Investigating Auditory Concepts in Deep Neural Networks

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics
by Research

by

Pratyaksh Gautam
2020114002
pratyaksh.g@research.iiit.ac.in

Advised by Dr. Vinoo Alluri and Dr. Makarand Tapaswi

International Institute of Information Technology Hyderabad
500 032, India

June 2026

© Copyright Pratyaksh Gautam, 2026

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis titled ***Investigating Auditory Concepts in Deep Neural Networks*** by *Pratyaksh Gautam* has been carried out under our supervision and is not submitted elsewhere for a degree.

Date	Dr. Vinoo Alluri	Dr. Makarand Tapaswi

To Mom, Dad, and Piya di.

Acknowledgements

This thesis has been a challenging undertaking, one in which I have received the support of many people in my life, more than I can fully acknowledge in this space. I'm grateful to my advisors, Dr. Vinoo Alluri and Dr. Makarand Tapaswi, whose careful consideration and guidance was instrumental in bringing the thesis to its current shape. They have pushed me to think critically and to persevere through challenges, which has led this journey of research to be an invaluable life experience for me. In addition, I would also like to thank all of my lecturers, who have instilled even more curiosity in me, and the college staff whose quiet contributions allow us a college lifestyle we sometimes take for granted.

I have immense gratitude for all of my friends in college; their presence made it so that I truly felt at home in a distant city, and they are the reason I carry a heart full of fond memories with me about a special time in my life wherever I go. My friends Abhinav, Shashwat and Lakshmanan have been true inspirations to me, people I can look up to while also never having to put my guard up around. I have shared incredible experiences with Aneesh, Pranjal, and Manav, who have become like brothers to me. People in my life like Priyanshul and Anirudh have opened my eyes in more ways than one, and been pivotal to my journey not just as a student, but as a human being. I'm grateful to my seniors - Aakash, Raghav and Aditya - who have made me feel like family, as well as my juniors - Nanda, Vedansh and Hardik - to whom I hope I have been able to extend the same feeling. I'm grateful for the vibrant club culture, and would specially like to thank all the members of the Music Club, Theory Group and OSDG.

I've been grateful to have the support of many of my other friends as well - Aaghaz, Sneha, Amrtanshu, Riti, Ananya, Priyanshi, Shubh and Shiksha - who have been there through thick and thin. Finally, I feel blessed to have a family whose support always allowed me to feel reassured, and keep my confidence in my abilities throughout this arduous journey.

Abstract

Deep neural networks have achieved strong performance across a wide range of auditory tasks, from speech recognition to music classification. Despite this success, their internal decision-making processes remain poorly understood, limiting both scientific insight and practical trust. In real-world settings, models that rely on opaque or spurious cues can behave unpredictably, making interpretability a central challenge in modern deep learning. This thesis addresses this gap by asking a central question: do auditory deep neural networks organize sound in a hierarchical manner similar to human auditory perception?

To test this hypothesis, we conduct a large-scale probing analysis of several widely used audio architectures, including three CNN models (VGGish, CLAP, MobileNetV3), and an Audio Spectrogram Transformer (AST). Using linear probes across six tasks of increasing abstraction, we examine how different forms of auditory information are distributed across network depth. Across all convolutional architectures, we observe a clear and consistent hierarchy. Low-level tasks such as note name classification peak in early layers, while higher-level semantic tasks, including genre classification and speaker count estimation, depend on deeper representations. Although this hierarchy is less sharply delineated in the AST due to its global attention mechanism, the same overall progression remains evident. This suggests that hierarchical organization is a robust property of effective auditory models, even when architectural constraints differ.

Beyond identifying this structure, we also show complex auditory concepts emerge within a deep neural network. Focusing on musical audio, we show that an AST trained solely for genre classification develops explicit representations of musical instruments in intermediate layers. Steering vector interventions further show that these instrument representations actively influence genre predictions, demonstrating that they are causally involved in the model's decisions. These findings collectively establish a principled framework for analyzing, comparing, and causally testing internal representations in auditory deep neural networks, while also grounding modern audio models in theories of hierarchical auditory perception.

