博士学位论文

拓扑方法在神经网络架构中的应用

METHODS OF ALGEBRAIC TOPOLOGY IN THE STUDY OF NEURAL NETWORK ARCHITECTURE

研　　究　　生：于智旺

指　导　教　师：方复全讲席教授

朱一飞助理教授

南方科技大学

二〇二五年六月

国内图书分类号：O189.2　　学校代码：14325
国际图书分类号：515.1　　密级：公开

理学博士学位论文

拓扑方法在神经网络架构中的应用

学 位 申 请 人：于智旺
指 导 教 师：方复全讲席教授
朱一飞助理教授
学 科 名 称：数学
答 辩 日 期：2025 年 5 月
培 养 单 位：数学系
学位授予单位：南方科技大学

METHODS OF ALGEBRAIC TOPOLOGY IN THE STUDY OF NEURAL NETWORK ARCHITECTURE

A dissertation submitted to
Southern University of Science and Technology
in partial fulfillment of the requirement
for the degree of
Doctor of Science
in
Mathematics

by
Yu Zhiwang

Supervisor: Chair Prof. Fang Fuquan
Assistant Prof. Zhu Yifei

May, 2025

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

无（全匿名评阅）

答辩委员会名单

主席	严质彬	教授	哈尔滨工业大学（深圳）
委员	方复全	讲席教授	南方科技大学
	邬龙挺	助理教授	南方科技大学
	于天舒	助理教授	香港中文大学（深圳）
	朱一飞	助理教授	南方科技大学
秘书	刘鲁一	博士后	南方科技大学

DECLARATION OF ORIGINALITY AND AUTHORIZATION OF THESIS, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Declaration of Originality of Thesis

I hereby declare that this thesis is my own original work under the guidance of my supervisor. It does not contain any research results that others have published or written. All sources I quoted in the thesis are indicated in references or have been indicated or acknowledged. I shall bear the legal liabilities of the above statement.

Signature: [redacted] Date: 2025.06.02

Declaration of Authorization of Thesis

I fully understand the regulations regarding the collection, retention, and use of the thesis of the Southern University of Science and Technology.

1. Submit the electronic version of the thesis as required by the University.

2. The University has the right to retain and send the electronic version to other institutions that allow the thesis to be read by the public.

3. The University may save all or part of the thesis in certain databases for retrieval and may save it with digital, cloud storage, or other methods for teaching and scientific research. I agree that the full text of the thesis can be viewed online or downloaded within the campus network.

(1) I agree that once submitted, the thesis can be retrieved online and the first 16 pages can be viewed within the campus network.

(2) I agree that ☑upon submission/ ☐_____ months after submission, the full text of the thesis can be viewed and downloaded by the public.

4. This authorization applies to the decrypted confidential thesis.

Signature of Author: [redacted] Date: 2025.06.02

Signature of Supervisor: [redacted] Date: 2025.06.02

摘 要

拓扑数据分析作为一项新兴的数学工具，近年来在深度学习领域展现了巨大的潜力，尤其在神经网络的设计与优化方面表现出色。在本论文中，我们用提取拓扑信息的卷积核来构造卷积神经网络，其在多个语音数据集均表现出高准确率。

首先，研究详细分析了以 Klein 瓶特征初始化的卷积核在 CNN 中的独特表现，这种设计能够高效捕获数据中的拓扑信息，显著提升网络的表征能力与泛化性能。在复现基础实验的同时，本研究还优化了实验条件，为探讨这些方法的广泛适用性奠定了基础。同时，本研究结合 Gabrielsson 与 Carlsson 提出的理论框架，尝试提取了卷积权重分布的拓扑特征。

进一步，也是本文的核心内容是，对卷积核进行正交群作用，然后通过获得的轨道空间与矩阵空间之间的关系，获得了近似纤维丛的结构，从而将矩阵空间分解成底空间与正交群两部分，为生成滤波器提供了理论基础。

之后，我们生成了正交特征层 (OF)，其在音素识别任务中的性能显著优于传统方法，尤其在低噪声环境下展现出极高的生命力。此外，这些卷积核在单词分类和图像分类任务中同样表现出良好的适应性，展示了拓扑方法跨领域应用的潜力。

最后，研究初步将构造卷积核的理论推广至黎曼几何框架下，提出基于几何正则化的优化策略，为后续研究提供理论基础与实验支撑。本研究通过复现与拓展现有成果，进一步揭示了拓扑工具在神经网络优化中的可能性，为跨数学与深度学习的交叉领域开辟了新的研究方向。

关键词：拓扑数据分析；卷积神经网络；语音识别；群作用

ABSTRACT

Topological data analysis, as a burgeoning mathematical tool, has demonstrated immense potential in the field of deep learning in recent years, particularly excelling in the design and optimization of neural networks. In this dissertation, convolutional kernels that extract topological information are employed to construct convolutional neural networks, achieving high accuracy across multiple speech datasets.

The study begins with a detailed analysis of the unique performance of convolutional kernels initialized using Klein bottle features within CNNs. This design effectively captures the topological structures in data, significantly enhancing the representational capability and generalization performance of the network. Alongside reproducing foundational experiments, this work optimizes experimental conditions, laying the groundwork for exploring the broader applicability of these methods. Additionally, the research incorporates the theoretical framework proposed by Gabrielsson and Carlsson, attempting to extract topological features from the distribution of convolutional weights.

Furthermore, as the core contribution, the study explores orthogonal group actions on convolutional kernels. By uncovering the relationship between orbit space and matrix space, a structure approximating fiber bundles is established, decomposing matrix space into a base space and orthogonal group components, thus providing a theoretical foundation for filter generation.

Subsequently, the Orthogonal Feature Layer (OF) is developed, which demonstrates significantly superior performance in phoneme recognition tasks compared to traditional methods, particularly excelling in low-noise environments. Additionally, these kernels exhibit robust adaptability in word classification and image classification tasks, highlighting the cross-domain application potential of topological methods.

Finally, the study preliminarily extends the kernel construction theory into the Riemannian geometry framework, proposing geometry-regularized optimization strategies. This provides a theoretical and experimental foundation for subsequent research. By reproducing and expanding upon existing results, the work further unveils the potential of topological tools in neural network optimization, paving the way for new research directions in the intersection of mathematics and deep learning.

Keywords: Topological data analysis; Convolutional neural network; Speech recognition; Group action

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

Topological Data Analysis provides a foundational exploration of TDA's core concepts, such as simplicial complexes and persistent homology, making it an essential resource for understanding multi-scale topological features in data analysis. (Bass et al.[4])

1.1 Research Backgrounds and Motivations

1.1.1 Historical Evolution of Topological Data Analysis

From Algebraic Topology to Data Science

The mathematical foundations of Topological Data Analysis (TDA) trace back to classical algebraic topology, where concepts such as homology groups and Betti numbers were developed to characterize the connectivity of abstract spaces. Early attempts to apply these ideas to data analysis were largely theoretical, focusing on combinatorial representations of point clouds through simplicial complexes. However, the lack of scalable algorithms limited practical adoption until the early 2000s, when computational geometry intersected with topology to address real-world data challenges.

The Birth of Persistent Homology

A pivotal advancement occurred with the formalization of persistent homology by Edelsbrunner et al.[19], which introduced a multi-scale framework for analyzing topological features. Persistent homology uses a filtration process to incrementally construct nested simplicial complexes, enabling the systematic tracking of topological structures, such as clusters and cycles, across multiple spatial resolutions. Zomorodian and Carlsson[100] further refined this framework, demonstrating its stability under noise through algebraic representations. The publication of seminal review by Carlsson[7] in 2009 marked the transition of TDA from a niche mathematical tool to a mainstream data science methodology.

Algorithmic and Theoretical Maturation

Subsequent research focused on enhancing computational efficiency and theoretical robustness. Chazal et al.[11] established stability theorems, proving that small perturba-

tions in input data yield bounded changes in persistence diagrams, a critical property for noisy real-world datasets. Meanwhile, Oudot[56] developed efficient algorithms for computing persistence modules, enabling applications to large-scale datasets. These advances laid the groundwork for modern TDA software libraries, such as Gudhi and Ripser, which handle millions of data points in domains ranging from genomics to materials science.

The theoretical maturation of TDA also bridged gaps between pure mathematics and applied statistics. Concepts like the Euler characteristic curve and Mapper algorithm emerged as interpretable tools for high-dimensional data visualization, further expanding the utility of TDA beyond homology-based methods.

1.1.2 Applications of Topological Data Analysis

Materials Science and Chemistry

In materials science, TDA has revolutionized the characterization of porous structures. For instance, Kramar et al.[41] used persistence diagrams to quantify the connectivity of nanopore networks in catalytic materials, while Nakamura et al.[54] applied Morse theory to classify crystal defects. Recent work by Pike et al.[60] demonstrated how TDA-guided simulations predict polymer phase separation dynamics, offering insights for designing advanced composites. In drug design, Liu et al.[52] developed a hypergraph-based persistent cohomology model to capture complex protein-ligand interactions through atomic-level topological representations, showing superior performance in binding affinity prediction compared to traditional geometric and chemical descriptors. Building on this, Liu et al.[51] introduced persistent spectral hypergraphs to enhance machine learning predictions of molecular binding thermodynamics. Knot theory-based innovations by Shen et al.[74] further expanded the utility of TDA through multiscale Gauss link integrals for analyzing molecular entanglement and polymer chain dynamics.

Biomedical Imaging and Genomics

Medical applications leverage the ability of TDA to detect subtle structural anomalies. Qaiser et al.[61] employed persistent homology to distinguish malignant tumors from benign tissues in histopathology images, achieving higher specificity than traditional texture analysis. Similarly, Dindin et al.[16] utilized TDA for pathological pattern recognition in medical imaging. In genomics, Carrière and Rabadán[10] analyzed RNA sequencing data through topological descriptors to identify nonlinear gene interactions linked to can-

cer progression, extending earlier work by Yao et al.[95] on modeling gene expression dynamics via topological networks.

Sensor Networks and Transportation

De Silva and Ghrist[15] pioneered the use of Vietoris–Rips complexes to monitor connectivity in wireless sensor networks, enabling fault detection in harsh environments. Li et al.[49] extended these ideas to urban transportation, using TDA to model traffic flow patterns and optimize route planning in real time. TDA also addresses multivariate time series challenges: Seversky et al.[73] developed anomaly detection methods based on topological features, while Umeda[91] analyzed dynamical systems using evolving topological descriptors.

3D Shape and Image Analysis

In 3D shape analysis, Skraba et al.[77] introduced topological signatures for shape matching, Turner et al.[90] proposed persistent homology frameworks to quantify shape complexity, and Tralie and Perea[89] combined geometric-topological features for robust shape classification. The recent Zhuo et al.[102] advanced this field with PHTNet, a neural network architecture integrating multi-perspective topological features for enhanced 3D object recognition. For 2D image analysis, Rieck et al.[64] leveraged TDA to generate topological descriptors that enhance texture and structural feature representation. Finally, Tinarrage[87] contributed to the computation of persistent Stiefel–Whitney classes of line bundles, expanding the applicability of topological methods by providing novel tools that intersect algebraic topology with computational frameworks for 2D analysis.

Emerging and Niche Applications

Beyond these domains, TDA has found niche applications in art restoration,analyzing brushstroke patterns in paintings, and climate science, where persistent homology detects recurring atmospheric patterns in spatiotemporal data. Its versatility continues to inspire interdisciplinary innovations across mathematical physics, materials informatics, and computational biology.

1.1.3 The Rise of Convolutional Neural Networks

Biological Inspiration and Early Models

The development of CNNs was inspired by Hubel and Wiesel's[36] discovery of hierarchical visual processing in the mammalian cortex. Fukushima's[22] neocognitron, the first CNN-like architecture, mimicked this hierarchy with alternating layers of simple and complex cells. However, limited computational power and training data hindered progress until the 1990s, when LeCun et al.[44] introduced LeNet-5 for handwritten digit recognition, leveraging backpropagation (Rumelhart and McClelland[67]) for end-to-end training.

The Deep Learning Revolution

The 2012 ImageNet competition marked a turning point. Krizhevsky et al.[43] demonstrated that a deep CNN (AlexNet) could outperform traditional computer vision methods by a significant margin, catalyzing a paradigm shift. Subsequent architectures like ResNet (He et al.[32]) and Inception addressed vanishing gradients and overfitting, while applications expanded to object detection (Faster R-CNN (Sultana et al.[81]) and semantic segmentation (U-Net (Zaitoun and Aqel[96])).

Modern CNNs exploit spatial locality through convolutional kernels, pooling layers for translation invariance, and skip connections for gradient flow. These innovations enabled state-of-the-art performance in tasks ranging from medical image segmentation to autonomous driving, as evidenced by their adoption in FDA-approved diagnostic tools (Qaiser et al.[61]) and Tesla's Autopilot system.

1.1.4 Challenges and Limitations of CNNs

Interpretability and Robustness

The breakthrough of CNNs in image classification, catalyzed by the ILSVRC challenge (Russakovsky et al.[68]) and architectural innovations like those in (Sultana et al.[82]), has been tempered by critical limitations. Despite their hierarchical feature abstraction capabilities (Rawat and Wang[63]), CNNs remain vulnerable to adversarial attacks-subtle input perturbations that induce high-confidence misclassifications (Zheng et al.[98]). This fragility stems from their texture-biased feature learning, as demonstrated through stylized ImageNet experiments (Geirhos et al.[25]), where models failed to recognize shape-preserving texture-altered objects.

Data and Computational Demands

While CNNs revolutionized video analysis through benchmarks like UCF101 (Soomro et al.[79]) and KTH (Schuldt et al.[72]), their reliance on massive labeled datasets and GPU clusters creates barriers in resource-constrained domains. Medical imaging studies (Guo et al.[29]) reveal catastrophic performance drops when training data falls below critical thresholds. Even foundational spatiotemporal modeling approaches (Gorelick et al.[27]) require unsustainable computational resources, raising environmental concerns about the carbon footprint of large-scale training.

1.1.5 Synergizing TDA and CNNs: A Topological Deep Learning Paradigm

Topological Regularization and Interpretability

Integrating TDA with CNNs addresses both interpretability and robustness. By using persistence diagrams as regularizers (Adams et al.[2]), networks learn features aligned with topologically meaningful structures (e.g., loops in texture analysis), overcoming the texture bias observed in pure CNN architectures (Geirhos et al.[25]). This approach enhances model introspection while maintaining the hierarchical abstraction strengths of CNNs (Rawat and Wang[63]).

Hybrid Architectures for Video and Speech Analysis

Building on CNN-based video analysis foundations (Soomro et al.; Schuldt et al.; Gorelick et al.[27,72,79]), TDA-enhanced frameworks now track topological dynamics across frames. The sliding window persistence method (Khasawneh and Munch[39]) improves action recognition by encoding temporal coherence in sports analytics videos. Similarly, speech processing systems combine Mel-frequency cepstral coefficients with persistent homology (Brown and Knudson[6]), preserving harmonic structures in noisy environments while leveraging CNN spectral analysis capabilities. Liu et al.[50] explores the innovative integration of topological persistence with convolutional neural networks (CNNs), demonstrating its effectiveness in feature extraction and classification tasks for music audio signals. Robinson[65] highlights how persistent homology can be utilized in signal analysis to uncover structural patterns, bridging topology and signal dynamics in practical applications.

Future Directions: Topological Attention and Generative Models

Emerging architectures integrate topological attention mechanisms with CNN feature maps, guiding focus toward structurally critical regions identified through persistence-based saliency. In generative modeling, GANs trained with topological loss functions (Smith et al.; Lee et al.[47,78]) produce synthetic data with geometrically realistic properties, potentially overcoming data scarcity limitations that plague conventional CNNs (Guo et al.[29]).

1.1.6 Research Significance

This study advances both theoretical foundations and methodological frameworks in neural network analysis, with dual contributions articulated as follows.

Bridging the Mathematical Gap in Neural Network Interpretability

The rapid evolution of deep learning has precipitated a critical disconnect between empirical success and theoretical understanding. While convolutional neural networks (CNNs) have attained human-level proficiency in speech recognition tasks[3], the intricacies of their decision-making processes continue to elude comprehensive understanding. This ”black box” dilemma not only hinders model optimization but also forces architectural improvements to rely on heuristic trial-and-error approaches. Our work addresses this gap by establishing a mathematical scaffolding through TDA. Specifically, we:

- Develop a hybrid framework integrating persistent homology (Edelsbrunner and Harer[17]) and Morse theory to characterize convolutional operations in speech CNNs
- Quantify topological invariants (Betti numbers, persistence barcodes) of convolutional kernels in Mel-spectrogram space
- Reveal the geometric preservation of phonemic features through $\mathbb{Z}/2\mathbb{Z}$-homology analysis of pitch contour manifolds

This mathematical formalization provides unprecedented insights into how CNNs hierarchically extract speech patterns, moving beyond traditional statistical learning paradigms.

Cross-Modal Extension of TDA Methodology

While Carlsson et al.[8] pioneered TDA applications in visual CNNs, significant challenges emerge when adapting this approach to speech processing:

$$\mathcal{C}_{\text{speech}} = \underbrace{\text{Time-frequency entanglement}}_{\text{Non-Euclidean structure}} + \underbrace{\text{Non-stationary articulation dynamics}}_{\text{Topological instability}}. \quad (1\text{-}1)$$

Our key innovations include

• Designing temporal persistence modules for analyzing 1D convolutional filters in raw waveform processing

• Developing spiral barcode descriptors to capture formant transitions in vowel spaces

• Establishing a stability theorem for speech-specific topological signatures under ϵ-perturbations

This methodological transfer enables direct comparison of feature learning mechanisms across auditory and visual modalities, creating new opportunities for cross-domain neural architecture design.

1.2 Statement of Results

The main focus of this dissertation is the application of convolutional kernels, constructed using the newly defined Orthogonal Feature Layer (OF), to phoneme recognition. Furthermore, the approach is generalized to word recognition and image recognition, demonstrating its versatility and adaptability across multiple domains.

First, in Chapter 4, inspired by Love et al.[53], we transform the speech data to spectrogram by Short-Time Fourier Transform (STFT).

Second, in Chapter 5, we consider the space $M_{3\times3}(\mathbb{R})$, which is the most common space of convolutional kernels, as $\{[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]\}$. Without loss of generality, define the subspace $M = \{[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3] | \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 = 0\}$ with the Frobenius form of M is equal to one. Next, we define a group action on M by

$$\theta(\boldsymbol{Q}, \boldsymbol{m}) = \boldsymbol{Q}\boldsymbol{m} \text{ for } \boldsymbol{Q} \in \mathrm{SO}(3) \text{ and } \boldsymbol{m} \in M.$$

Thus, denote the quotient map from M to $M/\mathrm{SO}(3)$ by π, then the orbit space $M/\mathrm{SO}(3)$, denoted as B, is homeomorphic to a disk D^2. Moreover, there exist a stratified fiber bundle on B. The fiber is $\mathrm{SO}(3)/(\mathrm{SO}(2) \rtimes \mathbb{Z}_2) \cong \mathbb{R}P^2$ on the case when $\boldsymbol{v}_1 + \boldsymbol{v}_3 = 0$, and $\mathrm{SO}(3)/\mathrm{SO}(2) \cong S^2$ on the case when $\boldsymbol{v}_1$ and $\boldsymbol{v}_3$ are collinear, and $\mathrm{SO}(3)/\mathbb{Z}_2 \cong L(4,1)$ on the case when $\boldsymbol{v}_1$ and $\boldsymbol{v}_3$ are have equal magnitudes, and $\mathrm{SO}(3)$ on the other cases. The above provides a representation of M by special orthogonal group action.

Third, in Chapter 6, we define the Orthogonal Feature Layer (OF) by selecting element in B and $\mathrm{SO}(3)$. Then, we compare neural networks constructed using OF convolutional kernels to traditional neural networks and the networks proposed by Love et al. on phoneme datasets. The results indicate that OF achieves the highest accuracy under low noise conditions. However, in high noise environments, OF's performance declines, with KF (kernel filters) emerging as the superior approach.

Fourth, in Chapter 7, the applicability of OF convolutional kernels is further explored by extending their use to word datasets and image datasets. Results demonstrate consistent generalization properties, showcasing the versatility and robustness of the proposed methodology.

Finally, also in Chapter 7, this research attempts to generalize the convolutional kernel theory within the framework of Riemannian geometry.

1.3 Outline

To integrate the diverse theories, methodologies, and applications discussed throughout this dissertation, the structure is organized into eight chapters, each focusing on distinct aspects of topological deep learning. The detailed organization is as follows:

Chapter 2: Mathematical Tools for Topological Deep Learning This chapter develops the foundational concepts of topological data analysis (TDA) and its mathematical framework. It explores simplicial complexes, persistent homology, graph-based adaptations, and group actions, establishing the theoretical basis for integrating topology into convolutional neural networks.

Chapter 3: Interdisciplinary Tools for Topological Deep Learning This chapter investigates deep learning architectures, particularly CNNs, and their extension to topological convolutional neural networks (TCNNs). Additionally, it addresses linguistic concepts for phoneme analysis, enabling the seamless integration of topology with speech tasks.

Chapter 4: Topological Deep Learning: From Image Data to Speech Data This chapter examines topological properties in image datasets, such as MNIST and CIFAR10, while expanding analyses into spectrogram representations of speech data. It leverages both persistent homology and CNN architectures to evaluate the effectiveness of topological insights.

Chapter 5: The Space of Spectrogram Convolution Kernel This chapter con-

structs the kernel space for spectrogram convolution using mathematical constraints, such as contrast maximization and group actions. A geometric analysis highlights the kernel space's topological structure and its applicability in speech processing tasks.

