanisms hints to scientific explanation. When in reality, this project is about evaluating models based on their ability to predict (in the machine learning sense) neural activity. When they lament that a "large portion of the signal measured in the brain remains unexplained", they invoke the notion of explained variance. Rather than trying to develop a scientific explanation for a phenomenon of interest, they are concerned with statistically explaining, or in this case, being able to predict, the variance in the collected data—variance which may or may not be causally related to any number of different neural or cognitive phenomena.

Many of the issues described above can be subsumed under the notion of *generalization*. The philosopher's term *accommodation* does not imply any generalization beyond the observations used during the construction of the theory (or training of the model). The typical notion of generalization in psychology is akin to within-distribution generalization in machine learning. One assumes (or tries to ensure) that their sample of subjects represents a random sample from a population. The goal of statistical inference is to make general statements about the population from the measurements made on a sample. The notion of novel prediction in philosophy of science could be seen as an example of out-of-distribution generalization in machine learning.

The goal to explain as much variance as possible or to predict as accurately as possible expresses a desire for completeness. Philosophers of scientific explanation warn against over-completeness.

> It is important to note, in this connection, that particular [explananda] do not necessarily embody all of the features of the phenomena which are involved. For example, archaeologists are attempting to explain the presence of a particular worked bone at a site in Alaska. The relevant feature of the explanandum are the fact that the bone is thirty thousand years old, the fact that it was worked by a human artisan, and the fact that it has been deposited in an Alaskan site. Many other features are irrelevant to this sought-after explanation. The orientation of the bone with respect to the cardinal points of the compass at the time it was discovered, its precise size and shape (beyond the fact that it was worked), and the distance of the site from the nearest stream are all irrelevant. It is important to realize that we cannot aspire to explain particular phenomena in their full particularity. . . . In explanations of particular phenomena, the explanation-seeking why-question—suitably clarified and reformulated if necessary—should indicate those aspects of the phenomena for which an explanation is sought. (Salmon, 1984, pg.273-4)

The project of collecting large-scale neural datasets and building models that explain as much variance as possible in that data is one of mere description rather than explanation. Descriptive science is unambiguously crucially important to scientific progress. Recall the first aspect of Craver's notion of mechanical explanations is characterization of the phenomenon to be explained. However, the distinction between explanation and mere description is still important. Specific why-questions may eventually be motivated by such descriptive characterizations, but only if we don't mistake them for explanations prematurely.

6.4. General Conclusions

Comparing activations in biological and artificial neural networks is a promising approach to study the architectures and learning procedures that support brain-like representations and the nature of representations in intelligent systems. However, this scientific project is not just about chasing high accuracy. As much as one might like to, the scientific problems posed in neuroscience cannot be reformulated as engineering problems. The (long term) goal of science is to generate scientific explanations, which is not the same as statistically explaining the variance in our data. At this particular moment in computational neuroscience, there is a high degree of uncertainty about what such explanations might look like for many phenomena of interest. Therefore, the field may benefit from a closer relationship with philosophers of neuroscience concerned with scientific explanation. Philosophy of science offers conceptual scaffolds that can help refine a vision of an integrated science of intelligence.

Part of the value of deep learning from a neuroscience perspective could come from the fact that deep learning is theory-poor compared to other areas of machine learning. That there are a lot of open questions in deep learning theory may reflect generic challenges of studying learning in distributed networks. This positions deep learning science as a model science for neuroscience. The methods and concepts that prove useful for explaining phenomena in deep learning may inspire new methods and ways of thinking in neuroscience, due to the similar scientific problems posed in these two fields and the relatively ease with which artificial systems can be analyzed compared to their biological counterparts. In this way, the opportunities for transfer between deep learning and neuroscience span several scientific and meta-scientific levels.

The arguments presented here are not intended to advocate for a deep learning approach to neuroscience over other approaches. The purpose of the arguments presented is to justify and clarify the merits of a deep learning approach to neuroscience as one among many. The addition of a deep learning approach increasing the epistemic diversity of the set approaches employed. Innovative and diverse approaches in an epistemically humble research community will better lead us towards truth.

Bibliography

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv*, page 1610.01644v3.

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. (2018). Systematic Generalization: What Is Required and Can It Be Learned? pages 1–16.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439).

Bengio, Y., Lee, D.-h., Bornschein, J., and Lin, Z. (2014). Towards Biologically Plausible Deep Learning.

Brette, R. (2019). Neural coding : the bureaucratic model of the brain.

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walke, E. Y., Reimer, J., Bethge, M., Tolias, A. S., and Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Real Neurons & Hidden Units NeurIPS Workshop*.

Chemero, A. (2009). *Radical Embodied Cognitive Science.* The MIT Press, Cambridge, MA.

Cichy, R. M., Roig, G., Andonian, A., Dwivedi, K., Lahner, B., Lascelles, A., Mohsenzadeh, Y., Ramakrishnan, K., and Oliva, A. (2019). The Algonauts Project: A Platform for Communication between the Science of Biological and Artificial Intelligence. In *Computational Cognitive Neuroscience.*

Devezer, B., Nardin, L. G., Baumgaertner, B., and Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE*, 14(5):1–23.

Diedrichsen, J. and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, 13(4).

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial Examples that Fool both Human and Computer Vision.

Gelder, T. v. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioural and Brain Sciences*, 21:615–665.

Gerven, M. v. and Bohte, S. (2017). Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Front. Comput. Neurosci.*, 11(114).

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.*

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. (2019). Recurrent Independent Mechanisms.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *eLife*, 6(e22901).

Jonas, E. and Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*, 13(1).

Kell, A. J. and McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55:121–132.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2014). Random feedback weights support learning in deep neural networks. page 14.

Lillicrap, T. P. and Kording, K. P. (2019). What does it mean to understand a neural network? page arXiv preprint: 1907.06374.

Lindsay, G., Moskovitz, T., Yang, G. R., and Miller, K. (2019). Do Biologically-Realistic Recurrent Architectures Produce Biologically-Realistic Models? pages 779–782.

Lindsay, G. W. and Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, 7:1–29.

Love, B. C. (2019). Levels of Biological Plausibility. *PsyArXiv Preprints*.

Marcus, G. (2018). Deep Learning: A Critical Appraisal.

Marder, E. (2015). Understanding Brains: Details, Intuition, and Big Data. *PLoS Biology*, 13(5):1–6.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, 200(1140):269–294.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The Building Blocks of Interpretability.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement. *NeurIPS*.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):4968–4977.

Schölkopf, B. (2019). Causality for Machine Learning. pages 1–20.

Storrs, K. R. and Kriegeskorte, N. (2020). Deep Learning for Cognitive Neuroscience. In Poeppel, D., Mangun, G. R., and Gazzaniga, M. S., editors, *The Cognitive Neurosciences*. MIT Press, 6th edition.

Xu, Y. (2020). Limited correspondence in visual representation between the human brain and convolutional neural networks.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(3770).