Contents

Chapter Page

1 Introduction 1
1.1 Deep learning and the need for interpretability 1
1.2 A representational hierarchy for vision and audio 4
1.2.1 Vision 4
1.2.2 Audio 7
1.3 Thesis Objectives 11
1.3.1 Hypothesis 12
1.3.2 Key Contributions 12
1.4 Thesis Organization 12

2 Probing deep neural networks for auditory task performance 13
2.1 Prior work in interpretability of deep neural networks for audio 13
2.2 Probing intermediate layers in audio CNNs 14
2.2.1 Tasks and Datasets 14
2.2.2 Experimental setup 15
2.2.3 Results and discussion 17
2.2.4 Limitations 20
2.3 Probing intermediate layers in AST 21
2.3.1 Audio Spectrogram Transformer (AST) 21
2.3.2 Experimental setup 21
2.3.3 Results and discussion 22
2.3.4 Limitations 22

3 Instruments as emergent and causally relevant concepts for genre classification 25
3.1 Musical genre 25
3.2 Training and probing a genre classifier 27
3.2.1 Training procedure 27
3.2.2 Evaluating genre classification 28
3.2.3 Probing for instrument concepts 28
3.2.4 Results and Discussion 30
3.3 Instruments as steering vectors 30
3.3.1 Experimental setup 31
3.3.2 Results and discussion 34
3.3.3 Limitations 37

4 Conclusion: : : :

List of Figures

Figure Page

1.1 An image classi cation model might misclassify an image of a husky (1.1a) as that of a wolf, solely on the basis of the presence of snow in the image (1.1b). This is an example of a spurious correlation learnt by the model from the training data. Figure adapted from Ribeiro et al. [64]. 2

1.2 On the left, we have the chest X-ray scans of two patients, one who tested positive for COVID-19 (top), and one who tested negative (bottom). The orientation marker `R' is only present to indicate the patient's right side on the X-ray. On the right, we see the corresponding saliency maps for a classi er trained to predict the COVID test result from the X-ray, with the darkest red spots being the most important. We clearly see that in both cases, the orientation marker `R', (as highlighted with a black arrow on the saliency maps) is in uential in predicting the test result. Figures adapted from DeGrave et al. [27]. 3

1.3 Visualization of features learned by a convolutional neural network. Early layers detect simple edges (top left), while deeper layers assemble these into complex geometries and object parts (middle, bottom). Note how these features naturally form a hierarchy of composition. Figure from Zeiler and Fergus [79]. 5

1.4 Individual neurons from a convolutional neural network trained for scene classi cation (i.e. identifying whether an image is of a bedroom, a movie theater, a zoo, etc.) often act as object detectors. These neurons learn to re with high precision for instances of the same object, contributing to the understanding of a larger scene. This phenomenon is observed even with no explicit training for object labels. Figure from Zhou et al. [81]. 6

1.5 In the human brain, the processing of visual information is supported by two pathways: the dorsal or the `where' pathway, and the ventral or `what' pathway. The processing of auditory information is analogously supported by its own distinct dorsal and ventral pathways. Though these pathways are not completely isolated from each other, each one specializes for its function. 7

1.6 The Recognition by Components (RBC) framework [12] suggests that visual objects are recognized by parsing them into component parts, or “geons”. This includes objects like a cylinder, and a toroidal segment (or a curved cylinder) as shown on the left, combining to form a mug as shown on the right. 8

1.7 Conceptual illustration of temporal integration and hierarchical auditory processing for music. Low-level acoustic features (pitch, timbre) integrate over time to form interme- diate objects (melody, instruments), which further combine to form high-level semantic concepts (musical style or genre). We only enlist concepts here directly relevant to our discussion; this is not an exhaustive list. 9

1.8 An image showing the tonotopic map of the human cochlea. The different regions of the cochlea respond to speci c frequency regions, as shown with different colors in the image. Figure from Li et al. [51]. 10
1.9 Sequences of notes having the same timbre are perceived as belonging to the same auditory stream. Differing timbres are perceived as separating the auditory streams (shown here in two different colors), despite similarities in pitch. 11