Chapter 6: New Spectrogram Convolutional Kernel Building upon Chapter 5, this chapter introduces novel kernel designs optimized for phoneme recognition through CNN architectures. Experimental results demonstrate enhanced accuracy and noise robustness across multiple datasets.

Chapter 7: Further Applications and Extensions This chapter begins by examining the experimental results of various models under conditions where phonemes are not filtered. It further evaluates the models on words and images, demonstrating the strong generalization capability of the proposed model. Additionally, the chapter explores the extension of group action theory from Euclidean geometry to Riemannian geometry. Finally, the chapter concludes by addressing the limitations of this study.

Throughout the paper, every claim is supported by a robust set of references. In this dissertation, we treat the weight vector and convolutional kernel interchangeably, without differentiating between the two concepts. The integration of TDA and CNN is presented not only as a promising research direction but also as a practical solution that addresses some of the key challenges in modern data science. For details, all experiments can be seen at (https://github.com/ZhiwangYu/OrthogonalityFeatures).

CHAPTER 2 MATHEMATICAL TOOLS FOR TOPOLOGICAL DEEP LEARNING

This chapter presents a cohesive framework for topological analysis spanning discrete structures and continuous dynamical systems. Beginning with the mathematical bedrock, we first develop the core machinery of topological data analysis through simplicial complexes, combinatorial representations of topological spaces, and their multiscale interrogation via persistent homology. These foundational constructs naturally extend to graph-structured data through the lens of adjacency complexes, where relational connectivity patterns are translated into hierarchical topological signatures, an approach pioneered by Grigoryan et al.[28] in their work on persistent path homology.

Temporal dynamics enter the framework through sliding window embeddings, a technique that transforms time-evolving systems into geometrically structured point clouds. This methodology bridges discrete topology with continuous data streams, enabling the analysis of phenomena ranging from neural activity patterns (as demonstrated by Giusti et al.[26]) to social network evolution (per Sizemore et al.[76]). The progression culminates in the study of group actions on manifolds, where Horak et al.'s[35] insights into spectral invariants inform the analysis of symmetry-driven dynamics in homogeneous spaces.

By interweaving discrete combinatorial topology with geometric flows, this chapter establishes a unified scaffold for multiscale data analysis-from static graphs to temporally evolving networks and beyond. Each conceptual layer (simplicial complexes §1, persistent homology §2, graph adaptations §3, temporal embeddings §4, and geometric dynamics §5) builds dialectically upon its predecessor, creating an analytical continuum that respects both discrete data structures and continuous system behaviors.

2.1 Simplicial Complex

Datasets are often represented as collections of points, commonly referred to as **point clouds**. To understand the intrinsic shape of these data, one approximates the point cloud by constructing families of simplicial complexes. These abstract complexes serve as combinatorial representations of the underlying geometry. Different methods for "filling in"

higher-dimensional simplices in the proximity graph yield different global representations. These three methodologically fundamental frameworks emerge through rigorous analytical derivation:

Definition 2.1 (Čech Complex): Consider a finite family of points $\{x_\alpha\}_{\alpha\in\mathcal{A}}$ in the Euclidean space $\mathbb{E}^n$. The **Čech complex** $\mathcal{C}_\epsilon$, parameterized by a scale $\epsilon > 0$, is an abstract simplicial complex. Its k-simplices are formed by subsets $\{x_{\alpha_0}, x_{\alpha_1}, \dots, x_{\alpha_k}\}$ that satisfy the geometric criterion: the intersection of closed balls

$$\bigcap_{i=0}^{k} \overline{B}\left(x_{\alpha_i}, \frac{\epsilon}{2}\right)$$

contains at least one common point in $\mathbb{E}^n$. The Čech complex captures the topological connectivity of the point cloud at scale $\epsilon/2$ through the nerve theorem, which establishes a homotopy equivalence between the union of these balls and the abstract complex $\mathcal{C}_\epsilon$. Compared to the Vietoris-Rips complex — which only requires pairwise distances between points to be less than ϵ — the Čech complex imposes a stricter condition by demanding a global intersection of all $\epsilon/2$-neighborhoods.

Definition 2.2 (Vietoris–Rips Complex): The **Vietoris–Rips complex** $\mathcal{R}_\epsilon$ is algorithmically generated from a discrete point cloud via the following steps: A k-simplex $\{x_{\alpha_0}, \dots, x_{\alpha_k}\}$ is included in $\mathcal{R}_\epsilon$ if every pair of points in the set satisfies

$$\left\|x_{\alpha_i} - x_{\alpha_j}\right\| \le \epsilon, \quad \forall\, 0 \le i < j \le k.$$

This method is computationally more efficient but less accurate in preserving intricate topological details compared to Čech complexes.

Definition 2.3 (Alpha Complex): For a set of points in $\mathbb{E}^n$, the **Alpha complex** $\mathcal{A}_\epsilon$ is constructed using the Delaunay triangulation. A simplex is included in $\mathcal{A}_\epsilon$ if the intersection of the closed balls $B(x_i, \epsilon/2)$ with the corresponding Voronoi cells is non-empty (Edelsbrunner and Mucke[18]).

2.1.1 Mapper Method and Simplicial Complex Construction

The Mapper algorithm is widely recognized as a powerful technique in topological data analysis, particularly effective for summarizing and visualizing high-dimensional complex data structures. It combines clustering, filtration, and graph construction to extract topological insights from high-dimensional datasets. Simplicial complexes generated in Mapper often employ techniques such as discrete Morse theory to simplify and

compute homological features.

Steps in the Mapper Algorithm

（1）**Filter Function:** Apply a continuous filter function $f : X \to \mathbb{R}$ to segment the dataset X. Common filter functions include density measures, principal components, or scalar features derived from data.

（2）**Overlapping Cover:** Divide the range of f into overlapping intervals. For each interval $[a, b]$, identify corresponding data subsets $f^{-1}([a, b])$.

（3）**Clustering within Subsets:** Perform clustering (e.g., k-means or DBSCAN) on each subset $f^{-1}([a, b])$. Nodes in the Mapper graph correspond to clusters.

（4）**Graph Construction:** Connect nodes (clusters) if they share data points in overlapping intervals. This step yields a simplicial complex that approximates the topology of the original data.

Simplicial Complex Construction via Discrete Morse Theory

To optimize homological computations, Mapper frequently incorporates discrete Morse theory. This framework simplifies the topology of a dataset while preserving its essential features. Below are formal definitions that highlight its application:

Definition 2.4 (Discrete Morse Function): Let K be a simplicial complex. A mapping $f : K \to \mathbb{R}$ qualifies as a **discrete Morse function** if $\forall \sigma^{(p)} \in K$, the conditions:

- At most one lower-dimensional simplex $\tau^{(p-1)} \prec \sigma$ satisfies $f(\tau) \geq f(\sigma)$.
- At most one higher-dimensional simplex $\nu^{(p+1)} \succ \sigma$ satisfies $f(\nu) \leq f(\sigma)$.

This formulation establishes a pairing mechanism that simplifies homology computations. Detailed theoretical discussions on these functions can be found in Knudson et al.[40], which addresses algorithmic applications, and Gyulassy[30], which focuses on combinatorial constructions. For additional structural insights, refer to Rote[66].

Definition 2.5 (Discrete Morse Complex): The quotient complex $\mathcal{M}_K$, known as a **discrete Morse complex**, is constructed by collapsing gradient-adjacent simplices:

$$\mathcal{M}_K = K/\sim \quad \text{where } \sigma \sim \tau \text{ if } \sigma \text{ is paired with } \tau \text{ via } f.$$

This operation preserves the homotopy type of K while significantly reducing the number of simplices[21].

Advantages of Čech and Vietoris-Rips Complexes

Mapper typically utilizes Čech or Vietoris-Rips complexes for filtration:

• **Čech Complex:** Constructs simplices based on intersections of data-point neighborhoods. This method ensures homotopy equivalence under the Nerve theorem, offering precise topological representations.

• **Vietoris-Rips Complex:** Constructs simplices by linking points whose pairwise distances fall below a threshold. While computationally efficient, this method may overestimate topological features in dense regions.

Connecting Persistent Homology and Mapper

After constructing simplicial complexes, Mapper enables persistent homology computations. Persistent homology captures the evolution of topological features across filtration scales, providing invariants such as Betti numbers and persistence diagrams.

Remark 2.1: The discrete Morse construction integrated into Mapper facilitates efficient computation of persistent homology by reducing the complexity of simplicial complexes without altering their homotopy type.

A natural question is: **How should one select the parameter ϵ?** Various strategies have been proposed, including local adaptive methods and statistical thresholding, to best capture the intrinsic geometry of the data. In practice, one often considers a range of ϵ values in order to construct a filtration.

2.2 Persistent Homology

Persistent homology, a foundational technique in topological data analysis, provides a robust framework for examining how topological structures evolve and persist across different scales. Central to this approach is the construction of an **ϵ-filtration** — an ordered family $\{\mathcal{K}_\epsilon\}_{\epsilon\geq 0}$ of nested simplicial complexes — which enables the continuous monitoring of homological features. Within this filtration, each topological attribute (e.g., connected components, 1-dimensional cycles, or higher-dimensional cavities) is assigned a pair $(\epsilon_{\text{birth}}, \epsilon_{\text{death}})$ quantifying.

Example 2.1 (Persistence): Let $\mathcal{X} = \{x_\alpha\}_{\alpha\in\mathcal{A}}$ be a stationary point cloud in $\mathbb{E}^n$, and let $(\epsilon_i)_{i=1}^N$ denote a strictly ascending sequence of scale parameters where $0 < \epsilon_1 < \epsilon_2 < \cdots < \epsilon_N$. For each ϵ_i, we construct a Vietoris–Rips complex $\mathcal{R}_i$. The complexes are

linked by natural inclusions

$$\mathcal{R}_1 \overset{\iota}{\hookrightarrow} \mathcal{R}_2 \overset{\iota}{\hookrightarrow} \cdots \overset{\iota}{\hookrightarrow} \mathcal{R}_N.$$

Rather than studying the homology of each $\mathcal{R}_i$ individually, we examine the induced homomorphisms

$$\iota_* : H_*(\mathcal{R}_i) \to H_*(\mathcal{R}_j), \quad i < j,$$

to identify persistent topological features across varying scales.

Lemma 2.1: At every scale parameter $\epsilon > 0$, there is a canonical hierarchy of simplicial embeddings

$$\mathcal{R}_\epsilon \overset{\iota_1}{\hookrightarrow} \mathcal{C}_\epsilon \overset{\iota_2}{\hookrightarrow} \mathcal{R}_{\sqrt{2}\epsilon},$$

where ι_1 and ι_2 denote the natural inclusion maps between Rips and Čech complexes. This sandwich structure arises from the relationship $\mathrm{diam}(\sigma) \leq \sqrt{2}\epsilon$ for any simplex σ in $\mathcal{C}_\epsilon$, guaranteeing the rightmost inclusion.

Definition 2.6 (Persistent Complex): A **persistent complex** C in algebraic topology is defined as a tower of chain complexes $\{C_*^i\}_{i\in\mathbb{N}}$ equipped with connecting morphisms

$$x^{(i)} : C_*^i \to C_*^{i+1}$$

that form a directed system under composition, i.e., satisfying $x^{(i+1)} \circ x^{(i)} = x^{(i+1)}$ for all i. These structure-preserving homomorphisms encode the evolutionary dynamics of filtration construction across scale parameters, ensuring the commutativity of differentials $d^{i+1} \circ x^{(i)} = x^{(i)} \circ d^i$ throughout the persistence hierarchy.

Definition 2.7 (Persistent Homology): Given a persistent complex C and indices $i < j$, the (i, j)-**persistent homology group** is algebraically realized as the persistent image

$$\mathrm{Im}\left(x_*^{(i,j)} : H_k(C_*^i) \to H_k(C_*^j)\right),$$

where $x_*^{(i,j)}$ denotes the homomorphism induced by the composition of chain maps $C_*^i \to C_*^{i+1} \to \cdots \to C_*^j$ through the filtration. This subgroup of $H_k(C_*^j)$ precisely captures homological features born before scale i and surviving until at least scale j. We denote this by $H_*^{i\to j}(\mathsf{C})$.

Remark 2.2: A key result in this theory is the Structure Theorem for persistence modules over a field, which guarantees a unique decomposition into interval modules. This theorem underlies the barcode representation, where each interval corresponds to a topo-

logical feature that persists over a range of scales.

Theorem 2.1: For a finite persistence module over a field F, the homology $H_*(\mathsf{C}; F)$ decomposes as

$$H_*(\mathsf{C}; F) \cong \bigoplus_i \left(x^i \cdot F[x]\right) \oplus \bigoplus_j \left(x^{r_j} F[x]/(x^{s_j} F[x])\right).$$

This decomposition is the theoretical foundation of the persistent barcode.

Definition 2.8 (Barcode): In topological data analysis, a **barcode** serves as a multiset of closed intervals $[b_d, d_d) \subseteq \mathbb{R}$ that encodes the persistence of homological features: the left endpoint of each interval b_d marks the birth scale of a d-dimensional topological structure, while the right endpoint $d_d > b_d$ signifies its disappearance scale. This graphical representation provides a multiscale summary of feature longevity across the filtration.

Theorem 2.2: The algebraic rank of the persistent homology group $H_k^{i \to j}(\mathsf{C}; F)$ corresponds precisely to the cardinality of k-dimensional barcode intervals whose persistence windows contain $[i, j]$. Specially, $\dim H_*(C_*^i; F)$ equals the number of intervals covering the parameter corresponding to i.

2.2.1 Stability of Persistent Homology as Topological Features

An essential advantage of persistent homology is its stability against perturbations in input data, a trait rigorously established through stability theorems. These results ensure that small changes in data, such as noise or measurement errors, yield correspondingly small variations in persistence diagrams, making persistent homology a reliable tool for analyzing real-world data.

Stability of Persistence Diagrams

Let X and Y be two metric spaces, and let d_B represent the bottleneck distance that measures the disparity between the persistence diagrams D_X and D_Y. For any continuous function $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ with finite sublevel set filtrations, the following holds:

$$d_B(D_X, D_Y) \leq ||f - g||_\infty,$$

where $||f - g||_\infty = \sup_{x \in X \cup Y} |f(x) - g(x)|$ is the L^∞ norm. This theorem, proven by Cohen-Steiner et al.[13], guarantees that persistent homology features are stable under bounded perturbations of the filtration function.

Implications for Topological Features.

The stability of persistence diagrams implies that topological features derived from noisy or perturbed datasets retain their significance over small perturbations:

• **Noise Robustness:** Persistent homology captures meaningful topological structures that persist across multiple scales, while ignoring transient features introduced by noise.

• **Feature Reproducibility:** The robustness of persistence diagrams allows consistent extraction of key topological features from different subsamples of the same underlying dataset.

• **Computational Reliability:** Algorithms for computing persistence are designed to efficiently handle perturbations without compromising accuracy, enhancing their utility for large-scale applications.

Practical Considerations in Data Analysis.

When applying persistent homology in real-world scenarios, stability ensures the following:

（1）**Parameter Selection:** Stability justifies the use of parameter sweeps (e.g., varying scale thresholds ϵ) to construct filtration complexes, as small changes in ϵ do not drastically alter the results.

（2）**Handling Imperfect Data:** For datasets with inherent noise or measurement errors, stability enables the extraction of reliable topological descriptors.

（3）**Cross-Domain Applicability:** Stability theorems are particularly valuable in interdisciplinary applications, such as neural network weight analysis (Chapter 5) or audio feature extraction (Chapter 6).

This theoretical assurance underpins the use of persistent homology as a foundation for robust topological data analysis and further enhances its applicability in fields ranging from computational geometry to machine learning.

2.3 Graph Persistent Homology

While classical persistent homology applies to simplicial complexes derived from point clouds, many datasets are naturally represented as graphs. **Graph persistent homology** extends the ideas of persistent homology to graph-structured data by converting graphs into simplicial complexes via the construction of the **adjacency complex**.

2.3.1 Graph Filtration

Define $G = (V, E)$ as a graph equipped with edge weights via $w : E \to \mathbb{R}_{\geq 0}$, enabling weighted adjacency analysis. A common approach is to define a filtration by thresholding the edge weights.

Definition 2.9 (Graph Filtration): For each threshold $\tau \in \mathbb{R}$, define the subgraph

$$G_\tau = \Big(V, \{e \in E \mid w(e) \leq \tau\}\Big).$$

The collection $\{G_\tau\}_{\tau \in \mathbb{R}}$ forms a graph filtration since if $\tau_1 \leq \tau_2$, then $G_{\tau_1} \subseteq G_{\tau_2}$.

2.3.2 Adjacency Complex

The persistent homology framework requires a simplicial complex. For graph data, one common method is to associate each subgraph G_τ with its **adjacency complex**. This complex encodes higher-order interactions by including every clique as an abstract simplex.

Definition 2.10 (Adjacency Complex): Let $G = (V, E)$ be an undirected graph. The *adjacency complex* $\mathcal{X}(G)$ is constructed from the vertex set V such that a subset $\{v_0, \ldots, v_k\} \subset V$ constitutes a k-simplex if every pair $\{v_i, v_j\}$ forms an edge in E.

2.3.3 Definition and Computation of Graph Persistent Homology

Once a graph filtration $\{G_\tau\}$ is established, each subgraph G_τ is converted into its corresponding adjacency complex $\mathcal{X}(G_\tau)$. Graph persistent homology is then defined as follows.

Definition 2.11 (Graph Persistent Homology): Let $G = (V, E)$ be a weighted graph with filtration $\{G_\tau\}$. For each threshold τ, construct the associated adjacency complex $\mathcal{X}(G_\tau)$. The **graph persistent homology** consists of the homology groups

$$H_k(\mathcal{X}(G_\tau)), \quad \tau \in \mathbb{R},$$

together with the linear maps induced by the inclusions

$$\mathcal{X}(G_{\tau_1}) \hookrightarrow \mathcal{X}(G_{\tau_2}), \quad \tau_1 \leq \tau_2.$$

The persistence barcode, as a central object in persistent homology theory, provides a complete combinatorial encoding of the birth and death parameters for homological features (e.g., H_0-type connected components and H_1-type cycles) across the filtration $\{K_t\}_{t \in \mathbb{R}}$.

2.3.4 Computational Procedure

The computation of graph persistent homology typically involves the following steps:

（1）**Filtration Construction:** For the given weighted graph G, construct the filtration $\{G_\tau\}$ by varying the edge weight threshold τ.

（2）**Adjacency Complex Construction:** For each subgraph G_τ, compute its adjacency complex $\mathcal{X}(G_\tau)$. In practice, maximal complete subgraphs are enumerated through algorithms like the Bron–Kerbosch method, with all subsets of each resulting clique subsequently added as simplices.

（3）**Persistent Homology Computation:** Apply standard persistent homology algorithms (typically based on matrix reduction methods) to the nested sequence $\{\mathcal{X}(G_\tau)\}$ in order to compute the persistence barcode.

2.3.5 Example: Social Network Analysis

Example 2.2 (Graph Persistent Homology in Social Networks): Consider a weighted social network $G = (V, E)$, where nodes represent individuals and edge weights quantify interaction frequency. Construct the graph filtration $\{G_\tau\}$ by retaining edges with weight no more than τ. For low τ, the network is sparse, and the corresponding adjacency complex exhibits many isolated vertices (reflected in 0-dimensional homology). As τ increases, groups of nodes form cliques and cycles emerge, corresponding to overlapping communities (captured by 1-dimensional homology). The resulting persistence barcode provides multi-scale topological architecture of complex networks and offers insights into its robustness.

Remark 2.3: Recent studies have further advanced these techniques. For instance, Horak et al.[35] demonstrated the application of persistent homology to complex networks, while Giusti et al.[26] and Sizemore et al.[76] employed clique topology to reveal intrinsic geometric structure in neural and social data. In addition, the work by Grigoryan et al.[28] on homologies of path complexes and digraphs has provided a novel perspective that continues to inspire further developments in graph persistent homology.

In summary, graph persistent homology bridges classical TDA and network science by converting graph data into adjacency complexes and then computing persistent homology on the resulting filtrations. This approach provides a powerful tool for investigating the multi-scale topological features of networks in domains ranging from neuroscience to

social network analysis.

2.4 Time Series and Sliding Window Embedding

In modern data analysis, temporal dynamics embed critical invariants reflecting the structural organization of complex systems. This section establishes a mathematical framework for analyzing video, audio, and other time-dependent data through the lens of nonlinear dynamics and manifold learning (Perea and Harer; Perea et al.,[59];[58]). By combining time-delay embeddings with topological methods, we reveal how transient measurements can capture persistent geometric features of hidden state spaces.

2.4.1 Dynamical Foundation

Definition 2.12: (**Global Continuous Time Dynamical System**) A topological space M (state manifold) endowed with a continuous flow $\Phi : \mathbb{R} \times M \to M$ satisfying:

- **Initial Condition**: $\Phi(0, p) = p$ for all $p \in M$
- **Evolution Law**: $\Phi(s, \Phi(t, p)) = \Phi(s + t, p)$ for all $s, t \in \mathbb{R}$ and $p \in M$

termed a **global continuous-time dynamical system**. The system generates trajectories $\gamma_p(t) = \Phi(t, p)$ that foliate M into non-intersecting orbits (Perea et al.[58]). Common examples include Hamiltonian systems on symplectic manifolds and gradient flows on energy landscapes.