2.1 Accuracy using intermediate representations extracted from convolutional layers of increasing depth. Gray dashed line represents chance accuracy. Tasks are ordered in increasing abstraction for music (left) and speech (right). We observe that low-level tasks (row 1) are likely to perform well at shallow layers while semantic high-level tasks (row 3) perform well at deeper layers. 18
2.2 Effect of changing the value ofkfor CLAP on the Medley-solos-DB dataset. In general, changingk does not have a major effect on the overall layer-wise trend, across all models and datasets. 19
2.3 Effect of truncating all input audio les to1 s on CLAP. The high-level tasks display an earlier saturation, but the general trends remain. Similar results are observed for VGGish and MobileNetV3 as well. The six tasks are presented in the same order as Fig. 2.1. 20
2.4 Classi cation accuracy for representations extracted at different layers of AST. We see a less pronounced effect here. The note name classi cation task peaks at the earliest layers with performance dropping over time, and the low- and mid-level speech tasks also show the best classi cation accuracy for representations from the middle layers. Notably, the performance on speaker count estimation is comparable regardless of what layer the representation is extracted from. Results shown fork-Nearest Neighbor classi cation with k which gives maximum peak classi cation accuracy. 23

3.1 Normalized confusion matrix for our AST model on GTZANLike (test set). Model performance is above chance across all genres, though slightly weaker forblues and jazz. All misclassi cations> 10% are highlighted for brevity and often occur between closely related genres such asrock andmetal. 29
3.2 Macro average F1-score on MedleySolosDBLike (test set) for multi-class instrument probes. Probes are trained on intermediate layer representations with balanced sampling. Performance peaks at the middle layers. 30
3.3 Macro average F1-score on MedleySolosDBLike (test set) for binary instrument probes. Probes are trained on intermediate layer representations with 100 random subsamples (250 positives, 250 negatives); error bands show95% con dence intervals. Performance rises from early layers and peaks at intermediate layers. Results are shown forguitar, violin, andpiano; similar results are observed for other instrument probes. 30
3.4 Effect on the probability of being classi ed as a particular genre when intervened on with the corresponding instrument. We intervene on instances from GTZANLike (test) labeled with the genre (blue) as well as instances labeled with any of the remaining genres (orange). In all cases, we see that adding the instrument vector has stronger effects on instances which are not from the corresponding genre, and subtracting the instrument vector has stronger effects on instances which are from the corresponding genre. We use the one-to-one steering strategy for these instruments. 32

3.5 Effect on the probability of being classified as a particular genre when intervened on with piano, an instrument not strongly associated with any one genre. We see classification probabilities for classical (top), rock-metal (middle), and jazz (bottom). We intervene on instances from GTZANLike (test) labeled with the genre (blue) as well as instances labeled with any of the remaining genres (orange). In all cases, no strong trend is observed in terms of the changes in the probability on intervening. We use the one-to-one steering strategy for these instruments. 33
3.6 Effect on the probability of being classified as a particular genre when intervened on with the corresponding instrument. We intervene on instances from GTZAN labeled with the genre (blue) as well as instances labeled with any of the remaining genres (orange). In all cases, we see that adding the instrument vector has stronger effects on instances which are not from the corresponding genre, and subtracting the instrument vector has stronger effects on instances which are from the corresponding genre. We use the one-to-one steering strategy for these instruments. 35
3.7 Effect on the probability of being classified as a particular genre when intervened on with piano, an instrument not strongly associated with any one genre. We see classification probabilities for classical (top), rock-metal (middle), and jazz (bottom). We intervene on instances from GTZAN labeled with the genre (blue) as well as instances labeled with any of the remaining genres (orange). In all cases, no strong trend is observed in terms of the changes in the probability on intervening. We use the one-to-one steering strategy for these instruments. 36
3.8 The effect of intervening with the guitar vector on rock-metal using the one-to-all strategy. We show results for guitar vectors extracted from different layers. Interventions using instrument vectors from early layers have a negligible effect. In contrast, adding the guitar vector from later layers increases the probability of classification as rock-metal, while subtracting it decreases this probability. This behavior mirrors the results observed with the one-to-one strategy. Similar trends are observed for other instrument–genre pairs. 38
3.9 Cosine similarities between the vectors for guitar and rock-metal at different layers. Note how the similarity is greatest at the earlier layers, but we see the effect of the intervention is strongest at later layers (Fig. 3.8). 39