Definition 2.13: (**Observation and Time Series**) Given a dynamical system (M, Φ), an **observation function** $F \in C^1(M, \mathbb{R})$ maps states to scalar measurements. For an initial condition $p \in M$, the induced **time series** is the composition

$$\begin{aligned} \varphi_p : \mathbb{R} &\to \mathbb{R} \\ t &\mapsto F \circ \Phi(t, p) \end{aligned}$$

Practical implementations sample φ_p at discrete times $t_k = k\Delta t$, yielding $\{\varphi_p(t_k)\}_{k=0}^{N}$. The measurement F often represents partial observations (e.g., a single sensor output), making state reconstruction essential (Gakhar and Perea[24]).

2.4.2 Delay Embedding Theory

Takens' theorem[84] provides a principled approach to reconstructing hidden state spaces from scalar time series. The key insight is that successive measurements contain implicit information about the the history of system.

Theorem 2.3 (Takens' Embedding Theorem): Let M be a smooth m-dimensional compact manifold, $\tau > 0$ a delay time, and $d \geq 2m$ an embedding dimension. For generic,

- Dynamics $\Phi \in C^2(\mathbb{R} \times M, M)$ with non-degenerate periodic orbits
- Observation function $F \in C^2(M, \mathbb{R})$ transverse to system trajectories

the delay coordinate map

$$\varphi : M \to \mathbb{R}^{d+1}$$
$$p \mapsto \big(\varphi_p(0), \varphi_p(\tau), \dots, \varphi_p(d\tau)\big)$$

is a diffeomorphic embedding of M into $\mathbb{R}^{d+1}$. Here, "generic" means the property holds for an open dense set of pairs (Φ, F) in the C^2 Whitney topology.

Remark 2.4: Each delayed measurement $\varphi_p(k\tau)$ encodes information about the state of system at time $k\tau$. The embedding dimension $d+1$ must be sufficiently large to "unfold" the manifold, with $2m$ being the minimal requirement to avoid self-intersections (Perea and Harer[59]). The delay parameter τ should be chosen to balance between redundancy (τ too small) and decorrelation (τ too large).

2.4.3 Sliding Window Implementation

Practical applications require adapting Takens' theorem to finite, noisy data streams (Kennel et al.[38]). The sliding window embedding operationalizes delay reconstruction through local averaging and overlap.

Definition 2.14: (**Sliding Window Embedding**) Given a time series $f : \mathbb{R} \to \mathbb{R}$, define the **sliding window operator** with parameters $d \in \mathbb{N}$ (window length) and $\tau > 0$ (stride) as

$$SW_{d,\tau} f : \mathbb{R} \to \mathbb{R}^{d+1}$$
$$t \mapsto (f(t),\ f(t+\tau),\ \dots,\ f(t+d\tau))$$

For discrete sampling times $T = \{t_0, t_0 + \Delta t, \dots, t_0 + N\Delta t\}$, the **sliding window point cloud**, as formulated in (Salas[71]), is expressed as

$$\mathbb{SW}_{d,\tau} f = \big\{(f(t_k), f(t_k+\tau), \dots, f(t_k+d\tau)) \in \mathbb{R}^{d+1} \;\big|\; t_k \in T\big\}.$$

The product $w = d\tau$ is the **window size**, controlling the temporal context captured by each vector.

Parameter Selection Guidelines:

• **Embedding Dimension (d)**: Start with $d = \lceil 2m \rceil$ where m is the estimated manifold dimension. In practice, use false nearest neighbor methods (Kennel et al.[38]) to determine minimal d.

• **Delay (τ)**: Choose using mutual information (first minimum of $I(f(t), f(t+\tau))$) or autocorrelation time.

• **Window Size (w)**: Should span characteristic system timescales. For quasiperiodic signals, w must exceed the beat frequency between incommensurate periods (Perea et al.[58]).

2.4.4 Case Study: Quasiperiodic Dynamics on Torus

Example 2.3 (Irrational Flow on Torus): Consider angular coordinates $(\theta_1, \theta_2) \in \mathbb{T}^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ with dynamics

$$\frac{\mathrm{d}\theta_1}{\mathrm{d}t} = 1, \quad \frac{\mathrm{d}\theta_2}{\mathrm{d}t} = \omega \quad (\omega \notin \mathbb{Q}).$$

The observation function $F(\theta_1, \theta_2) = \cos\theta_1 + \cos\theta_2$ generates a quasiperiodic time series

$$f(t) = \cos t + \cos(\omega t).$$

Applying $SW_{d,\tau}$ with $d = 4$, $\tau = \pi/2$ yields a 5-dimensional point cloud $\mathbb{SW}_{4,\pi/2} f \subset \mathbb{R}^5$. Despite the high ambient dimension, persistent homology reveals the topology of point cloud matches $\mathbb{T}^2$ (Betti numbers $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$), successfully recovering the hidden state space structure (Perea et al.[58]).

2.5 Group Action on Manifolds and Homogeneous Space

The study of group actions on manifolds provides a unifying framework for understanding symmetry and geometric structure in differential geometry and topology. A **homogeneous space** is a topological space X endowed with a continuous transitive action by a topological group G. This section develops the theory of group actions, homogeneous spaces, and their canonical examples, such as Stiefel manifolds. (See Pelliott and Dawber[57]; Lee and Lee[46])

2.5.1 Group Actions and Basic Definitions

Definition 2.15 (Smooth Group Action): Given a Lie group G and smooth manifold M, a **smooth (left) action** of G on M is a smooth map $\theta : G \times M \to M$ satisfying

（1）$\theta(\boldsymbol{e}, x) = x$ for all $x \in M$, where $\boldsymbol{e}$ is the identity in G.

（2）$\theta(\boldsymbol{g}, \theta(\boldsymbol{h}, x)) = \theta(\boldsymbol{gh}, x)$ for all $\boldsymbol{g}, \boldsymbol{h} \in G$ and $x \in M$.

We often write $\boldsymbol{g} \cdot x$ instead of $\theta(\boldsymbol{g}, x)$.

Definition 2.16 (Orbit and Stabilizer): For a group action $\theta : G \times M \to M$ and a point $x \in M$:

- The **orbit** of x is the set $G \cdot x = \{\boldsymbol{g} \cdot x \mid \boldsymbol{g} \in G\} \subset M$.
- The **stabilizer** (or **isotropy subgroup**) of x is the subgroup $G_x = \{\boldsymbol{g} \in G \mid g \cdot x = x\} \subset G$.

Definition 2.17 (Transitive Action): A group action is **transitive** if for any two points $x, y \in M$, $\exists\, \boldsymbol{g} \in G$ with $\boldsymbol{g} \cdot x = y$. Equivalently, M consists of a single orbit: $M = G \cdot x$.

2.5.2 Homogeneous Spaces

A manifold M is called a **homogeneous space** if it admits a transitive smooth action by a Lie group G. Homogeneous spaces are central to geometry because they provide a ”uniform” structure where every point is geometrically indistinguishable.

Theorem 2.4 (Structure of Homogeneous Spaces): Let G act transitively on M, and fix $x \in M$. Then the map

$$\phi : G/G_x \to M, \quad \phi(\boldsymbol{g}G_x) = \boldsymbol{g} \cdot x$$

is a G-equivariant diffeomorphism. Here, G/G_x is the quotient manifold of left cosets.

Proof: The map ϕ is well-defined because $\boldsymbol{g}G_x = \boldsymbol{h}G_x$ implies $\boldsymbol{h}^{-1}\boldsymbol{g} \in G_x$, so $\boldsymbol{h}^{-1}\boldsymbol{g} \cdot x = x$, hence $\boldsymbol{g} \cdot x = \boldsymbol{h} \cdot x$. Transitivity ensures surjectivity, and smoothness follows from the quotient manifold structure. Equivariance is immediate: $\phi(\boldsymbol{g} \cdot \boldsymbol{h}G_x) = \boldsymbol{g} \cdot \phi(\boldsymbol{h}G_x)$. ■

Example 2.4 (Sphere as a Homogeneous Space): The sphere S^n is a homogeneous space under the transitive action of $SO(n+1)$. For any $x \in S^n$, the stabilizer $SO(n+1)_x$ is isomorphic to $SO(n)$, yielding

$$S^n \cong SO(n+1)/SO(n).$$

Similarly, $S^{2n+1} \cong SU(n+1)/SU(n)$.

2.5.3 Orbit Spaces and Quotient Manifolds

When a group action is not transitive, the manifold partitions into orbits. The family of orbits constitutes the **orbit space** M/G, which carries the quotient topology. However,

M/G need not be a manifold unless the action is **free and proper**.

Definition 2.18 (Free and Proper Action):

- An action is **free** if $\boldsymbol{g} \cdot x = x$ implies $\boldsymbol{g} = \boldsymbol{e}$ (i.e., all stabilizers are trivial).
- A **proper action** requires the map $\Phi : G \times M \to M \times M$ with $\Phi(\boldsymbol{g}, x) = (x, \boldsymbol{g} \cdot x)$ to be proper, i.e., $\Phi^{-1}(K)$ remains compact for all compact $K \subseteq M \times M$.

Theorem 2.5 (Quotient Manifold): If G acts freely and properly on M, the orbit space M/G becomes a smooth manifold. The map $\pi : M \to M/G$ establishes a smooth submersion.

Example 2.5 (Real Projective Space): The group $\mathbb{Z}_2$ acts freely and properly on S^n by antipodal maps. The orbit space is the real projective space

$$\mathbb{RP}^n \cong S^n/\mathbb{Z}_2.$$

2.5.4 Stiefel Manifolds

Stiefel manifolds are fundamental examples of homogeneous spaces arising in the study of frame bundles and Grassmannians.

Definition 2.19 (Stiefel Manifold): The **Stiefel manifold** $V_k(\mathbb{R}^n)$ is the set of all orthonormal k-frames in $\mathbb{R}^n$

$$V_k(\mathbb{R}^n) = \left\{(\boldsymbol{v}_1, \dots, \boldsymbol{v_k}) \in (\mathbb{R}^n)^k \,\middle|\, \boldsymbol{v_i} \perp \boldsymbol{v_j}\ \forall i \neq j, \text{ and } \|\boldsymbol{v_i}\| = 1\ \forall i\right\}.$$

Similarly, the complex Stiefel manifold $V_k(\mathbb{C}^n)$ consists of unitary k-frames.

Theorem 2.6 (Stiefel Manifold as a Homogeneous Space): The orthogonal group $O(n)$ acts transitively on $V_k(\mathbb{R}^n)$. For a fixed frame $(\boldsymbol{e}_1, \dots, \boldsymbol{e_k})$, the stabilizer is $O(n-k)$, yielding

$$V_k(\mathbb{R}^n) \cong O(n)/O(n-k).$$

Similarly, $V_k(\mathbb{C}^n) \cong U(n)/U(n-k)$.

Proof: The action $A \cdot (\boldsymbol{v}_1, \dots, \boldsymbol{v_k}) = (\boldsymbol{Av}_1, \dots, \boldsymbol{Av_k})$ is transitive, as every orthonormal frame can form part of an orthonormal basis. The stabilizer of the standard frame consists of matrices in $O(n)$ with a block identity matrix in the first k columns, isomorphic to $O(n-k)$. ■

Remark 2.5: The Stiefel manifold $V_k(\mathbb{R}^n)$ is a compact manifold of dimension $\frac{1}{2}k(2n-k-1)$. It fibers over the Grassmann manifold $Gr_k(\mathbb{R}^n)$ with fiber $O(k)$.

2.5.5 Applications and Further Examples

Example 2.6 (Grassmann Manifold): The Grassmann manifold $Gr_k(\mathbb{R}^n)$, consisting of k-dimensional subspaces of $\mathbb{R}^n$, is a homogeneous space

$$Gr_k(\mathbb{R}^n) \cong O(n)/(O(k) \times O(n-k)).$$

The action $\boldsymbol{A} \cdot \boldsymbol{W} = \boldsymbol{AW}$ is transitive, and the stabilizer of $\mathbb{R}^k \subset \mathbb{R}^n$ is $O(k) \times O(n-k)$.

Example 2.7 (Flag Manifolds): A **flag manifold** parameterizes nested sequences of subspaces $V_1 \subset V_2 \subset \cdots \subset V_k \subset \mathbb{R}^n$. It generalizes Grassmannians and admits a homogeneous structure

$$\mathcal{F}\ell(n; k_1, \ldots, k_m) \cong O(n)/(O(k_1) \times \cdots \times O(k_m)).$$

2.5.6 Principal Bundles and Geometry of Homogeneous Spaces

Homogeneous spaces often arise as base spaces of principal fiber bundles. A key example is the **frame bundle**.

Theorem 2.7 (Frame Bundle as a Principal Bundle): Let M be a smooth n-manifold. The frame bundle $F(M)$, consisting of all bases for tangent spaces T_pM, is a principal $GL(n, \mathbb{R})$-bundle over M. If M is Riemannian, the orthonormal frame bundle $F_O(M)$ is a principal $O(n)$-bundle.

Example 2.8 (Hopf Fibration): The Hopf fibration $S^3 \to S^2$ is a principal $U(1)$-bundle, where $S^3 \cong SU(2)$ and $S^2 \cong SU(2)/U(1)$.

2.5.7 Invariant Metrics and Geodesics

A Riemannian metric on the homogeneous space G/H is **G-invariant** whenever the left G-action preserves the metric through isometries. Such metrics are determined by their value at the identity coset eH.

Theorem 2.8: There exists a bijection between G-invariant metrics on G/H and $\mathrm{Ad}(H)$-invariant inner products on $\mathfrak{m}$.

Example 2.9 (Symmetric Spaces): A **symmetric space** is realized as G/H with involution $\sigma \in \mathrm{Aut}(G)$ satisfying $H = \{\boldsymbol{g} \in G \mid \sigma(\boldsymbol{g}) = \boldsymbol{g}\}$. Examples include spheres, Grassmannians, and classical Lie groups.

2.5.8 Manifold-Frequency Duality Theorem

Theorem 2.9 (Manifold-Spectral Isomorphism): Based on the high-contrast analysis framework from Lee et al.[45] and the stability theory from Chazal et al.[11], let (M, g)

be a compact Riemannian manifold of convolutional features with Laplace-Beltrami operator Δ_g, and $\mathcal{F}$ the Fourier transform over $L^2(M)$. There exists an isometric isomorphism:

$$\Phi : \Gamma(TM) \to \bigoplus_{k=0}^{\infty} \mathcal{H}_k, \tag{2-1}$$

where $\mathcal{H}_k = \{f \in C^\infty(M) \mid \Delta_g f = \lambda_k f\}$ are eigenspaces, satisfying:

$$||\nabla_g f||^2_{L^2(M)} = \sum_{k=0}^{\infty} \lambda_k |\mathcal{F}(f)(k)|^2.$$

Proof: Let $\{\phi_k\}$ be orthonormal eigenfunctions of Δ_g. For any $f \in C^\infty(M)$, expand:

$$f = \sum_{k=0}^{\infty} \langle f, \phi_k \rangle \phi_k.$$

Compute gradient energy:

$$\begin{aligned} ||\nabla_g f||^2 &= \int_M g(\nabla_g f, \nabla_g f) dV_g \\ &= \int_M f \Delta_g f \, dV_g \quad \text{(Green's identity)} \\ &= \sum_{k=0}^{\infty} \lambda_k |\langle f, \phi_k \rangle|^2. \end{aligned}$$

Define $\mathcal{F}(f)(k) = \langle f, \phi_k \rangle$, then:

$$||\nabla_g f||^2 = \sum_{k=0}^{\infty} \lambda_k |\mathcal{F}(f)(k)|^2. \qquad \blacksquare$$

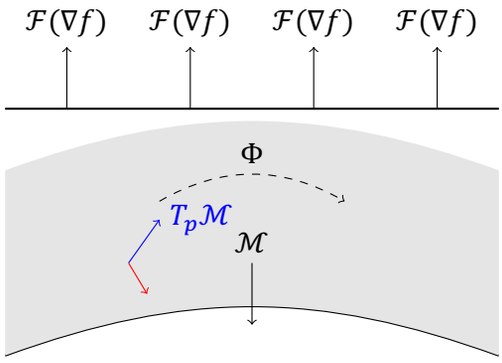


Figure 2-1 Manifold-spectral duality: Gradient flows on feature manifold $\mathcal{M}$ correspond to high-frequency components in Fourier domain

2.5.9 Orthogonal Basis on Feature Manifolds

Building upon the metric g in Theorem 2.9, we construct orthogonal bases through geometric Gram-Schmidt process:

（1）Project raw kernels $\{K_i\}$ onto tangent space $T_p\mathcal{M}$.

（2）Iterative orthogonalization: For each K_i,

$$K_i^{\perp} = K_i - \sum_{j=1}^{k-1} \frac{\langle K_i, B_j \rangle_g}{\langle B_j, B_j \rangle_g} B_j,$$

where $\langle \cdot, \cdot \rangle_g$ is induced by Theorem 2.9.

（3）Retain $B_k = K_i^{\perp} / ||K_i^{\perp}||_g$ if $||K_i^{\perp}||_g > \epsilon$

Proposition 2.1 (Approximation Completeness): [45] For any $\delta > 0$, $\exists$ orthogonal basis $\{B_j\}_{j=1}^{m}$ such that:

$$\min_{c_j} \left\| K - \sum c_j B_j \right\|_g < \delta, \quad \forall K \in \mathcal{M}.$$

CHAPTER 3 INTERDICIPLINARY TOOLS FOR TOPOLOGICAL DEEP LEARNING

This chapter examines the intersection of topological data analysis and speech recognition through two methodological perspectives.

The first section introduces the fundamental concepts of neural networks.

The second section explores various topological neural networks, with a primary focus on Topological Convolutional Neural Networks (TCNNs), a framework proposed by Love et al.[53]. TCNNs integrate persistent homology with convolutional operations to analyze multiscale topological features of spectrograms. This approach aims to enhance feature extraction capabilities within non-Euclidean data spaces, as theorized in their foundational study.

The third section outlines the core components of speech recognition technology. Beginning with the linguistic basis of phonemes—the discrete sound units that constitute spoken language—the discussion transitions into computational methods. Established architectures, including GMM-HMM, Recurrent Neural Networks (RNNs) (Hochreiter and Schmidhuber[33]) and Deep Fully Convolutional Neural Networks (DFCNNs) (Abdel-Hamid et al.[1]), are reviewed to highlight their roles in modeling temporal dependencies and hierarchical acoustic patterns.

While TCNNs propose novel interactions between topology and speech processing, their practical effectiveness remains an active subject of research. This chapter synthesizes these concepts while maintaining a focus on rigorously documented mechanisms, avoiding assumptions of unvalidated synergies.

3.1 Neural Networks Architectures

3.1.1 Convolutional Neural Network

Definition 3.1: (**Feed-Forward Neural Network (FFNN)**) The **Feed-Forward Neural Network (FFNN)** is abstractly characterized by a directed acyclic graph Γ with vertex set $V(\Gamma)$, formally defined through three structural axioms:

（1）The vertex set $V(\Gamma)$ is partitioned into disjoint layers:

$$V(\Gamma) = V_0(\Gamma) \sqcup V_1(\Gamma) \sqcup \cdots \sqcup V_r(\Gamma).$$

（2）For any vertex $v \in V_i(\Gamma)$, every edge (v, w) in Γ satisfies $w \in V_{i+1}(\Gamma)$.

（3）For any non-initial node $w \in V_i(\Gamma)$ where $i > 0$, at least one vertex $v \in V_{i-1}(\Gamma)$ is required to form an edge (v, w) within Γ.

The elements of $V(\Gamma)$, when viewed as graph components, are alternatively termed **nodes**. Each $V_i(\Gamma)$ denotes **layer-i nodes**, with $V_0(\Gamma)$ and $V_r(\Gamma)$ being the **input** and **output** layers respectively.

The edge connections between adjacent layers V_i and V_{i+1} are encoded by a **correspondence** $C_i \subseteq V_i \times V_{i+1}$, where $(v, w) \in C_i \Leftrightarrow (v \to w) \in E(\Gamma)$. The neighborhood mappings are defined through:

$$C(v_0) := \{w \in V_{i+1} \mid (v_0, w) \in C_i\} \quad \text{(Forward activation domain)}$$

$$C^{-1}(w_0) := \{v \in V_i \mid (v, w_0) \in C_i\} \quad \text{(Backward dependency set).}$$

Definition 3.2: (**Fully Connected Layer**) Within the FFNN architecture, a layer V_{i+1} is **fully connected** precisely when its edge correspondence $C \subseteq V_i \times V_{i+1}$ becomes maximal, i.e., $C_c = V_i \times V_{i+1}$. This complete bipartite connectivity pattern implies:

- Every node $v \in V_i$ connects to all nodes $w \in V_{i+1}$
- The adjacency matrix $\mathbf{A}_i \in \{0, 1\}^{|V_i| \times |V_{i+1}|}$ has all entries equal to 1.

Definition 3.3: (**Normal One Layer (NOL)**) A layer V_{i+1} in an FFNN attains the structure of a **convolutional layer** (termed **normal one layer**) under the following conditions:

- **Vertex decompositions**: $V_i = \chi \times \mathbb{Z}^N$ and $V_{i+1} = \chi' \times \mathbb{Z}^N$ for finite channel sets χ, χ' and spatial dimension $N \in \mathbb{N}^*$
- **Edge correspondence**: Determined by a spatial radius parameter $s \geq 0$, with

$$C = C_c \times C_{d,N}(s) \subset (\chi \times \chi') \times (\mathbb{Z}^N \times \mathbb{Z}^N),$$

where:

  - $C_c = \chi \times \chi'$ enforces full channel-wise connectivity.
  - $C_{d,N}(s)$ constrains spatial neighbors via

$$C_{d,N}(s)(\boldsymbol{x}') := \left\{\boldsymbol{x} \in \mathbb{Z}^N \,\middle|\, \max_{1 \leq k \leq N} |x_k - x'_k| \leq s\right\}.$$

The L^∞-metric governing spatial proximity is canonically defined as:

$$d_{\mathbb{Z}^N}(\boldsymbol{x}, \boldsymbol{x}') := \max_{1 \leq k \leq N} |x_k - x'_k|.$$

Definition 3.4: (**Pooling Layer**) A layer V_{i+1} in a FFNN is termed a **pooling layer** if $V_i = \chi \times \mathbb{Z}^N$ and $V_{i+1} = \chi \times \mathbb{Z}^N$ for some finite set χ, and positive integers N and s. The

edge-defining correspondence $C \subset V_i \times V_{i+1}$ is defined by

$$C = C_{id} \times C_{N,s},$$

where $C_{id} = \chi \times \chi$ is the identity correspondence, given by

$$C_{id}^{-1}(\kappa) = \{\kappa\}$$

for all $\kappa \in \chi$, and $C_{N,s} \subset \mathbb{Z}^N \times \mathbb{Z}^N$ is defined as

$$C_{N,s}^{-1}(x_1', x_2', \dots, x_N') := \left\{(x_1, x_2, \dots, x_N) \in \mathbb{Z}^N \mid 0 \leq x_i - s x_i' \leq s - 1 \text{ for } 1 \leq i \leq N\right\}$$

for all $(x_1', x_2', \dots, x_N') \in \mathbb{Z}^N$.

A **Convolutional Neural Network (CNN)** is a feedforward architecture where hierarchical feature extraction governs layer composition:

- **Feature Encoding Phase**: Layers $V_0, V_1, \dots, V_i$ implement convolutional operations
- **Decision Phase**: Layers $V_{i+1}, V_{i+2}, \dots, V_{r-1}$ establish full connectivity

Remark 3.1: In standard CNN architectures, **pooling layers** are systematically interleaved with convolutional layers to perform spatial downsampling. This dimensional reduction serves dual purposes: decreasing computational load through dimensionality contraction while simultaneously enhancing feature abstraction by enforcing local translation invariance. The combined effect yields hierarchical representations invariant to minor spatial perturbations. In the context of Topological Convolutional Neural Networks (TCNNs), pooling layers can be utilized in the same manner and for similar purposes as in traditional CNNs.

Information propagation in the network is governed by a dual mathematical framework operating over Γ:

- **Weight parameters** $\Lambda = \{\lambda_{v,w} \in \mathbb{R} \mid (v,w) \in E(\Gamma),\ v \in V_{i-1},\ w \in V_i\}$ modulate inter-layer signal transmission
- **Activation functions** $\{f_w\}_{w \in V(\Gamma)}$ assigned to nodes transform propagated signals

In the context of a CNN, let $V_{i-1} = \chi \times \mathbb{Z}^N$ and $V_i = \chi' \times \mathbb{Z}^N$. We denote nodes as $v = (\kappa, \boldsymbol{x}) \in V_{i-1}$ and $w = (\kappa', \boldsymbol{x}') \in V_i$. A hallmark of CNNs is the homogeneity of weights, expressed through translational invariance

$$\lambda_{(\kappa,\boldsymbol{x}),\ (\kappa',\boldsymbol{x}')} = \lambda_{(\kappa,\boldsymbol{x}+\boldsymbol{z}),\ (\kappa',\boldsymbol{x}'+\boldsymbol{z})}.$$

The activation system $\mathcal{A} = \{(u_v, f_v)\}_{v \in V(\Gamma)}$ associates each node v with a scalar state $u_v \in \mathbb{R}$ and a nonlinear transform $f_v : \mathbb{R} \to \mathbb{R}$. Data propagation from V_{i-1} to V_i

entails calculating states u_w at nodes $w \in V_i$ through the synaptic integration formula:

$$u_w = f_w\left(\sum_{\substack{v\in V_{i-1}\\(v,w)\in\Gamma}} \lambda_{v,w}u_v\right).$$

In neural networks, the **activation functions** f_v typically map real numbers $x \in \mathbb{R}$ to ranges such as $0 < x < 1$ or $x \geq 0$. The output layer activations are constrained to form a probability distribution over $V_r(\Gamma)$, necessitating non-negativity ($u_v \geq 0$) and normalization ($\sum_{v\in V_r} u_v = 1$). For hidden layer computations, the Rectified Linear Unit (ReLU) $f(x) = \max(0, x)$ serves as the canonical activation function, introducing sparsity by thresholding negative pre-activations while preserving linear response in the positive domain. This piecewise-linear nonlinearity balances expressivity with gradient stability during backpropagation. For the terminal layer, we use the softmax function

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}, \quad i, j \in \{1, \dots, n\},\ x_i \in \mathbb{R}^+,$$

to produce a probability distribution over the outputs.

3.2 Topological Convolutional Neural Network

Consider a manifold M, and let $\chi, \chi' \subset M$ be two finite subsets serving as discretizations of M. Given successive FFNN layers $V_i = \chi \times \mathbb{Z}^N$ and $V_{i+1} = \chi' \times \mathbb{Z}^N$ equipped with a metric $d : \chi \times \chi' \to \mathbb{R}_+$, the s-threshold correspondence $C(s) \subseteq \chi \times \chi'$ is characterized by its inverse images:

$$\forall \kappa' \in \chi', \quad C(s)^{-1}(\kappa') := \{\kappa \in \chi \mid d(\kappa, \kappa') \leq s\}.$$

This constructs a parameterized family of adjacency relations where s controls the receptive field radius in the feature space χ.

By introducing dual thresholds $s \geq 0$ (channel) and $s' \geq 0$ (spatial), the composite correspondence $C \subset V_i \times V_{i+1}$ emerges as a tensor product:

$$C = C(s) \times C_{d,N}(s'),$$

where $C(s) \subset \chi \times \chi'$ governs channel-wise connectivity through metric d, and $C_{d,N}(s') \subset \mathbb{Z}^N \times \mathbb{Z}^N$ controls spatial locality via L^∞-metric. For any output node $(\kappa', \boldsymbol{x}') \in \chi' \times \mathbb{Z}^N$,

the pre-image decomposes orthogonally:

$$C^{-1}(\kappa', \boldsymbol{x}') = \underbrace{\{\kappa \in \chi \mid d(\kappa, \kappa') \leq s\}}_{\text{channel filter}} \times \underbrace{\{\boldsymbol{x} \in \mathbb{Z}^N \mid \max_k |x_k - x'_k| \leq s'\}}_{\text{spatial patch}}.$$

This Cartesian product structure simultaneously enforces feature similarity in χ-space and spatial proximity in $\mathbb{Z}^N$, creating a convolutional template that slides across the input lattice while maintaining channel-specific pattern matching.

3.2.1 Sheaf-Theoretic Foundations

Sheaf Neural Networks: A sheaf $\mathcal{F}$ on a graph G assigns data spaces to vertices and edges, with restriction maps $\rho_{uv} : \mathcal{F}(u) \to \mathcal{F}(v)$ for edges (u, v). This framework generalizes graph neural networks by enforcing local consistency through sheaf cohomology[31].

Speech Sheaf Construction: For spectrogram patches $U \subset \mathbb{R}^2$, define the sheaf $\mathcal{S}$ as:

$$\mathcal{S}(U) = \{f : U \to \mathbb{C} \mid f \text{ is locally stationary}\}$$

with restriction maps $\rho_{UV}(f) = f|_V$. Cohomology groups $H^1(\mathcal{S})$ classify topological obstructions in speech dynamics[14].

3.2.2 Topological Signatures in Speech

Persistence Landscapes: Given a spectrogram $\mathcal{G}$, compute its persistence landscape $\Lambda_{\mathcal{G}}(t)$ as:

$$\Lambda_{\mathcal{G}}(t) = \sup\{\lambda \mid (t - \lambda, t + \lambda) \in \text{Barcode}(H_1(\mathcal{G}))\}.$$

This signature provides multiscale topological features for phoneme classification[34].

Regularization via Signatures: Augment the loss function with:

$$\mathcal{L}_{\text{topo}} = \sum_{i=1}^{n} \|\Lambda_{\mathcal{G}_i} - \Lambda_{\mathcal{G}_i^{\text{clean}}}\|_2$$

penalizing deviations from clean topological profiles[34].

3.2.3 Klein Bottle Convolution

2D Images

Definition 3.5: (**Circle Correspondence**) Let $\chi, \chi' \subset S^1$ constitute finite discrete approximations of the base circle. Define adjacent network layers in the FFNN as $V_i = \chi \times \mathbb{Z}^2$ and $V_{i+1} = \chi' \times \mathbb{Z}^2$. Fix a threshold $s \geq 0$.

The **circle correspondence** $C_S(s) \subset \chi \times \chi'$ is defined by

$$C_S(s)^{-1}(\kappa') = \{\kappa \in \chi \mid d_S(\kappa, \kappa') \leq s\}$$

for all $\kappa' \in \chi'$, where the metric d_S is given by

$$d_S(\kappa, \kappa') = \cos^{-1}(\kappa \cdot \kappa'), \quad \kappa, \kappa' \in S^1.$$

A vertex set V_{i+1} attains the **circle one layer (COL)** designation when, given an auxiliary threshold parameter $s' \geq 0$, its edge correspondence $C \subset V_i \times V_{i+1}$ exhibits the product structure:

$$C = C_S(s) \times C_{d,2}(s'),$$

where $C_{d,2}(s')$ denotes the convolutional component performing neighborhood aggregation in $\mathbb{Z}^2$. This implies that for all $(\kappa', x', y') \in \chi' \times \mathbb{Z}^2$,

$$\begin{aligned} C^{-1}(\kappa', x', y') &= C_S(s)^{-1}(\kappa') \times C_{d,2}(s')^{-1}(x', y') \\ &= \left\{(\kappa, x, y) \in \chi \times \mathbb{Z}^2 \mid d_S(\kappa, \kappa') \leq s \text{ and } d_{\mathbb{Z}^2}((x, y), (x', y')) \leq s'\right\}. \end{aligned}$$

We subsequently establish a weight localization mechanism on the Klein bottle $\mathcal{K}$, leveraging its geometric structure through coordinate-dependent parametrization to constrain network parameters. Recall that $\mathcal{K}$ is the two-dimensional manifold formed by taking $\mathbb{R}^2$ and applying the identifications $(\theta_1, \theta_2) \sim (\theta_1 + 2k\pi, \theta_2 + 2l\pi)$ for $k, l \in \mathbb{Z}$ and $(\theta_1, \theta_2) \sim (\theta_1 + \pi, -\theta_2)$.

We construct a geometric embedding $F_{\mathcal{K}} : \mathcal{K} \hookrightarrow \mathcal{Q}([-1, 1]^2)$, where $\mathcal{Q}([-1, 1]^2)$ denotes the space of quadratic functions over the unit square, extending the framework for manifold representations in CNNs established by (Carlsson and Gabrielsson; Carlsson et al.[8-9]). Each image patch $F_{\mathcal{K}}(\theta_1, \theta_2)$ encodes orientation information parameterized by the angular coordinate $\theta_1 \in S^1$. Geometrically, this manifests as directional features orthogonal to the central axis of θ_1, with visual representations exhibiting line patterns rotated by $\theta_1 + \pi/2$ radians relative to the image plane.

The embedding is defined by

$$F_{\mathcal{K}}(\theta_1, \theta_2)(x, y) = \sin(\theta_2)(\cos(\theta_1)x + \sin(\theta_1)y) + \cos(\theta_2)Q\,(\cos(\theta_1)x + \sin(\theta_1)y)\,,$$

where $Q(t) = 2t^2 - 1$.

This mapping $F_{\mathcal{K}}$ is inherently a function on the torus $\mathbb{T}^2$, parameterized by θ_1 and θ_2. However, due to the identifications $F_{\mathcal{K}}(\theta_1, \theta_2) = F_{\mathcal{K}}(\theta_1 + 2k\pi, \theta_2 + 2l\pi)$ and $F_{\mathcal{K}}(\theta_1 + \pi, -\theta_2) = F_{\mathcal{K}}(\theta_1, \theta_2)$, it naturally descends to a function on the Klein bottle $\mathcal{K}$.

Definition 3.6: (**Klein Correspondence**) Consider finite discrete samplings $\chi, \chi' \subset \mathcal{K}$ of the parameter space of Klein bottle. Within a feed-forward neural architecture, we define adjacent vertex layers $V_i = \chi \times \mathbb{Z}^2$ and $V_{i+1} = \chi' \times \mathbb{Z}^2$. Let $s \geq 0$ specify the neighborhood radius for connection establishment.

The **Klein correspondence** $C_{\mathcal{K}}(s) \subset \chi \times \chi'$ is defined by

$$C_{\mathcal{K}}(s)^{-1}(\kappa') = \{\kappa \in \chi \mid d_{\mathcal{K}}(\kappa, \kappa') \leq s\}$$

for all $\kappa' \in \chi'$, where the metric $d_{\mathcal{K}}$ is given by

$$d_{\mathcal{K}}(\kappa, \kappa') = \left(\int_{[-1,1]^2} \left[F_{\mathcal{K}}(\kappa)(x, y) - F_{\mathcal{K}}(\kappa')(x, y) \right]^2 \mathrm{d}x\, \mathrm{d}y \right)^{\frac{1}{2}}$$

for $\kappa, \kappa' \in \mathcal{K}$.

We designate V_{i+1} as a **Klein one layer (KOL)** when the edge correspondence $C \subset V_i \times V_{i+1}$ admits the decomposition

$$C = C_{\mathcal{K}}(s) \times C_{d,2}(s'),$$

where $s, s' \geq 0$ are distance thresholds. For any node $(\kappa', x', y') \in \chi' \times \mathbb{Z}^2$, the preimage satisfies

$$\begin{aligned} C^{-1}(\kappa', x', y') &= \bigcup_{\substack{\kappa \in \chi \\ d_{\mathcal{K}}(\kappa, \kappa') \leq s}} \{\kappa\} \times C_{d,2}(s')^{-1}(x', y') \\ &= \left\{ (\kappa, x, y) \in \chi \times \mathbb{Z}^2 \;\middle|\; \begin{array}{l} d_{\mathcal{K}}(\kappa, \kappa') \leq s \\ \wedge\; d_{\mathbb{Z}^2}((x, y), (x', y')) \leq s' \end{array} \right\}. \end{aligned}$$

Definition 3.7: (**CF or KF Layer**) Let the base manifold M be either S^1 or $\mathcal{K}$, equipped with a finite discrete sampling $\chi \subset M$. In a feed-forward neural architecture, we construct adjacent layers $V_i = \mathbb{Z}^2$ (input grid) and $V_{i+1} = \chi \times \mathbb{Z}^2$ (output layer). The layer V_{i+1} attains convolutional functionality when endowed with a neighborhood radius parameter $s \geq 0$.

We define V_{i+1} as a **Circle Features (CF) layer** when $M = S^1$, or a **Klein Features (KF) layer** when $M = \mathcal{K}$, if the weights $\lambda_{-,(\kappa,-,-)}$ for $\kappa \in \chi$ are derived via convolution over V_i. Specifically, the filter of size $(2s+1) \times (2s+1)$ has values

$$\text{Filter}(\kappa)(n,m) = \int_{-1+\frac{2m}{2s+1}}^{-1+\frac{2(m+1)}{2s+1}} \int_{-1+\frac{2n}{2s+1}}^{-1+\frac{2(n+1)}{2s+1}} F_M(\kappa)(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

for integers $0 \leq n, m \leq 2s$.

A Reason for the Klein Bottle

To contextualize our analytical framework (cf. Carlsson[7]), we adopt a function-theoretic viewpoint of the Klein bottle. The 3×3 image patches are interpreted as discrete samples obtained by evaluating smooth functions $f : D \to \mathbb{R}$ at nine predetermined grid points $\{p_k\}_{k=1}^{9} \subset D$. Our investigation focuses on identifying closed subspaces $\mathcal{F} \subset C(D, \mathbb{R})$ that satisfy the approximation property:

$$\sup_{f \in \mathcal{F}} ||f||_{L^2(\{p_k\})} \approx ||f||_{L^2(D)},$$

where the left-hand norm corresponds to patch space measurements.

Let Q denote the space of bivariate quadratic polynomials, explicitly parametrized as

$$f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2 \quad (A, \ldots, F \in \mathbb{R}).$$

This constitutes a six-dimensional real vector space. Our analysis focuses on the constrained subspace $\mathcal{P} \subseteq Q$ defined by the conditions

$$\int_D f(x,y)\,\mathrm{d}x\mathrm{d}y = 0 \quad \text{(mean centering)}, \qquad \int_D f(x,y)^2\,\mathrm{d}x\mathrm{d}y = 1 \quad \text{(contrast normalization)}.$$

The linear constraint alone reduces Q to a five-dimensional affine subspace, while the quadratic normalization further restricts $\mathcal{P}$ to a four-dimensional ellipsoid embedded within this subspace.

We subsequently characterize the submanifold $\mathcal{P}_0 \subseteq \mathcal{P}$ consisting of functions with the specialized form

$$f(x,y) = q(\lambda x + \mu y),$$

where q is a single-variable quadratic function, and $\lambda^2 + \mu^2 = 1$. The space of such functions within Q is 4-dimensional-three parameters define q, and (λ, μ) lies on the unit circle, which is one-dimensional. Incorporating the two additional constraints reduces this to a 2-dimensional complex $\mathcal{P}_0$.

We demonstrate that $\mathcal{P}_0$ is homeomorphic to the Klein bottle $\mathcal{K}$ via the following construction. Define the function space A as containing all univariate quadratic polynomials of the form

$$q(t) = c_0 + c_1 t + c_2 t^2 \quad (c_i \in \mathbb{R})$$

subject to the integral constraints

$$\int_{-1}^{1} q(t)\mathrm{d}t = 0 \quad \text{(zero mean)}, \qquad \int_{-1}^{1} q^2(t)\mathrm{d}t = 1 \quad \text{(unit energy)}.$$

These lead to the equations

$$c_0 + \frac{c_2}{3} = 0 \quad \text{and} \quad c_0^2 + \frac{2c_0c_2 + c_1^2}{3} + \frac{c_2^2}{5} = \frac{1}{2}.$$

Simplifying, we obtain

$$3c_0 + c_2 = 0 \quad \text{and} \quad \frac{8c_0^2}{5} + \frac{2c_1^2}{3} = 1.$$

The constrained solution set constitutes an elliptical manifold in $\mathbb{R}^3$, exhibiting circular topology through standard diffeomorphism.

Given a direction vector $\boldsymbol{v} \in \mathbb{R}^2$ with $||\boldsymbol{v}|| = 1$ and a polynomial $q \in A$, we construct the directional function $q_{\boldsymbol{v}} : \mathbb{R}^2 \to \mathbb{R}$ via the parameterization

$$q_{\boldsymbol{v}}(\boldsymbol{w}) = q(\boldsymbol{v} \cdot \boldsymbol{w}) \quad \forall \boldsymbol{w} \in \mathbb{R}^2,$$

where $\boldsymbol{v} \cdot \boldsymbol{w}$ denotes the Euclidean inner product. For a unit vector $\boldsymbol{v}$ and $q \in A$, this statement is verifiable with straightforward reasoning,

$$\int_D q_{\boldsymbol{v}} = 0 \quad \text{and} \quad \int_D q_{\boldsymbol{v}}^2 \neq 0.$$

Thus, the mapping

$$(q, \boldsymbol{v}) \mapsto \frac{q_{\boldsymbol{v}}}{||q_{\boldsymbol{v}}||_2}$$

defines a continuous function θ from $A \times S^1$ to $\mathcal{P}_0$. However, θ is not a homeomorphism. The normalized mapping

$$\theta : A \times S^1 \to \mathcal{P}_0, \quad (q, \boldsymbol{v}) \mapsto \frac{q_{\boldsymbol{v}}}{||q_{\boldsymbol{v}}||_2}$$

is continuous but fails to be a homeomorphism due to the non-injective nature of the parameterization. This can be seen by introducing the involution $\rho : A \to A$ defined by

$$\rho(c_0 + c_1 t + c_2 t^2) = c_0 - c_1 t + c_2 t^2.$$

The mapping θ exhibits the symmetry relation

$$\theta(q, \boldsymbol{v}) = \theta(\rho(q), -\boldsymbol{v}),$$

which induces a well-defined quotient mapping over the orbit space $\mathcal{O} = (A \times S^1)/\sim$, where $\sim$ denotes equivalence under the involution $(q, \boldsymbol{v}) \sim (\rho(q), -\boldsymbol{v})$.

The quotient mapping $\bar{\theta} : \mathcal{O} \to \mathcal{P}_0$ satisfies two fundamental properties:

(a)$\bar{\theta}$ constitutes a homeomorphism preserving the quotient topology,

(b)The orbit space $\mathcal{O}$ exhibits the topological structure $\mathcal{O} \cong \mathcal{K}$.

Remark 3.2: There is an alternative way to consider the $\mathcal{K}$ by quotient maps. The original space $\mathcal{Q}$ is homeomorphic to $\mathbb{R}^3 \times S^1$. The mean centering can be considered as a quotient θ_1 as

$$\theta_1(q) = q_1,$$

where $q(t) = c_0 + c_1 t + c_2 t^2$ and $q_1(t) = c_{01} + c_1 t + c_2 t^2$ satisfying the mean centering condition. The unit energy can be considered as a quotient θ_2 as

$$\theta_1(q) = q_2,$$

where $q_2 = \frac{q}{||q||_2}$.