List of Tables

Table Page

2.1 Speci c layers at which the intermediate representations are extracted for each model and the corresponding layer name used in Fig. 2.1. The layers are selected to be equally spread from the model input up to the rst fully-connected layer. For CLAP, the pre x audio _encoder.base is omitted for brevity. 16

2.2 Datasets used in our work from musEc)(and human speech (domains. #C is the number of classes and Dur. (s) is the typical audio duration in seconds. *The dataset is modi ed slightly to be class-balanced when possible. 19

3.1 Class counts for genres in GTZANLike and instruments in MedleySolosDBLike. . . . 27

3.2 Data augmentation con guration. size/rate/steps are drawn from a uniform distribution, and then the augmentation is applied with probabilityp 28

Chapter 1

Introduction

1.1 Deep learning and the need for interpretability

With the advent of deep learning [50], deep neural networks (DNNs) have shown a remarkable ability to perform complex tasks across a wide range of domains. These successes span language modeling [17, 29, 76], image classi cation [30, 48], and automatic speech transcription [9]. In the domain of non-speech audio speci cally, DNNs have been deployed with great success to classify audio instances, identifying sounds, instruments, and musical genres [8, 31, 44]. Furthermore, multi-modal large language models (MLLMs) capable of processing both audio and text [24, 28, 37] have enabled capabilities such as captioning and question-answering for audio inputs. Beyond analysis, DNNs are now used to clone voices [5, 19] and generate music directly as raw waveforms [2, 25, 36, 53]. Collectively, these advances highlight the growing breadth and sophistication of deep learning in processing auditory information.

Despite their empirical success, the interpretability of DNNs remains a signi cant challenge; the internal mechanisms that give rise to model behavior are not yet fully understood. Because these models build understanding through many layers of non-linear transformations on high-dimensional data, it remains dif cult to discern how their internal representations give rise to the behaviors we observe.

Studying how models arrive at their conclusions is often just as revealing as the conclusions themselves. A model might provide the correct output for a given task but rely on awed reasoning hidden beneath the surface. Consider an image classi er that occasionally erroneously labels huskies as wolves, as shown in Fig. 1.1. At rst glance, this seems like an understandable error, given the visual similarity between the two animals. However, a closer analysis shows that the misclassi cation is not driven by confusion in their physical features at all [64]. Instead, the model has learned to associate snow in the background with the label `wolf'. This is a spurious correlation, a coincidental pattern in the training set where wolves tended to appear in snowy environments. Because this cue was statistically reliable during training, the model adopted it as a shortcut, even though it bears no meaningful relation to the actual task of distinguishing between the animals.

While this example is relatively benign, the stakes can be markedly higher for other applications. Consider a medical image classi er that correctly diagnoses diseases, not by detecting patient features,

(a) An image of a husky. An image classi cation model mistakenly classi es it as an image of a wolf.

(b) The parts of the image found to be most important for the model's classi cation.

Figure 1.1: An image classi cation model might misclassify an image of a husky (1.1a) as that of a wolf, solely on the basis of the presence of snow in the image (1.1b). This is an example of a spurious correlation learnt by the model from the training data. Figure adapted from Ribeiro et al. [64].

but by identifying orientation markers like `left' or `right' on the X-ray scans (see Fig. 1.2). This is yet another spurious correlation, a `shortcut' learned from the training data [27]. Such success is deceptive; these shortcuts can fail catastrophically when the model is deployed in a new hospital or imaging setup where those speci c markers are absent.