Define the involution $f : \mathcal{Q} \to \mathcal{Q}$ by

$$f(q)(t) = q_0(t) = c_0 - c_1 t + c_2 t^2,$$

which reverses the sign of the linear term c_1. This satisfies $f^2 = \mathrm{id}$.

The quotient θ_1 enforces $\int_{S^1} q(t)dt = 0$, eliminating c_0. The reduced space is:

$$\theta_1(\mathcal{Q}) \cong \mathbb{R}^2 \times S^1 \quad \text{(parameters } (c_1, c_2) \in \mathbb{R}^2,\ t \in S^1\text{)}.$$

Under f, the coefficients transform as $(c_1, c_2) \mapsto (-c_1, c_2)$.

The quotient θ_2 normalizes the energy:

$$\theta_2(q) = \frac{(c_1, c_2)}{||(c_1, c_2)||_2} \in S^1 \quad \text{(unit circle)}.$$

The resulting space after θ_2 is a fiber bundle over S^1 with fiber S^1.

The involution f acts on the normalized coefficients as:

$$f : (c_1, c_2) \mapsto (-c_1, c_2) \Longrightarrow (\cos\theta, \sin\theta) \mapsto (\cos(\pi - \theta), \sin(\pi - \theta)).$$

This corresponds to a reflection $\theta \mapsto \pi - \theta$ on S^1. Simultaneously, the base S^1 (original $t \in S^1$) is twisted by a half-period shift $t \mapsto t + \pi$ due to the phase dependency in $\mathcal{Q}$. The total space is constructed by gluing the fibers S^1 over the base S^1 with a reflection map.

This gluing is equivalent to the Klein bottle:

$$\mathcal{K} \cong (S^1 \times S^1)/\sim, \quad (\theta, t) \sim (\pi - \theta, t + \pi).$$

Since the involution f introduces a non-orientable twist in both the fiber and base, the quotient space is the Klein bottle.

Videos

Let us denote coordinates in $\mathbb{R}^3 \times \mathbb{R}^3$ by the variables $(\theta_1, \theta_2, r, u, v, w)$. The triplet (θ_1, θ_2, r) parameterizes $\mathcal{K}^t$, while (u, v, w) parameterize its tangent spaces. Specifically, $\mathcal{K}^t$ is defined as the quotient of $\mathbb{R}^3$ under the relations $(\theta_1, \theta_2, r) \sim (\theta_1 + 2k\pi,\ \theta_2 + 2l\pi,\ r)$ for all $k, l \in \mathbb{Z}$, and $(\theta_1, \theta_2, r) \sim (\theta_1 + \pi,\ -\theta_2,\ -r)$. Similarly, A quotient construction over $\mathbb{R}^3 \times \mathbb{R}^3$ characterizes the tangent bundle $T(\mathcal{K}^t)$. We omit further discussion of these identifications, as they are relevant only insofar as they are preserved by the embeddings $F_{\mathcal{K}^t}$ and $F_{T(\mathcal{K}^t)}$.

Define $I = [-1, 1]$. Let $C(I^2, I)$ denote the space of continuous functions from I^2 to I, representing image patches at infinite resolution. Similarly, let $C(I^2 \times I, I)$ denote the space of video patches. The embeddings

$$F_{\mathcal{K}^t} : \mathcal{K}^t \to C(I^2, I)$$

and

$$F_{T(\mathcal{K}^t)} : T(\mathcal{K}^t) \to C(I^2 \times I, I)$$

are defined by

$$\begin{aligned} F_{\mathcal{K}^t}(\theta_1, \theta_2, r)(x, y) = \sin(\theta_2)\left(\cos(\theta_1)\big(x + r\cos(\theta_1)\big) + \sin(\theta_1)\big(y + r\sin(\theta_1)\big)\right) \\ + \cos(\theta_2)\, Q\left(\cos(\theta_1)\big(x + r\cos(\theta_1)\big) + \sin(\theta_1)\big(y + r\sin(\theta_1)\big)\right), \end{aligned}$$

and

$$F_{T(\mathcal{K}^t)}(\theta_1, \theta_2, r, u, v, w)(x, y, t) = F_{\mathcal{K}^t}\big(\theta_1 + tu,\ \theta_2 + tv,\ r + tw\big),$$

where $Q(z) = 2z^2 - 1$.

The embedding $F_{T(\mathcal{K}^t)}$ induces a metric structure

$$d_{T(\mathcal{K}^t)}(\kappa, \kappa') = \left(\int_{I^2 \times I} \Big(F_{T(\mathcal{K}^t)}(\kappa)(x, y, t) - F_{T(\mathcal{K}^t)}(\kappa')(x, y, t)\Big)^2 \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t\right)^{\frac{1}{2}}$$

for $\kappa, \kappa' \in T(\mathcal{K}^t)$ on $T(\mathcal{K}^t)$ via pullback of the L^2-metric defined on the function space $C(I^2 \times I, I)$. This metric $d_{T(\mathcal{K}^t)}$ enables us to define a new type of layer in a neural

network.

Definition 3.8: (**6D Moving Klein Correspondence and 6MKOL**) Let $\chi, \chi' \subset T(\mathcal{K}^t)$ be two finite subsets. For consecutive layers $V_i = \chi \times \mathbb{Z}^3$ and $V_{i+1} = \chi' \times \mathbb{Z}^3$ in a FFNN, given a fixed threshold $s \geq 0$, the **6D Moving Klein correspondence**

$$C_{T(\mathcal{K}^t)}(s) \subset \chi \times \chi'$$

is characterized by the correspondence condition

$$C_{T(\mathcal{K}^t)}(s)^{-1}(\kappa') := \left\{ \kappa \in \chi \;\middle|\; d_{T(\mathcal{K}^t)}(\kappa, \kappa') \leq s \right\}$$

holding for all $\kappa' \in \chi'$, where the metric $d_{T(\mathcal{K}^t)}$ is induced by the pullback construction discussed previously.

A vertex set V_{i+1} is designated as a **6D Moving Klein one layer (6MKOL)** provided an auxiliary threshold $s' \geq 0$ exists, with its edge correspondence $C \subset V_i \times V_{i+1}$ defined via

$$C = C_{T(\mathcal{K}^t)}(s) \times C_{d,3}(s'),$$

specifically requiring that for every $(\kappa', x', y', t') \in \chi' \times \mathbb{Z}^3$,

$$\begin{aligned} C^{-1}(\kappa', x', y', t') &= C_{T(\mathcal{K}^t)}(s)^{-1}(\kappa') \times C_{d,3}(s')^{-1}(x', y', t') \\ &= \left\{ (\kappa, x, y, t) \in \chi \times \mathbb{Z}^3 \;\middle|\; \begin{array}{l} d_{T(\mathcal{K}^t)}(\kappa, \kappa') \leq s \\ \wedge\, d_{\mathbb{Z}^3}\big((x, y, t),\ (x', y', t')\big) \leq s' \end{array} \right\}. \end{aligned}$$

Submanifold Selection Principle. Within the fiber bundle $T(\mathcal{K}^t)$, there exist distinguished submanifolds $\mathcal{M}_\alpha \hookrightarrow T(\mathcal{K}^t)$ whose associated video patches $\Psi(\mathcal{M}_\alpha) \subset C(I^2 \times I, I)$ are hypothesized to be critical for spatiotemporal pattern recognition. The 6MKOL architecture implements this geometrically through the layer design:

$$\chi = \bigcup_{\alpha \in A} \mathcal{D}(\mathcal{M}_\alpha), \quad \chi' = \bigcup_{\beta \in B} \mathcal{D}(\mathcal{M}_\beta),$$

where $\mathcal{D}$ denotes adaptive discretization operators preserving topological invariants of $\mathcal{M}_\alpha$.

Complexity-accuracy Tradeoff. Full discretization of $T(\mathcal{K}^t)$ leads to computationally prohibitive filter cardinalities. For example, with parameter quantization:

$$\theta_1, \theta_2 \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}, \quad \xi_j \in \{-1, 0, 1\}\ (j = 1, \ldots, 4)$$

yields χ with $|\chi| = 4^2 \times 3^4 = 1296$ coordinate charts. Each chart encodes spatiotemporal features as $5 \times 5 \times 5$ tensors (spatial resolution × temporal depth), totaling 125 pixels per

filter.

Regularization through Geometry. This submanifold focus is philosophically aligned with our **geometric Occam's razor**: Minimal sufficient parameterization of observable dynamics via

$$\dim \mathcal{M}_\alpha \ll \dim T(\mathcal{K}^t) = 6$$

intrinsically controls model capacity while preserving discriminative power, thereby mitigating overfitting in video classification tasks.

We focus on five significant 2-dimensional submanifolds of $T(\mathcal{K}^t)$, defined as:

$$\tilde{\mathcal{K}} := \{(\theta_1, \theta_2, 0, 0, 0, 0) \in T(\mathcal{K}^t)\},$$
$$S_\tau^\pm := \{(\theta_1, \theta_2, 0, 0, 0, \pm 1) \in T(\mathcal{K}^t)\},$$
$$S_\rho^\pm := \{(\theta_1, \theta_2, 0, \pm 1, 0, 0) \in T(\mathcal{K}^t)\}.$$

Through the embedded parametrization $F_{T(\mathcal{K}^t)}$, the submanifold $\tilde{\mathcal{K}}$ is parameterized by temporally invariant Klein bottle embeddings. The dynamical regimes $S_\tau^\pm$ encode rigid translations of Klein bottle configurations orthogonal to their primary symmetry axis, where the superscript $\pm$ specifies the translation polarity. Analogously, $S_\rho^\pm$ characterizes axial rotation with chirality determined by the $\pm$ index.

Definition 3.9: (**M-F Layer**) Given an admissible subset $M \subset T(\mathcal{K}^t)$ (topologically characterized as submanifolds $\tilde{\mathcal{K}}$, $S_\tau^\pm$, $S_\rho^\pm$, or their stratified unions), the discrete configuration space χ is defined as a finite discretization $\chi := D(M) \subset M$ via an operator D with cardinality constraints, where D preserves key topological invariants of M.

For a FFNN with adjacent layers $V_i = \mathbb{Z}^3$ and $V_{i+1} = \chi \times \mathbb{Z}^3$, where V_i operates as a convolutional module with activation threshold $s \geq 0$, the subsequent layer V_{i+1} acquires the classification of a **Manifold-Features (M-F) layer** if the weights $\lambda_{-,(\kappa,-,-,-)}$ for each $\kappa \in \chi$ are determined by convolving over V_i using a filter of dimensions $(2s+1) \times (2s+1) \times (2s+1)$, where the filter values are defined as

$$\text{Filter}(\kappa)(n, m, p) = \int_{-1+\frac{2n}{2s+1}}^{-1+\frac{2(n+1)}{2s+1}} \int_{-1+\frac{2m}{2s+1}}^{-1+\frac{2(m+1)}{2s+1}} \int_{-1+\frac{2p}{2s+1}}^{-1+\frac{2(p+1)}{2s+1}} F_{T(\mathcal{K}^t)}(\kappa)(x, y, t)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}t$$

for integers $0 \leq n, m, p \leq 2s$.

3.2.4 Equivariant Neural Network Architectures

Equivariant neural networks incorporate symmetry constraints directly into the architecture. Taco Cohen and Max Welling's work[12] formalized G-equivariance for arbitrary

groups G, enabling advanced feature extraction in neural networks. These architectures ensure that features transform predictably under group actions, making them highly effective for structured data.

Definition 3.10 (Equivariance): A function $f : \mathcal{X} \to \mathcal{Y}$ is G**-equivariant** if:

$$f(g \cdot x) = g \cdot f(x), \quad \forall g \in G, x \in \mathcal{X},$$

where G, acting on input space $\mathcal{X}$, is a symmetry group. This property enforces consistency in transformations, ensuring that structural relationships in the data are preserved.

Theoretical Foundations

Group-equivariant architectures extend classical convolutional layers by encoding group symmetries such as rotations, reflections, or translations into the network's structure. Representation theory plays a critical role here by enabling the decomposition of high-dimensional inputs into invariant components under group actions ([12]). This leads to more efficient computations and improved generalization.

Applications in Computer Vision and Physics

Equivariant neural networks have demonstrated remarkable success in domains requiring symmetry-aware analysis:

- **Pose Estimation**: Effectively identifying object orientations and spatial alignments in images.
- **Molecular Dynamics**: Modeling physical interactions where symmetry groups like SO(3) describe rotational behaviors.
- **Astronomical Data Processing**: Classifying galaxies and analyzing trajectories in datasets with inherent rotational and translational symmetries.
- **Medical Imaging**: Detecting rotationally invariant patterns in 3D scans and other volumetric data.

Advantages Over Traditional Architectures

Equivariant neural networks offer significant advantages over standard architectures, including:

- Reduced parameter count by sharing weights across symmetric transformations.
- Enhanced robustness to perturbations by preserving invariance under group actions.

- Improved efficiency in extracting features from high-dimensional data.

3.2.4.1 Future Directions

Future advancements in equivariant neural networks may include:

- Integration with attention mechanisms to refine symmetry-aware processing.
- Extending equivariant principles to graph neural networks, enabling symmetry analysis on relational data structures.
- Application to dynamic systems using non-compact groups such as $SE(3)$ to model continuous transformations.

Equivariant neural networks remain a cornerstone in bridging mathematical theory with practical applications, paving the way for innovations across disciplines ([12]).

3.3 Foundations of Speech Recognition Technology

As a core research pillar in intelligent systems, speech signals, paralleling visual data modalities like images and videos, underpin critical applications ranging from automatic recognition to noise suppression and synthetic generation. Notable breakthroughs include: visual-assisted speech enhancement through lip movement analysis (Zheng et al.[99]), The optimized end-to-end recognition pipeline achieving benchmark performance of Microsoft (Li[48]), and their contemporaneous innovations in neural speech synthesis architectures (Tan et al.[85]). The convergence of articulatory phonetics with deep learning has enabled systems achieving $95\%+$ word accuracy on clean speech, though challenges persist in noisy environments and low-resource languages.

3.3.1 Phonetic Building Blocks

Phonemes are systematically classified into vowels and consonants according to articulatory characteristics. The rhythmic interplay between these units forms the structural basis of spoken language. This hierarchical organization drives research emphasis toward suprasegmental analysis (words/sentences), where expanded contextual dependencies enable more reliable pattern identification.

Modern systems employ a three-tiered processing hierarchy:

- **Phoneme Level**: 40-60 basic units (English: 44 phonemes) with 50-200ms duration
- **Syllable Level**: 10,000+ possible combinations through phoneme concatenation

- **Prosodic Level**: Pitch contours and stress patterns conveying semantics

The precise alignment between transient acoustic features and discrete phonetic symbols remains challenging, particularly for coarticulated phonemes where adjacent sounds blend spectrally.

3.3.2 Phonetic Classification via IPA Standards

The International Phonetic Alphabet (IPA) categorizes phonemes into three primary classes: pulmonic consonants, non-pulmonic consonants, and vowels. Our analysis focuses exclusively on pulmonic consonants and vowels, as non-pulmonic consonants exhibit negligible prevalence in English. Pulmonic consonants are produced by constricting airflow at the glottis (the space between vocal folds) or oral cavity while coordinating pulmonary airflow, exemplified by symbols such as [b], [p], [m], and [n].

Consonants are further specified through three articulatory dimensions:

（1）**Place**: Bilabial [p], Alveolar [t], Velar [k]

（2）**Manner**:

- •Plosives [ptk]
- •Fricatives [szf]
- •Nasals [mnŋ]
- •Approximants [jw]

（3）**Voicing**: Vocal fold vibration (e.g., [z] vs. [s])

Vowels are systematically mapped in IPA based on lingual positioning (Figure 3-1), quantified through:

- **Height**:
  - High [i]
  - Mid [e]
  - Low [a]
- **Backness**:
  - Front [i]
  - Central [ə]
  - Back [u]
- **Roundedness**:
  - Rounded [y]
  - Unrounded [i]

To streamline English phonetic notation, ARPABET emerged as a practical alterna-

tive, mapping 39 English phonemes to ASCII combinations (Figure 3-2). This system enables efficient computational processing through:

- Single-letter vowels: AA [ɑ], AE [æ]
- Two-letter consonants: SH [ʃ], TH [θ]
- Stress markers: Primary (ˈ), secondary (ˌ)

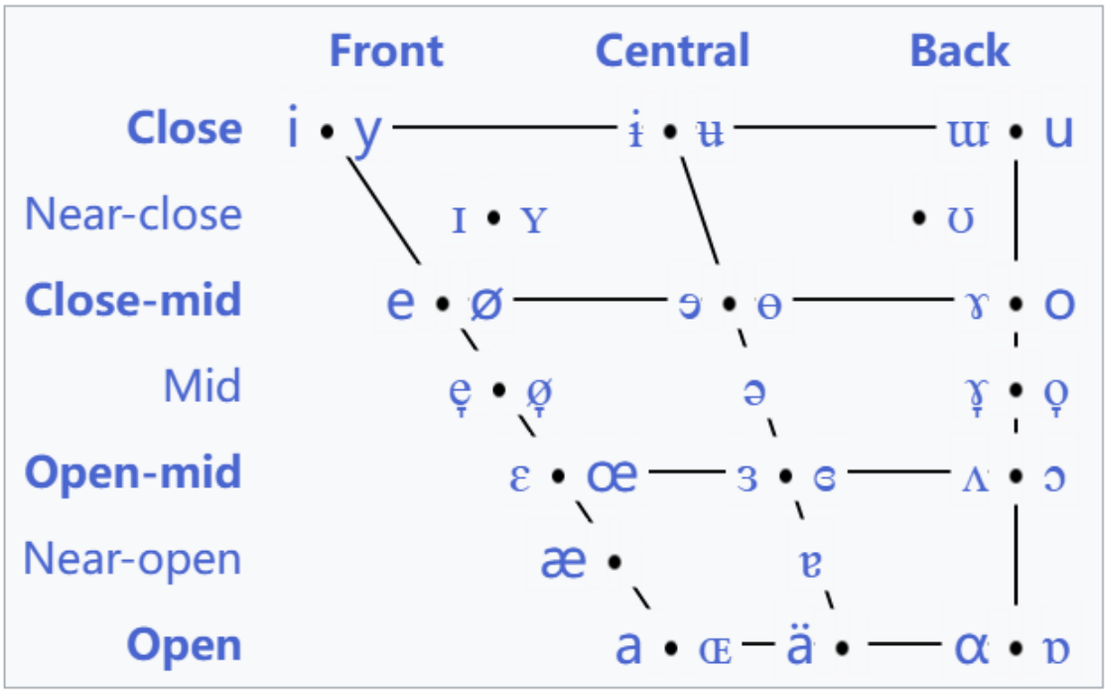


Figure 3-1 Positioning of vowels in oral cavity

AA	ɑ~ɒ	balm, bot (with father–bother merger)
AE	æ	bat
AH	ʌ	butt
AO	ɔ	caught, story
AW	aʊ	bout
AX	ə	comma
AXR[3]	ɚ	letter, forward
AY	aɪ	bite
EH	ɛ	bet
ER	ɝ	bird, foreword
EY	eɪ	bait
IH	ɪ	bit
IX	ɨ	roses, rabbit
IY	i	beat
OW	oʊ	boat
OY	ɔɪ	boy
UH	ʊ	book
UW	u	boot
UX[3]	ʉ	dude

(a) ARPABET vowel-phoneme mapping table

B	b	buy
CH	tʃ	China
D	d	die
DH	ð	thy
DX	ɾ	butter
EL	l	bottle
EM	m	rhythm
EN	n	button
F	f	fight
G	g	guy
HH *or* H[3]	h	high
JH	dʒ	jive
K	k	kite
L	l	lie
M	m	my
N	n	nigh
NX *or* NG[3]	ŋ	sing
NX[3]	ɾ̃	winner

(b) ARPABET consonant mapping table I

P	p	pie
Q	ʔ	uh-oh
R	ɹ	rye
S	s	sigh
SH	ʃ	shy
T	t	tie
TH	θ	thigh
V	v	vie
W	w	wise
WH	ʍ	why (without wine–whine merger)
Y	j	yacht
Z	z	zoo
ZH	ʒ	pleasure

(c) ARPABET consonant mapping table II

Figure 3-2 ARPABET phonetic notation system

The above two figures (see Figure 3-1 and Figure 3-2) are both from Wikipidea (ht

tps://en.wikipedia.org/wiki/International_Phonetic_Alphabet; https://en.wikipedia.org/wiki/ARPABET).

3.3.3 Historical Context: GMM-HMM Frameworks

Before the rise of deep learning, the GMM-HMM framework, combining Gaussian Mixture Models and Hidden Markov Models, was considered the benchmark for speech recognition, as highlighted by Rabiner in his foundational work[62]. Additionally, this classical framework has been further extended and analyzed in various robust speech recognition contexts, as discussed by Sun et al.[83]. These models characterized sequential relationships between phonemes and acoustic features via probabilistic state transitions:

- **Gaussian Mixture Models (GMMs):** Modeled frame-level acoustic feature distributions using parameterized mean and covariance.
- **Hidden Markov Models (HMMs):** Captured temporal dynamics of phoneme sequences through state-based Markov chains.

Although GMM-HMM achieved early success in simple recognition tasks, its limitations included:

- Ineffectiveness in capturing long-term temporal dependencies.
- Reliance on hand-crafted features like MFCCs, which constrained generalizability to broader contexts.

The introduction of Recurrent Neural Networks revolutionized this paradigm by enabling end-to-end learning and integrating sequential memory propagation directly into model architectures.

3.3.4 Recurrent Neural Networks (RNNs)

Prior to computational advancements, GMM-HMM frameworks dominated speech recognition systems, despite their mechanistic divergence from human neural processing. Recurrent Neural Networks (RNNs) revolutionized this paradigm by inherently modeling sequential dependencies through contextual memory propagation. The temporal characteristics of speech signals can be observed through their waveforms, which exhibit amplitude variations over time. These patterns serve as the initial step in processing raw audio data for further analysis, such as phoneme recognition or spectrogram generation. The following figure displays waveforms corresponding to nine spoken commands in the Mini Speech Commands dataset, offering a visual representation of the differences in acoustic patterns.

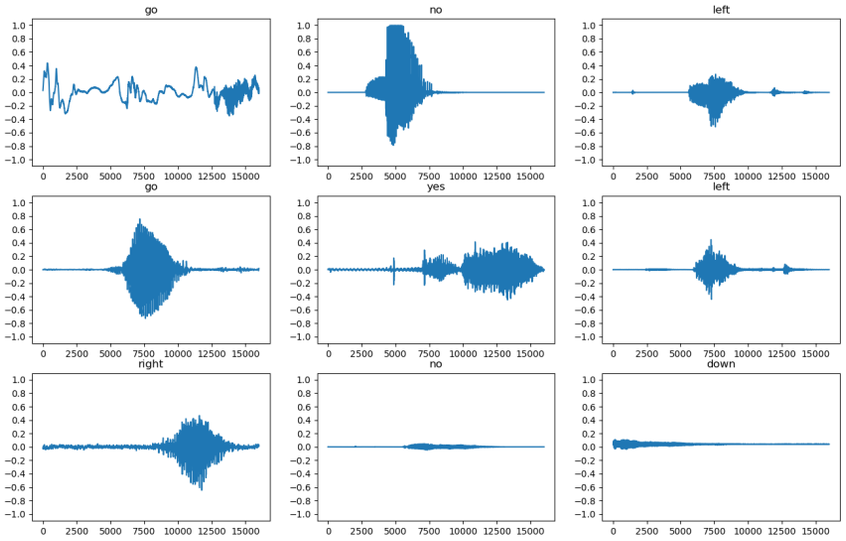


Figure 3-3 Waveforms of Mini Speech Commands

The LSTM variant introduced gated memory cells:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),$$

where f_t is forget gate activation, enabling selective retention of phonetic context across hundreds of time steps. Bidirectional LSTMs further improved phone error rates (PER) to $< 15\%$ by processing sequences forwards and backwards.

Transformer Integration: The self-attention mechanism in Transformers (Vaswani et al.[93]) revolutionized phonetic modeling through:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q, K, V represent query, key, and value matrices. This enables:

- Parallel processing of entire utterances
- Direct modeling of phoneme-syllable-word dependencies
- State-of-the-art PER $< 8\%$ on TIMIT benchmark

3.3.5 Time Delay Embedding for Speech Signals

The transformation of raw speech signals into structured representations often employs time delay embedding, a method rooted in Takens' Theorem[59]. This technique reconstructs the hidden state space of dynamic systems from scalar time series, enabling the analysis of phonetic structures.

Let $s(t)$ denote a speech signal over time. A time-delay embedding constructs vectors $\mathbf{s}_k$ as:

$$\mathbf{s}_k = [s(t_k), s(t_k + \tau), \ldots, s(t_k + (d-1)\tau)],$$

where τ is the delay parameter, d is the embedding dimension, and t_k represents discrete sample times. This framework captures temporal dependencies and is foundational in generating time-frequency representations for spectrogram computation.

Applications:

- **Phoneme Dynamics:** Capturing periodic and quasi-periodic behaviors in articulatory signals.
- **Feature Generation:** Creating meaningful point clouds for time-frequency analysis.

3.3.6 Speech Signal to Image Representation

Convolutional Neural Networks (CNNs) offer effective feature extraction paradigms for acoustic processing. The Time Delay Neural Network (TDNN), among the earliest CNN-based speech recognition architectures, performs simultaneous convolutions along both frequency and temporal axes, thereby capturing variable-length contextual dependencies (Sainath et al.[69];[70]).

The Deep Fully Convolutional Neural Network (DFCNN) introduced by iFLYTEK processes spectrograms as 2D images through:

- Frequency-axis convolution: Learns Mel-filterbank equivalents
- Time-axis convolution: Discovers triphone patterns
- Residual blocks: 40+ convolutional layers

As shown in Figure 3-1 and Figure 3-2, this resembles spectral analysis methodologies in phonetics. Zhang et al.[97] achieved 4.7% WER on Switchboard using:

- 128-channel log-Mel inputs (300ms context)
- 15 convolutional blocks with batch normalization
- Connectionist Temporal Classification (CTC) output

Short-Time Fourier Transform (STFT) for Speech Signals

The conversion of raw speech waveforms into spectrograms begins with the Short-Time Fourier Transform (STFT), which decomposes the signal into its frequency components across time intervals[55].

Formally, the STFT of a signal $x(t)$ is given by:

$$X(f,t) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f \tau} \, \mathrm{d}\tau,$$

where $w(t)$ denotes a window function (such as Hamming or Gaussian windows) centered at each temporal point t, and f corresponds to the frequency domain. This approach captures localized frequency content while preserving temporal resolution.

To illustrate the transformation of speech signals from waveforms to spectrograms, we apply the Short-Time Fourier Transform (STFT). This process captures temporal and frequency-domain features, providing a foundation for subsequent audio analysis. The following figure demonstrates an example of a speech waveform (top) and its corresponding spectrogram (bottom), offering a clear visualization of how sound evolves across time and frequency domains.

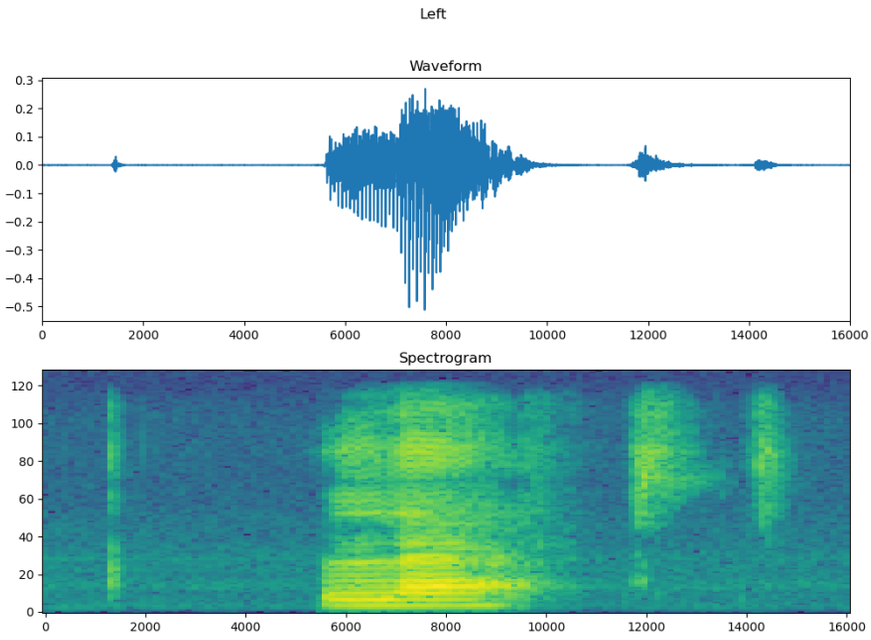


Figure 3-4 Waveform and Corresponding Spectrogram on mini speech commands

Spectrograms are widely used in speech recognition to capture unique frequency patterns associated with spoken commands. For instance, in the Mini Speech Commands dataset[94], spectrograms reveal distinct features for commands such as ”go”, ”stop”, and ”yes”. The following figure provides a comparison of spectrograms for nine different commands, arranged in a 3×3 grid, highlighting the frequency characteristics that aid in differentiating these spoken instructions.

Applications in Speech Recognition:

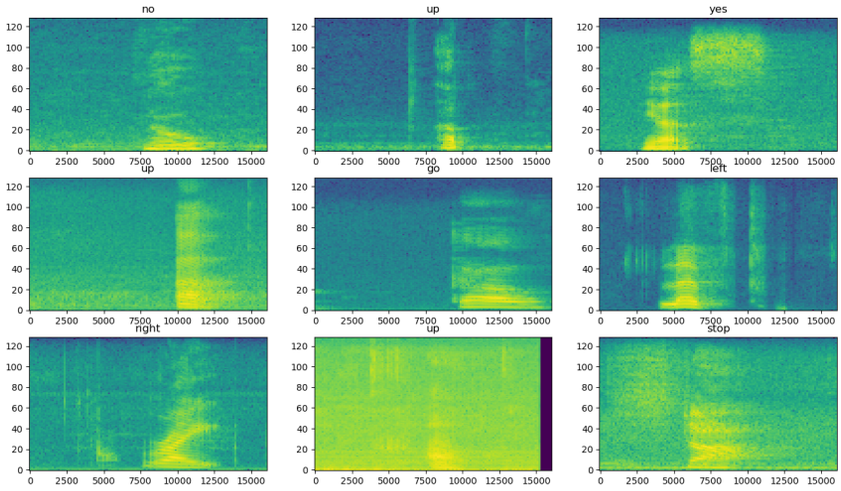


Figure 3-5 Spectrograms of Mini Speech Commands

- **Feature Extraction:** Capturing pitch, formant, and harmonic structures.
- **Spectrogram Analysis:** Enabling input preparation for convolutional neural networks (CNNs).

3.4 Topological Fusion of Audio Features

MFCC-HPCP Fusion: For a time-frequency patch P, compute both MFCCs $\mathbf{m} \in \mathbb{R}^{13}$ and HPCPs $\mathbf{h} \in \mathbb{R}^{12}$, then concatenate into a topological descriptor:

$$\mathbf{v}_P = [\mathbf{m}, \mathbf{h}, \text{pers}(H_1(P))] \in \mathbb{R}^{25+n},$$

where $\text{pers}(H_1(P))$ encodes the persistence of harmonic loops[88].

Cover Song Analogy: The fusion method is inspired by cover song identification, where topological consistency across variations is critical[88].

CHAPTER 4 TOPOLOGICAL DEEP LEARNING: FROM IMAGE DATA TO SPEECH DATA

In this chapter, we aim to replicate partial experimental results and explore the role of topological information in both image and speech data. Although the overall outcomes are somewhat modest, this chapter serves as an important bridge between conventional performance evaluations and the discovery of intrinsic data structures.

We begin by replicating experiments on widely used image datasets such as MNIST[5] and CIFAR10[42]. In these sections, we conduct a comparative analysis of the representational efficacy between canonical Convolutional Neural Networks (CNNs) and that of Topologically Configured CNNs (TCNNs) by analyzing loss curves and accuracy metrics. This replication confirms existing findings and establishes a baseline for understanding the potential benefits of integrating topological methods.

The chapter then shifts focus to the domain of speech processing using the SpeechBox dataset. Here, we detail the process of phoneme segmentation and spectrogram generation, followed by training CNNs for phoneme recognition. By examining confusion matrices and utilizing an ensemble of network weight vectors, we employ Principal Component Analysis (PCA) alongside persistent homology to uncover latent topological structures within the data.

While the direct impact of infusing topological insights on model accuracy remains modest, the exploratory analyses presented herein lay a promising foundation for future work. This investigation not only replicates prior results but also offers a novel perspective by linking traditional performance metrics with the underlying topology of the data.

4.1 Topological Data Analysis in Natural Images

This section delves into Carlsson's seminal works (De Silva and Carlsson[75]), (Carlsson et al.[9]), which explore the local topological properties of spaces formed by natural images. By investigating these spaces, we aim to uncover the underlying geometric and topological structures that shape natural image datasets.

4.1.1 Datasets and Preprocessing

The primary dataset under consideration, denoted as $\mathcal{M}$, comprises 4×10^6 high-contrast image patches of size 3×3, sampled from the still image database curated using the van Hateren-Schaaf natural image ensemble ([92]). Additionally, $\mathcal{M}$ constitutes a proper subcollection within the ambient superset $\tilde{\mathcal{M}}$ containing approximately 8×10^6 patches, provided by Pedersen. To extract high-contrast patches, the following multi-step procedure was implemented:

（1）**Initial Sampling**:

- Randomly select 5000 image patches of size 3×3 from a chosen image.
- Embed local image regions into a 9-dimensional Euclidean space via vectorization of pixel intensities.

（2）**Intensity Transformation**:

- Perform pixel-wise logarithmic mapping $\log : \mathcal{I} \to \mathbb{R}$ on the intensity domain $\mathcal{I} \subset \mathbb{R}_{\geq 0}$.
- Perform component-wise demeaning through the linear transformation $\mathbf{v} \mapsto \mathbf{v} - \mu_{\mathbf{v}} 1$, where $\mu_{\mathbf{v}} := \frac{1}{n} \sum_{i=1}^{n} v_i$ denotes the empirical mean, and $\mathbf{1} \in \mathbb{R}^n$ denotes the canonical basis vector with all components equal to unity (explicitly $\mathbf{1} = (1, 1, \ldots, 1)^{\top}$).

（3）**Contrast Selection**:

- For each mean-centered vector, calculate its contrast (or "$\boldsymbol{D}$-norm") defined as

$$||\boldsymbol{x}||_{\boldsymbol{D}} = \sqrt{\boldsymbol{x}^T \boldsymbol{D} \boldsymbol{x}},$$

  where $\boldsymbol{D}$ is a 9×9 matrix that is both symmetric and positive-definite.
- Retain only the patches that rank in the top 20% in terms of $\boldsymbol{D}$-norm.

（4）**Normalization and Embedding**:

- Spatiotemporal patches are normalized onto a 7-dimensional ellipsoidal manifold through quadratic constraints, embedding the data in $\mathbb{R}^8$ while intrinsically reducing dimensionality via curvature-driven parametrization.
- Apply a coordinate transformation to project the data onto the unit sphere in $\mathbb{R}^8$.

This process yields the curated dataset $\mathcal{M}$, optimized for topological analysis. In the following sections, we examine the local and global topological characteristics embedded in this dataset.

4.1.2 Persistent Homology Analysis

To uncover robust topological features, persistent homology provides a framework for studying local density variations and global connectivity within $\mathcal{M}$. The focus is on leveraging density filtration, denoising techniques, and the construction of witness complexes to compute and interpret homological invariants.

Density Filtration

The density of a data point $x \in X$ is estimated using the k-nearest neighbor method. For a given $k > 0$, let $\rho_k(x)$ represent the distance to the k-th nearest neighbor. Smaller values of ρ_k correspond to higher local density. The dataset is filtered by ordering points by density and extracting the top p percent for analysis, forming subsets $X(k, p)$. This method highlights core regions of the data, which often contain significant topological information that may be obscured in the full dataset.

Denoising

To improve computational efficiency and minimize noise in the data, the following denoising process was applied:

（1）**Nearest Neighbor Averaging**:

•Replace each data point with the mean value of its neighboring points.

（2）**Iterative Smoothing**:

•Perform the averaging process iteratively, typically repeating it twice, to generate a cleaner and more refined representation of the data.

Witness Complex Construction

The witness complex $W_\infty(D)$ is utilized as a primary tool for constructing simplicial complexes by leveraging the distances between n landmarks and N data points. The key steps in this process are as follows:

（1）**Construction of Edges and Higher-Dimensional Simplices:**

•Edges and simplices are built based on the ordering of distances between data points and landmarks.

（2）**Efficient Approximation Using Lazy Witness Complexes:**

•Computationally efficient approximations, such as the lazy witness complex $W_1(D)$, are employed to reduce complexity in practical applications.

Landmarks are selected through random sampling or the maxmin algorithm, ensuring representative coverage of the dataset.

4.1.3 Homological Insights

By applying the methodologies outlined above, we computed the persistent homology of subsets within $\mathcal{M}$. The key findings are summarized as follows:

（1）**Smaller Subset Analysis:**

- Using $n = 50$ landmarks and $k = 15$, the first Betti number (β_1) was calculated as 5, consistent with the three-circle model C_3.
- When k was increased to 300, the structure simplified, resulting in a single dominant feature with a first Betti number of 1.

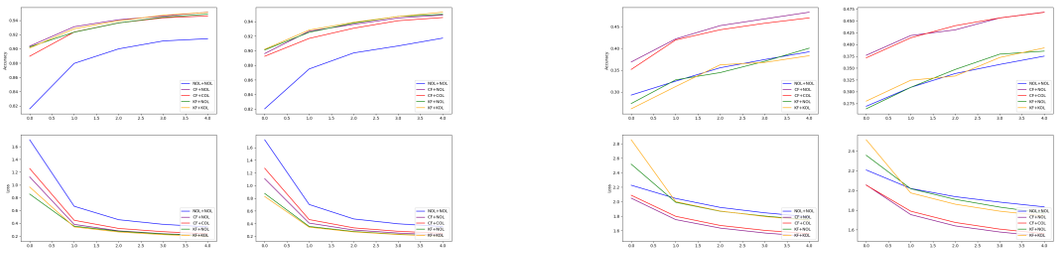
（2）**Larger Dataset Insights:**

- The persistent homology of the subset $X(100, 10) \cup Q \subset \mathcal{M}$ revealed two prominent a set of 1-cycles $\{\gamma_i\}_{i=1}^{n}$ forming a free abelian group basis, together with a 2-cocycle ω spanning the second cohomology.
- This result aligns with the topological structure of a Klein bottle rather than a torus, as suggested by experimental and theoretical considerations.

Remark 4.1: The three-circle model C_3 can be embedded into the Klein bottle $\mathcal{K}$, providing a theoretical link between local and global topological phenomena.

4.2 Reproduction of Main Image Results

This methodological segment faithfully replicates the experimental findings documented in the source study, applying both CNN and TCNN to the MNIST and CIFAR10 datasets. The outcomes are depicted in Figure 4-1, with all experimental parameters aligned with those described in (Love et al.[53]).



(a) MNIST Loss and Accuracy (b) CIFAR10 Loss and Accuracy

Figure 4-1 Comparison of Loss and Accuracy between CNN and TCNN on Two Datasets

4.2.1 Observations and Analysis

MNIST Dataset

For the MNIST dataset, all TCNN configurations demonstrated noticeable improvements in accuracy compared to conventional CNNs. These improvements were consistent across various setups, suggesting that TCNNs effectively capture the critical features of simpler datasets with lower complexity, such as MNIST. The enhanced performance underscores the capacity of TCNN for achieving better generalization and optimization.

CIFAR10 Dataset

When applied to the more complex CIFAR10 dataset, the performance trends differed significantly. The accuracy of all configurations experienced a drop due to the increased data diversity and complexity of dataset. However, among the TCNN variants, the combination corresponding to COL exhibited the most significant accuracy gains, outperforming both the other TCNN configurations and the baseline CNN. Notably, the KOL configuration, despite achieving relatively lower accuracy within the TCNN family, still matched or slightly exceeded the performance of the standard CNN. This finding highlights the resilience and adaptability of TCNN when applied to datasets with higher visual and structural complexity.

4.2.2 Conclusions

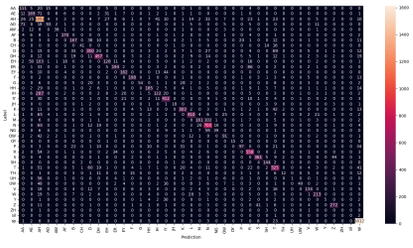
The experimental results illustrate that TCNN offers effective and tangible improvements over conventional CNN models. For simpler datasets like MNIST, TCNN consistently outperforms traditional methods, demonstrating its robustness in extracting key features. On the more challenging CIFAR10 dataset, TCNN continues to show potential, with certain configurations achieving substantial gains. This supports the argument that TCNN, as a novel architecture, offers an advantage in performance across a range of scenarios, making it a promising direction for further exploration and optimization.

4.3 Exploration of Topological Information in Speech Data

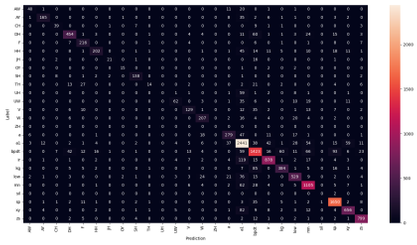
We begin by addressing the task of phoneme recognition within the domain of speech processing. Using the SpeechBox dataset provided by Northwestern University (SpeechBox[80]), which comprises recordings from 26 volunteers reading various passages, we embarked on a detailed experiment. The dataset is remarkable in its level of annotation,

providing not only timestamps for sentences and words but also for individual phonemes through specialized tools. This granularity allowed us to segment the original English recordings into distinct phoneme fragments. Subsequently, these segments were converted into spectrogram representations via customized programming pipelines.

In our experimental framework, a convolutional neural network (CNN) was trained on these spectrograms. The CNN architecture consisted of three convolutional layers with a kernel size of 3×3 and feature maps of sizes 64, 128, and 256, respectively. The training process yielded confusion matrices that characterize the performance of network on phoneme classification. Figure 4-2 displays these confusion matrices for two distinct scenarios.