Additionally, achieving greater mechanistic insight into DNNs does more than prevent failure modes; it can advance our understanding of the domain itself. For instance, examining the features relied upon by protein-folding models has helped biologists identify previously overlooked structural patterns in real proteins [67]. The researchers originally aimed to apply interpretability tools to better understand the internal representations of protein language models, yet this analysis unexpectedly uncovered coherent biological features that highlighted missing or incomplete annotations in existing protein databases. By applying supervised pruning to retain only the speci c subspaces (subsets of features) that best match human similarity ratings for a given category, Bavaresco et al. [10] were able to modify a deep image classi cation model to signi cantly improve its performance at identifying AI-generated images that aligned with human aesthetic and semantic preferences. The authors achieved this without any ne-tuning, demonstrating promise in the role of model interpretability in improving both alignment and performance of deep learning models.

Thus, the ability to interpret the models we use is crucial both for trust and for gaining insight into the tasks we use such models for. We see a clear need to understand the internal mechanisms of deep neural

(a) A chest X-ray scan of a COVID-19 positive patient.(b) Saliency map for the X-ray shown on the left (a).

(c) A chest X-ray scan of a COVID-19 negative patient(d) Saliency map for the X-ray shown on the left (c).

Figure 1.2: On the left, we have the chest X-ray scans of two patients, one who tested positive for COVID-19 (top), and one who tested negative (bottom). The orientation marker `R' is only present to indicate the patient's right side on the X-ray. On the right, we see the corresponding saliency maps for a classi er trained to predict the COVID test result from the X-ray, with the darkest red spots being the most important. We clearly see that in both cases, the orientation marker `R', (as highlighted with a black arrow on the saliency maps) is in uential in predicting the test result. Figures adapted from DeGrave et al. [27].

networks better. Arising from this need, there has been increased interest in work on interpretability methods to explain the behavior of DNNs.

1.2 A representational hierarchy for vision and audio

1.2.1 Vision

Interpretability studies in vision have demonstrated that convolutional neural networks (CNNs) develop structured, human-interpretable representations that mirror a compositional hierarchy. Research into the internal activations of these networks reveals that early layers form simple feature detectors (see Fig. 1.3), sensitive to low-level primitives such as edges and curves [79]. As data progresses through the network, higher layers encode increasingly complex concepts, such as shapes and textures [58].

Strikingly, this emergence of structure often occurs without explicit supervision for those specific concepts. For example, CNNs trained strictly for scene classification, identifying whether an image depicts a bedroom, a hotel lobby, or a theater, have been found to spontaneously develop `object detector' neurons [81]. A neuron might learn to activate strongly only in the presence of a bed, despite the model never being told what a `bed' is. The model implicitly learns that a `bedroom' is a scene composed of objects like beds and lamps, as shown in Fig. 1.4.

Quite notably, this computational hierarchy mirrors the hierarchical processing observed in human vision. Once the information of a visual scene travels from the retina of the observer and through the optic nerve, it eventually reaches the primary visual cortex. From there, processing splits into dorsal and ventral pathways (see Fig. 1.5), dual pathways that are thought to specialize to information about where something is (localization) and what it actually is (recognition and semantics) [57]. Neurons in early areas respond to features relevant to object identification (color, shape), with receptive fields (i.e. the region of visual space a neuron responds to) and feature selectivity (the specific visual attributes a neuron is tuned to) progressively increasing along the pathway [34].

A strikingly similar compositional logic is visible in deep networks. As shown in Fig. 1.3, the earliest CNN layers respond to simple edges and color contrasts; intermediate layers assemble these into curves, textures, and recurring part-like motifs reminiscent of geons; while the deepest layers activate for whole objects such as faces and animals. Prominent theoretical frameworks in human vision are also consistent with this compositional view. Recognition by Components (RBC) [12] proposes that we perceive objects by parsing them into volumetric primitives called “geons”. A mug, for instance, is recognized as a cylinder joined to a toroidal handle (see Fig. 1.6).

Human visual processing also involves top-down influences such as feedback connections, attentional modulation, and prior knowledge, but these do not diminish the substantial evidence for hierarchical processing and the important role of compositionality in vision.

Thus, prior work suggests that both deep vision models, much like the human visual system, learn compositional hierarchies: they assemble low-level inputs into intermediate concepts, which in turn inform high-level semantic decisions.