(a) SpeechBox Phoneme Confusion Matrix (Stress Not Distinguished)



(b) SpeechBox Phoneme Confusion Matrix (Stress Not Distinguished and Merged Easily Confused Items)

Figure 4-2 SpeechBox Phoneme Confusion Matrices

We observe that in the confusion matrix, the column corresponding to the vowel *AH* shows a significant number of misclassifications. The reason lies in the imbalance of phoneme distributions during segmentation of speech into phonemes. The frequency of phonemes depends on the occurrence rates provided by words and sentences, and *AH* happens to be the most frequent phoneme. This results in the neural network being more inclined to predict *AH* to achieve higher accuracy. For the confusion matrix shown in Figure 4-2(b), certain phoneme categories were merged to address frequent misclassifications. The specific merges are outlined as follows.

（1）**Vowels:**

•The following groups of vowels were merged. The first element in each

group represents the merged class

$$\begin{array}{cccc} a & AA & AO & \\ a1 & AE & AH & EH \quad IH \\ ir & ER & IR & R \\ low & L & OW & \\ xy & EY & IY & Y \end{array}$$

.

（2）**Consonants:**

•Similarly, for consonants, the following groupings were merged

$$\begin{array}{ccccc} bpdt & B & P & D & T \\ kg & K & G & & \\ mn & M & N & NG & \end{array}$$

.

After merging, our experiments indicate that recognition accuracy increased modestly from approximately 81% to around 83%. Analysis of the confusion matrices further revealed that the CNN tends to perform better on consonant identification than vowels, and its ability to discriminate between voiced and voiceless consonants shared at the same articulatory position remains only moderate.

Inspired by the work presented in (Gabrielsson and Carlsson[23]), we extended our analysis by running the CNN 100 times to create an ensemble of weight vectors across the replicates. Focusing on the first and third convolutional layers, we applied de-meaning and normalization to their respective weight vectors. After a subsequent density filtering step, we visualized the resulting datasets using principal component analysis (PCA) and Vietoris–Rips complex computations. Through a rounding process of the three principal component scores for each weight vector, three representative vectors emerged

$$\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} -3 & 0 & 3 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 1 \\ -2 & -2 & -2 \\ 1 & 1 & 1 \end{bmatrix}$$

PCA revealed that, among the first three components, the first and third jointly capture features associated with horizontal line detection, whereas the second component appears to encode vertical structures. To further isolate high-contrast features, we projected all weight vectors onto these three principal directions. Specifically, using a parameter n to denote the number of farthest points considered and a parameter p (defined as the

proportion of points relative to the distance from the n-th farthest point), we extracted subsets of high-contrast points. Figure 4-3 shows the results with $p = 10\%$ and Figure 4-4 presents the outcomes for $p = 30\%$. In each figure, the left panels correspond to the PCA visualizations of the first convolutional layer, while the right panels represent those of the third layer. Moreover, the upper rows use $n = 15$ and the lower rows use $n = 300$.

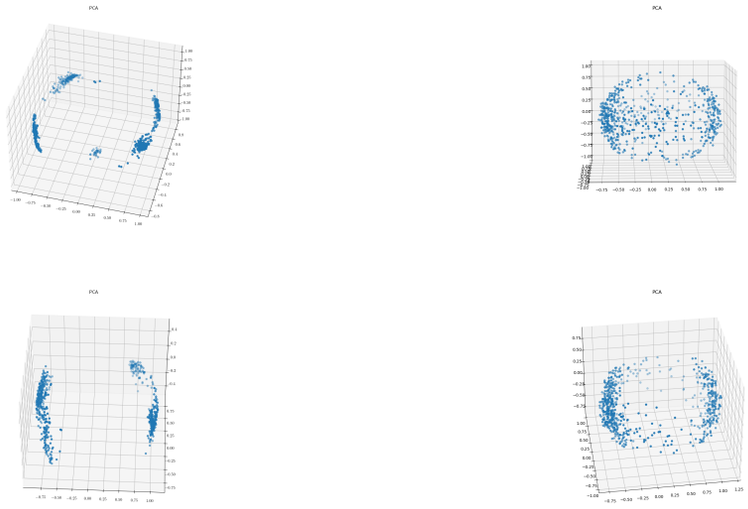


Figure 4-3 Principal Component Analysis of Weight Vectors within the Top 10% Distance

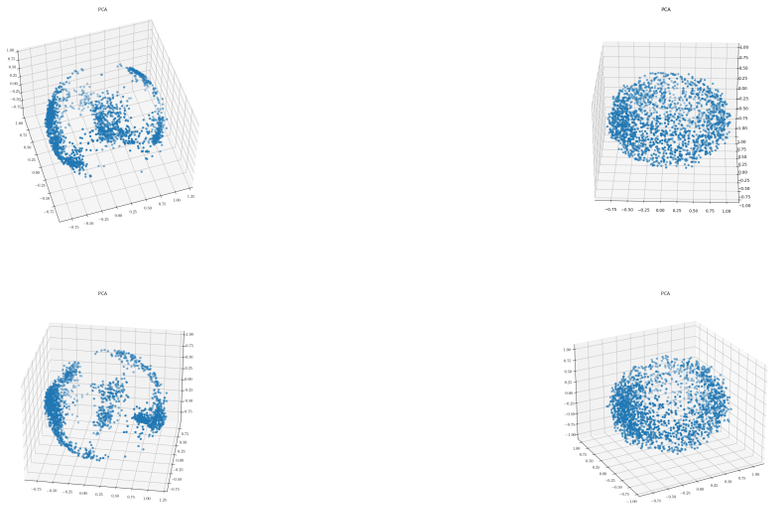


Figure 4-4 Principal Component Analysis of Weight Vectors within the Top 30% Distance

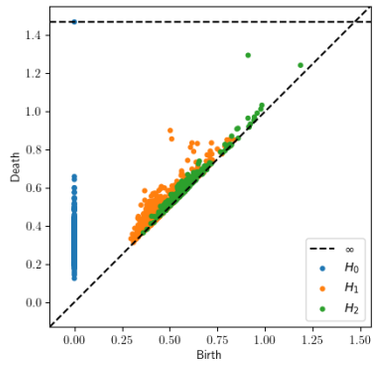
Inspection of these PCA plots reveals a gradual transition in the weight vector representations from being concentrated near two distinct points to adopting a distribution

that approximates a spherical surface. This transition is largely a consequence of the normalization applied during preprocessing. In particular, the PCA visualization of the first convolutional layer shows that the weight vectors are relatively concentrated. When considering only the top 10% of the data points based on distance, the vectors primarily cluster into two disjoint circular regions. However, when extending the selection to include the top 30%, these circular regions expand and gradually connect at the origin, forming a more unified structure. In contrast, the third convolutional layer exhibits significantly more dispersed representations, with weight vectors spreading broadly across the spherical surface. This divergence highlights the evolution of feature representation across layers, where deeper layers encode increasingly complex and diverse patterns.

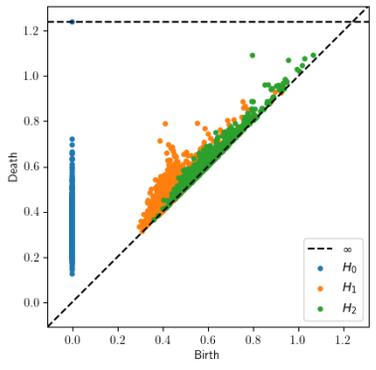
Following this geometric visualization, we computed the persistent homology of the filtered, high-contrast weight vector datasets. Due to constraints related to computational memory and complexity, we display results from only four representative scenarios in Figure 4-5. Among these, three cases exhibit a first Betti number (β_1) of 2 and a second Betti number (β_2) of 1; in the remaining scenario, both β_1 and β_2 are identified as 1. Based on these topological invariants, a toroidal structure is a plausible interpretation, especially as varying the homology coefficients produced no noticeable differences that might indicate a Klein bottle.

Finally, Figures 4-6 and 4-7 illustrate comparisons of loss and accuracy for different model variants. In these experiments, ”Normal” refers to the baseline CNN; ”Sphere” denotes a model where the initial convolution kernels are distributed on a spherical manifold; ”Torus” indicates that the kernels are selected along a torus; and ”W-Torus” represents the configuration in which kernels are chosen on a torus with an increased density along two circular trajectories. Unfortunately, based on these performance metrics, the incorporation of topological information in kernel selection does not yet lead to improvements that are as compelling as those achieved with the conventional CNN architecture.

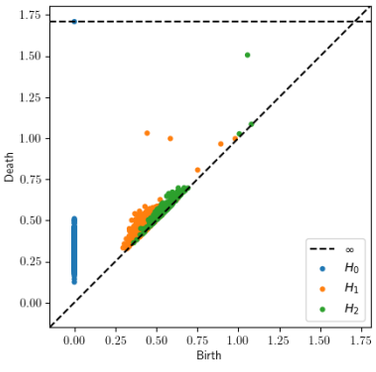
This expanded exposition not only details the experimental setup and observations but also contextualizes the topological analyses within broader challenges in speech processing. Additional inquiries might explore alternative density filtering parameters, more refined persistent homology computations, or even extended architectures that better harness topological prior information for improved performance.



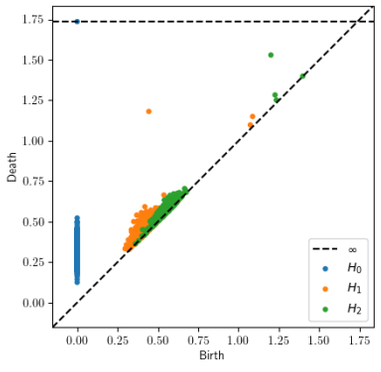
n=15, p=0.1



n=15, p=0.2



n=100, p=0.1



n=300, p=0.1

Figure 4-5 Persistent Homology of Weight Vectors

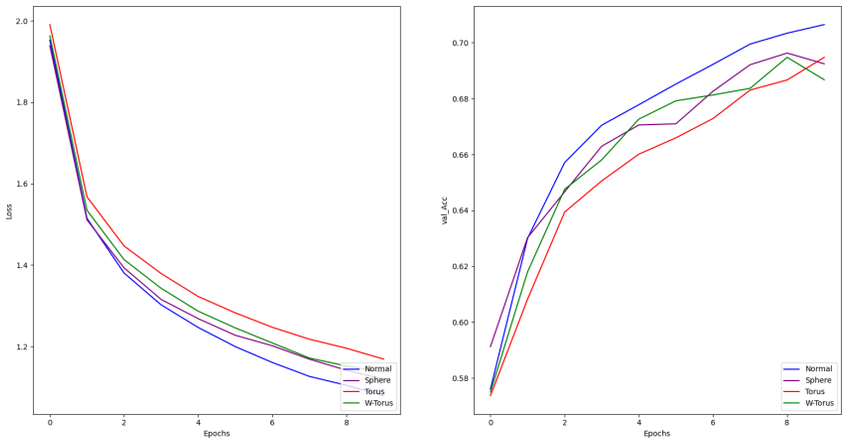


Figure 4-6 Loss and Accuracy

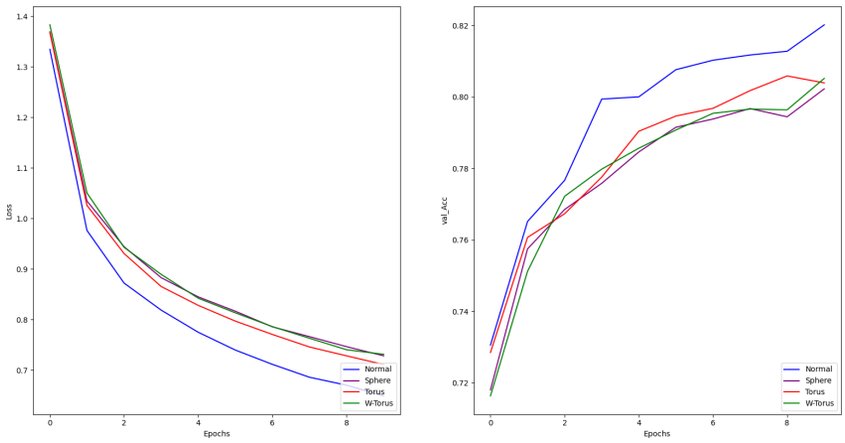


Figure 4-7 Loss and Accuracy (After Merging Confusing Items)

CHAPTER 5 THE SPACE OF SPECTROGRAM CONVOLUTION KERNELS

In this chapter, we consider the group action of the third-order special orthogonal group on the space of 3×3 real matrices. By leveraging the invariance properties of the group action, we first reduce the dimensionality of the matrix space to five. Subsequently, a new representation of the matrix space is introduced through orbit spaces and the special orthogonal group.

5.1 The Space of High-Contrast Spectrogram Convolution Kernels

Spectrograms, unlike ordinary images, lose their semantic interpretation under rotation. Thus, when considering convolution kernels for spectrograms, which we interpret as local fragments of speech where the variation is predominantly along the temporal axis, it is natural to restrict our attention to kernels that reflect this asymmetry.

Definition 5.1 (Kernel Norm): Let

$$\boldsymbol{A} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3] \in M_{3\times3}(\mathbb{R})$$

be a 3×3 convolution kernel with column vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 \in \mathbb{R}^3$. Define the norm of $\boldsymbol{A}$ by

$$||\boldsymbol{A}|| = \sqrt{||\boldsymbol{v}_1||^2 + ||\boldsymbol{v}_2||^2 + ||\boldsymbol{v}_3||^2}.$$

Note that this norm is equivalent to the standard L^2-norm up to a constant factor.

Definition 5.2 (Contrast): The **contrast** of a convolution kernel $\boldsymbol{A} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3] \in M$ is defined by

$$\mathrm{con}(\boldsymbol{A}) = \sqrt{||\boldsymbol{v}_1 - \boldsymbol{v}_2||^2 + ||\boldsymbol{v}_2 - \boldsymbol{v}_3||^2}.$$

Remark 5.1: The use of the contrast measure is motivated by the observation that spectrograms are inherently directional. Since rotation typically destroys the temporal structure of a spectrogram, a high-contrast convolution kernel (with respect to the temporal axis) is desirable for effectively capturing local speech features.

We now introduce a constrained space of convolution kernels that are both normal-

ized and optimized for high contrast.

Definition 5.3 (Normalized Convolution Kernels): We consider the subspace of $M_{3\times3}(\mathbb{R})$ consisting of convolution kernels $\boldsymbol{A} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$ satisfying the unit norm condition

$$||\boldsymbol{A}|| = 1.$$

Definition 5.4 (Contrast-Maximizing Constraint): In order to maximize contrast, we further impose the constraint that the kernel belongs to the orthogonal complement of the zero-contrast subspace. Concretely, we require

$$\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 = 0.$$

Definition 5.5 (The Kernel Space M): Let M denote the set of all 3×3 convolution kernels satisfying

$$||\boldsymbol{A}|| = 1 \quad \text{and} \quad \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 = 0.$$

Then,

$$M = \{\, \boldsymbol{A} \in M_{3\times3}(\mathbb{R}) \mid ||\boldsymbol{A}|| = 1,\ \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 = 0 \,\}.$$

Theorem 5.1: The space M is homeomorphic to the 5-dimensional sphere S^5.

Sketch of Proof: The constraints $||\boldsymbol{A}|| = 1$ and $\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 = 0$ define a smooth submanifold of $M_{3\times3}(\mathbb{R})$. One may show via dimension counting and the implicit function theorem that this submanifold has dimension $9 - 3 - 1 = 5$ (since $M_{3\times3}(\mathbb{R}) \cong \mathbb{R}^9$, and the two constraints remove 4 degrees of freedom). An explicit construction or application of known results then shows that this 5-dimensional manifold is in fact diffeomorphic to (and hence homeomorphic to) the standard sphere S^5. ■

5.2 Group Action and Quotient Space

Since $M \subset M_{3\times3}(\mathbb{R})$, the group of orthogonal transformations acts naturally on M.

Definition 5.6 (Orthogonal Group Action): Let $\theta : \mathrm{SO}(3) \times M \to M$ be defined by

$$\theta(\boldsymbol{Q}, \boldsymbol{m}) = \boldsymbol{Q}\boldsymbol{m}, \quad \text{for } \boldsymbol{Q} \in \mathrm{SO}(3) \text{ and } \boldsymbol{m} \in M.$$

Then θ is a smooth group action.

Theorem 5.2 (Contrast Projection for General Matrices): For any matrix $\boldsymbol{A} \in M_{3\times3}(\mathbb{R}^3)$ except when the three column vectors are identical, in which case the contrast

is defined as 0, such as the matrix $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$, the following procedure projects it onto the constrained subspace M:

（1）**Orthogonal Transformation**: Apply an orthogonal matrix $\boldsymbol{Q} \in \mathrm{SO}(3)$ to transform the sum of column vectors into a uniform vector:

$$\boldsymbol{Q}(\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3) = \lambda 1, \quad \lambda \in \mathbb{R},\ 1 = (1,1,1)^{\top}. \tag{5-1}$$

（2）**Centering**: Subtract the mean value from each component:

$$\tilde{\boldsymbol{A}} = \boldsymbol{Q}\boldsymbol{A} - \frac{\lambda}{3} 1 1^{\top}. \tag{5-2}$$

The resulting matrix $\tilde{\boldsymbol{A}}$ satisfies $\tilde{\boldsymbol{v}}_1 + \tilde{\boldsymbol{v}}_2 + \tilde{\boldsymbol{v}}_3 = 0$, i.e., $\tilde{\boldsymbol{A}} \in M$.

Rationale: This projection ensures:

- Invariance under orthogonal transformations: $\|\boldsymbol{Q}\boldsymbol{v}\| = \|\boldsymbol{v}\|$.
- Translation invariance: $\boldsymbol{v}_i \mapsto \boldsymbol{v}_i + \boldsymbol{c}$ cancels in (2).

The contrast $\mathrm{con}(\tilde{\boldsymbol{A}}) = \sqrt{\|\tilde{\boldsymbol{v}}_1 - \tilde{\boldsymbol{v}}_2\|^2 + \|\tilde{\boldsymbol{v}}_2 - \tilde{\boldsymbol{v}}_3\|^2}$ on M inherits these properties. ■

Note that for any $\boldsymbol{m} \in M$, and for any $\boldsymbol{Q} \in \mathrm{SO}(3)$, the group action defined above is compatible with the previously defined contrast, that is,

$$\mathrm{con}(\boldsymbol{Q}\boldsymbol{m}) = \mathrm{con}(\boldsymbol{m}).$$

Definition 5.7 (Quotient Space B): Define the homogeneous space (or orbit space)

$$B = M/\mathrm{SO}(3),$$

i.e., two kernels in M are identified if one can be obtained from the other by an orthogonal transformation.

Given coordinates derived from the columns of a kernel, let

$$x = ||\boldsymbol{v}_1||^2, \quad y = ||\boldsymbol{v}_3||^2, \quad z = \boldsymbol{v}_1 \cdot \boldsymbol{v}_3.$$

Then the constraints in M imply the following relations:

$$x + y + z = \frac{1}{2},$$
$$z^2 \le xy.$$

Proposition 5.1: The quotient space $B = M/\mathrm{SO}(3)$ is homeomorphic to the closed disk D^2.

Sketch of Proof: By introducing the coordinates (x, y, z) and considering the relations

$$x + y + z = \frac{1}{2} \quad \text{and} \quad z^2 \leq xy, \tag{5-3}$$

one can show that the set of equivalence classes is parametrized by two independent parameters satisfying inequalities that define a closed 2-dimensional disk. A more detailed study of the invariants associated with the SO(3)-action yields the claim that B is homeomorphic to D^2. ■

In particular, the boundary of B, called ∂B, corresponds to the case where the equality $z^2 = xy$ holds in the relation (5-3).

Remark 5.2: For an element $\left[\frac{1}{\sqrt{6}}\begin{bmatrix}1 & 0 & -1\\ 1 & 0 & -1\\ 1 & 0 & -1\end{bmatrix}\right] \in B$, the preimage set in M is $\left\{\begin{bmatrix}a & 0 & -a\\ b & 0 & -b\\ c & 0 & -c\end{bmatrix}\middle| a^2 + b^2 + c^2 = \frac{1}{2}\right\}$. However, the dimension of preimage is two, not three, which shows that $M \to B$ is not a fiber bundle.

Fortunately, there exists a stratified fiber bundle over B. Specifically:

• On the boundary of B, denoted ∂B, the fiber is $\mathrm{SO}(3)/\mathrm{SO}(2) \cong S^2$, quotienting out rotations around a fixed axis.

• On the region where $x = y$, the fiber is $\mathrm{SO}(3)/\mathbb{Z}_2 \cong L(4,1)$, quotienting out rotations by 180° about a fixed axis. Here $L(4,1)$ is a Lens space.

• On the intersection of the two aforementioned cases, i.e. $\boldsymbol{v}_1 + \boldsymbol{v}_3 = 0$, the fiber is given by $\mathrm{SO}(3)/(\mathrm{SO}(2) \rtimes \mathbb{Z}_2)$, which is isomorphic to $\mathbb{R}P^2$ (the real projective plane).

• On the remaining portion, the structure forms a principal SO(3)-bundle.

Proposition 5.2: For any (x, y) satisfying the relations (5-3), one corresponding kernel can be selected by $\begin{bmatrix}\boldsymbol{v}_1 & -\boldsymbol{v}_1 - \boldsymbol{v}_2 & \boldsymbol{v}_2\end{bmatrix}$, where $\boldsymbol{v}_1 = \sqrt{\frac{x}{3}}\begin{bmatrix}1\\1\\1\end{bmatrix}$ and $\boldsymbol{v}_2 = \sqrt{\frac{y}{3}}\cos\phi\begin{bmatrix}1\\1\\1\end{bmatrix} + \sqrt{\frac{y}{6}}\sin\phi\begin{bmatrix}1\\-2\\1\end{bmatrix}$, where $\sin\phi = \sqrt{1-\cos^2\phi}$ and $\cos\phi = \frac{\frac{1}{2}-x-y}{\sqrt{xy}}$ for $x, y \neq 0$. In particular, $y = \frac{1}{2}$ for $x = 0$ and $x = \frac{1}{2}$ for $y = 0$.

Remark 5.3: (x, y) satisfying the relations (5-3) if and only if

$$9\left(x + y - \frac{2}{3}\right)^2 + 3(x - y)^2 \leq 1.$$

5.3 Summary

To summarize, we have defined a notion of contrast for spectrogram convolution kernels and introduced rigid constraints (unit norm and zero-sum of column vectors) to define a space M of kernels that are well-suited for processing spectrograms. We have established that M is homeomorphic to S^5 and that the natural $\mathrm{SO}(3)$-action on M induces a quotient space B that is homeomorphic to a disk D^2. These results lay the foundation for further analysis and applications in spectrogram-based speech processing.

CHAPTER 6 NEW SPECTROGRAM CONVOLUTION FILTERS

As in Lee et al.[45], there exist 8 basic vectors in the image patch. However, up to constant factors, they will be reduced to just two, since two of them are of zero contrast, and the other of them can be reduced to two vectors through group actions.

We consider taking the orbit space of these two vectors under group actions as convolution kernels, i.e.,

$$\boldsymbol{A}_1 = \boldsymbol{Q}\begin{bmatrix}1 & 0 & -1\\ 1 & 0 & -1\\ 1 & 0 & -1\end{bmatrix}/\sqrt{6}, \text{ and } \boldsymbol{A}_2 = \boldsymbol{Q}\begin{bmatrix}1 & -2 & 1\\ 1 & -2 & 1\\ 1 & -2 & 1\end{bmatrix}/\sqrt{18}.$$

Additionally, this chapter focuses exclusively on phoneme-level recognition, as also mentioned in Chapter 4. Regarding the dataset, we cannot directly obtain phoneme-level annotations but instead employ segmentation tools. The Montreal Forced Aligner (MFA) tool from the SpeechBox dataset[80] is utilized in this study. All segmented phonemes undergo appropriate merging processes: stress variations are not differentiated and are combined, open/close vowel distinctions are eliminated, and highly similar vowel variants are merged. Notably, post-segmentation analysis revealed that certain phonemes with extremely low frequencies tend to be overlooked in prediction models, while over-represented phonemes create prediction biases. Therefore, all experiments in this chapter employ a balanced subset of 500 samples per phoneme class for classification tasks. Finally, the primary datasets used in this chapter are derived from the SpeechBox corpus, TIMIT[101] and LJSpeech[37] with specific implementation details provided in the experimental section. We selected only half of the LJSpeech dataset because the complete dataset contains a large number of speech signals, which exceeds the processing capability of my computer.

The overall procedure for all experiments in this chapter is as follows: First, segment the audio from the dataset into phonemes through **STFT spectrograms** (Short-Time Fourier Transform). Subsequently, convert the audio corresponding to each phoneme into spectrograms. These spectrograms are then fed into a convolutional neural network (CNN) for training, where the network architecture contains two convolutional layers with

64 filters each, ultimately yielding the classification accuracy.

6.1 Orthogonal Feature Layer Construction

Given two initial matrices $\{\boldsymbol{A}_1, \boldsymbol{A}_2\} \in M_{3\times3}(R)$, the layer is constructed through the following mathematical operations:

6.1.1 Matrix Augmentation

Extend the matrix set to ensure algebraic closure under inversion:

$$\mathcal{M} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, -\boldsymbol{A}_1, -\boldsymbol{A}_2\}.$$

6.1.2 SO(3)-Informed Kernel Generation

Let $\mathfrak{so}(3)$ denote the Lie algebra with basis generators:

$$\boldsymbol{L}_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \boldsymbol{L}_y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \boldsymbol{L}_z = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For stochastic kernel generation:

（1）Sample $\theta_x, \theta_y, \theta_z \sim \mathcal{N}(0, \sigma^2)$ independently.

（2）Construct Lie algebra element:

$$\boldsymbol{\theta} = \sum_{i=x,y,z} \theta_i \boldsymbol{L}_i \in \mathfrak{so}(3).$$

（3）Apply exponential map:

$$\boldsymbol{R} = \exp(\boldsymbol{\theta}) \in \mathrm{SO}(3), \quad \text{where } \exp(\boldsymbol{\theta}) = \boldsymbol{I} + \frac{\sin||\boldsymbol{\theta}||}{||\boldsymbol{\theta}||}\boldsymbol{\theta} + \frac{1-\cos||\boldsymbol{\theta}||}{||\boldsymbol{\theta}||^2}\boldsymbol{\theta}^2,$$

where $||\boldsymbol{\theta}|| = \sqrt{\sum_{i=x,y,z} \theta_i^2}$.

6.1.3 Convolutional Layer Definition

Definition 6.1 (Orthogonal Features(OF) Convolutional Layer): Given the kernel space M, each conlutional kernel of untrained **Orthogonal Features Convolutional**

Layer can be defined by

$$\boldsymbol{W}_k = \alpha \cdot \boldsymbol{R}_k \boldsymbol{M}_k, \quad \begin{cases} \boldsymbol{R}_k \in \mathrm{SO}(3), \\ \boldsymbol{M}_k \in \mathcal{M}, \\ \alpha \in \mathbb{R}^+ \text{ (adjustable scaling factor).} \end{cases}$$

6.2 Experimental Result I

In this section, we conducted experiments using the Speechbox dataset to compare the newly proposed OF convolutional layers with multiple other convolutional neural network architectures. The comparative results are presented in the accompanying Figure 6-1.

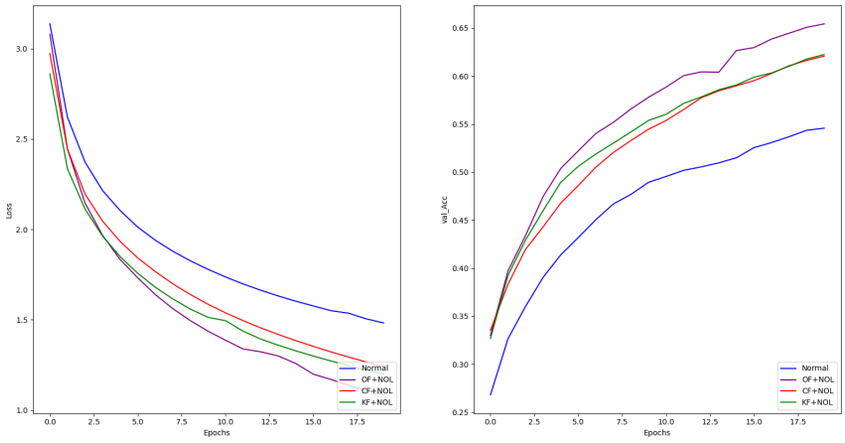


Figure 6-1 Comparisons of Loss and Accuracy on SpeechBox

The comparative analysis reveals two key observations. First, both KF and CF models demonstrate significantly superior performance in phoneme-balanced segmentation compared to traditional CNNs when evaluated against word-level phoneme frequency distributions. Second, and more critically, the proposed OF architecture exhibits marginally better effectiveness than both KF and CF configurations in these phoneme-aware classification tasks.

6.3 Experimental Result II

However, if we relax the orthogonality condition to the zero-contrast space, we can obtain a more canonical set of convolution kernels

$$\boldsymbol{Q}\begin{bmatrix}1 & 0 & -1\\ 1 & 0 & -1\\ 1 & 0 & -1\end{bmatrix}/\sqrt{6},\ \text{and}\ \boldsymbol{Q}\begin{bmatrix}1 & 0 & 1\\ 1 & 0 & 1\\ 1 & 0 & 1\end{bmatrix}/\sqrt{6}. \tag{6-1}$$

In essence, this set of convolution kernels corresponds to vertical stripe detectors with the middle column set to zero, structured as $[\boldsymbol{v}_1, 0, \pm\boldsymbol{v}_1]$, with which the sphere shares a homeomorphism. For simplicity, the neural network architectures constructed using this set of convolution kernels will retain the nomenclature OF convolutional layers.

First, let us analyze the performance of these convolutional kernel space on the Speechbox dataset (see Figure 6-2).

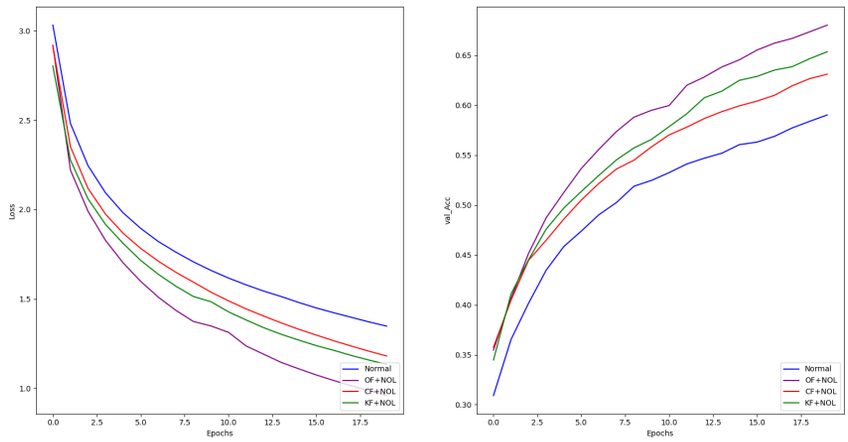


Figure 6-2 Comparisons of Loss and Accuracy on SpeechBox(Non-Orthogonal)

Here, we observe that the accuracy has approached 70%, outperforming both the previous orthogonal components and other comparative models.

Experimental results on the two additional datasets, TIMIT and LJSpeech, are also reported, yielding consistent findings (see Figure 6-3, Figure 6-4).

6.4 Noise

Analysis of the figure reveals that the datasets exhibit descending accuracy rankings: LJSpeech > SpeechBox > TIMIT, which is likely attributed to variations in acoustic clarity across the datasets. This section investigates the impact of introducing additive white

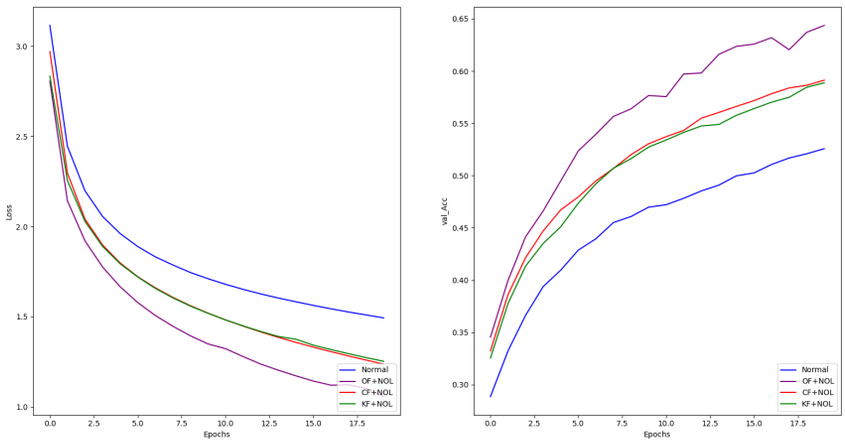


Figure 6-3 Comparisons of Loss and Accuracy on TIMIT(Non-Orthogonal)

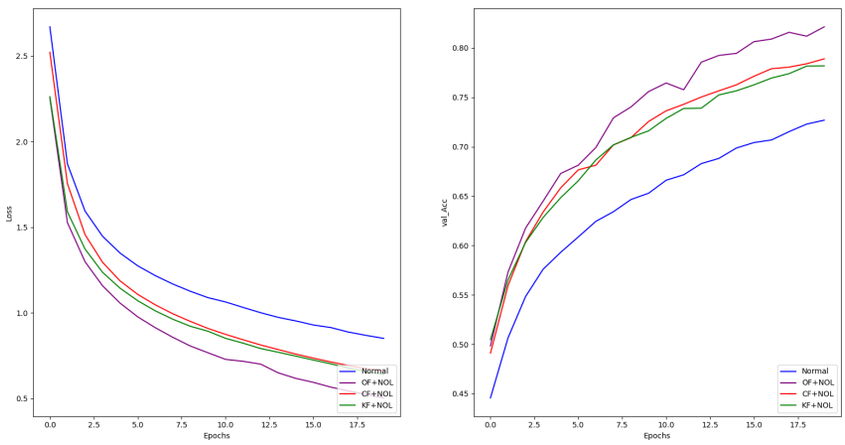


Figure 6-4 Comparisons of Loss and Accuracy on LJSpeech(Non-Orthogonal)

Gaussian noise (AWGN) on model performance.

The additive white Gaussian noise (AWGN) is systematically introduced under controlled signal-to-noise ratio (SNR) conditions, where SNR is mathematically expressed as:

$$\text{SNR (dB)} = 10\log_{10}\left(P_{\text{signal}}/P_{\text{noise}}\right)$$

with P_{signal} and P_{noise} representing the power of the original speech signal and the injected Gaussian noise, respectively. The implementation protocol comprises three phases:

（1）Data Partitioning: Split the speech corpus into training and validation subsets.

（2）Noise Injection: Apply AWGN exclusively to the training set across SNR levels ranging from **0 dB** to **20 dB**.

（3）Feature Extraction: Convert the noise-augmented training data into STFT spectrograms for downstream processing, while the validation set remains unaltered to preserve evaluation integrity.

Experimental results on the SpeechBox dataset under varying SNR conditions are as follows (see Figure 6-5, Figure 6-6).

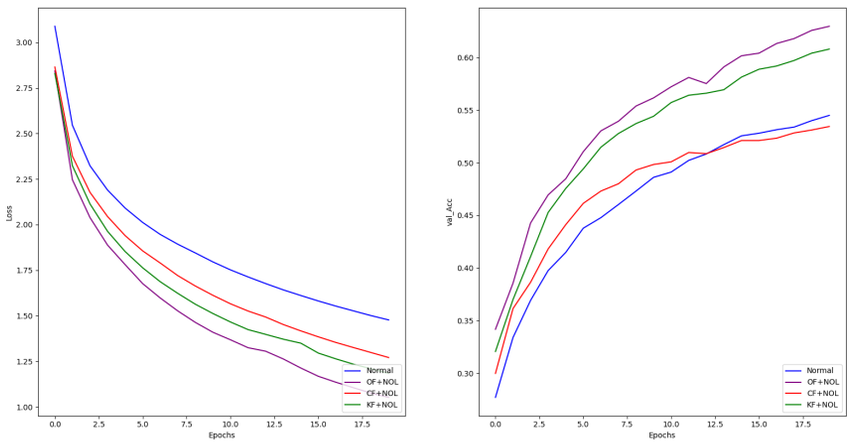


Figure 6-5 Comparisons of Loss and Accuracy on SpeechBox(SNR= 20)

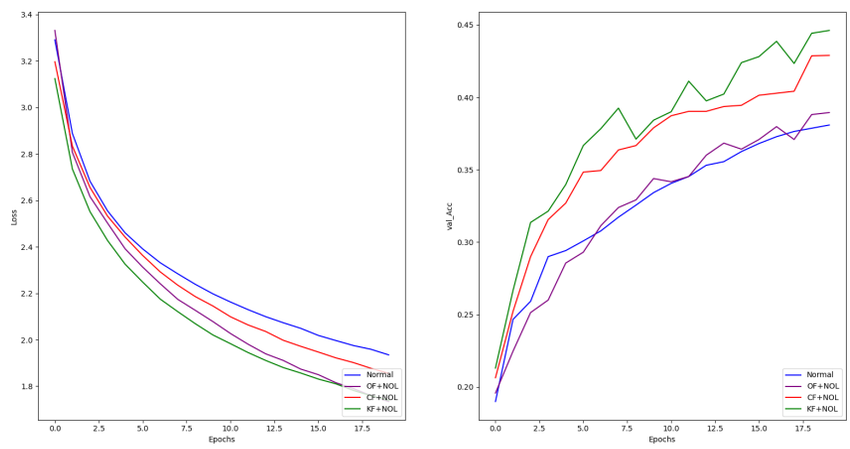


Figure 6-6 Comparisons of Loss and Accuracy on SpeechBox(SNR= 0)

The graphical comparison between the aforementioned diagrams demonstrates congruence between the SNR= 20 measurements and their noise-free counterparts. When SNR= 0, OF demonstrates moderate performance, CF exhibits inferior results, whereas KF achieves the optimal performance. This phenomenon might arise from the severe degradation of vertical stripe structures caused by additive noise, leading to reduced ac-

curacy. Consequently, in anti-noise experiments, KF manifests enhanced stability, while OF maintains superior accuracy under low-noise scenarios.

As for the convolutional kernel corresponding to this orthogonal group action, there exist multiple generation approaches, which we omit further elaboration here. In practice, our experiments with several such methods revealed accuracy rates nearly identical to those of the OF+NOL configuration across all aforementioned experimental groups.

CHAPTER 7 FURTHER APPLICATIONS AND EXTENSIONS

This chapter focuses on addressing gaps and extending prior experimental findings. It begins by supplementing earlier experiments with an analysis of scenarios where no phoneme filtering is applied, providing insights into performance under realistic conditions. Subsequently, it examines how different convolutional neural network architectures perform in word and image classification tasks, showcasing their versatility and efficiency across domains. The discussion then progresses to theoretical advancements, where the analysis of convolutional kernels is extended into the framework of Riemannian geometry, offering a novel perspective on optimization and robustness. Finally, the chapter concludes with an in-depth review of the study's limitations, acknowledging constraints in scope and methodology while outlining directions for future improvement. The convolutional neural network architecture discussed in this chapter is identical to the one in Chapter 6, consisting of two convolutional layers with 64 filters each.

7.1 Supplements on Phonemes

While previous noise robustness evaluations were conducted under phoneme-averaged conditions, an idealized scenario deviating from empirical requirements, this section implements dataset-averaged noise testing (without phoneme-level data filtering) to assess performance under more realistic conditions.

The following four figures illustrate the training performance of various neural network architectures across four datasets, SpeechBox (Figure 7-1), SpeechBox (SNR=0) (Figure 7-2), TIMIT (Figure 7-3), and LJSpeech (Figure 7-4), under conditions where no phoneme count filtering was applied.

The experimental results align with expectations in that our proposed convolutional kernel remains optimal, particularly under noise-free conditions. However, it is noteworthy that neural networks incorporating circular features and Klein features unexpectedly outperformed traditional architectures, despite prior assertions of their incompatibility with audio tasks. This apparent contradiction may stem from an overlooked preprocessing step: audio normalization was omitted in earlier implementations. Upon revisiting the

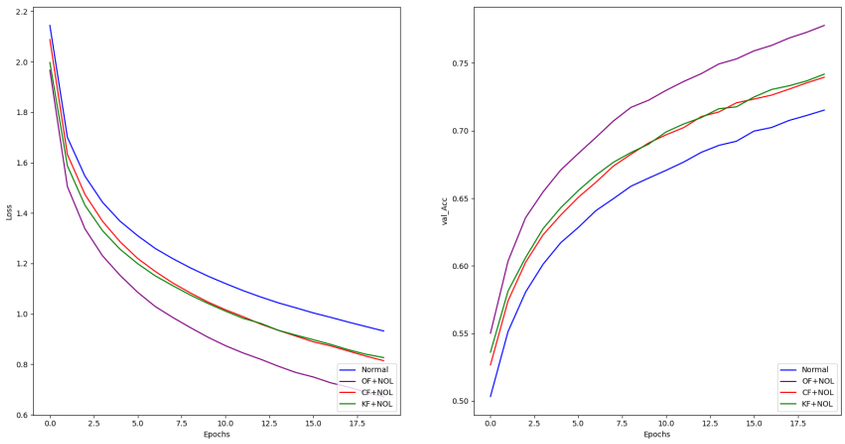


Figure 7-1 Comparisons of Loss and Accuracy on SpeechBox without Selection.

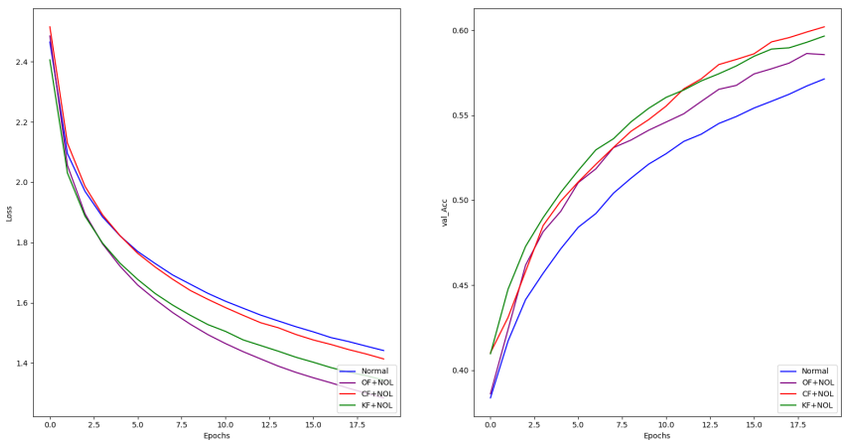


Figure 7-2 Comparisons of Loss and Accuracy on SpeechBox(SNR=0) without Selection.

codebase, we identified this omission as a plausible root cause for the previously observed accuracy degradation.

7.2 Applications to Words

Notably, the proposed convolutional layer demonstrates cross-linguistic efficacy, achieving excellent recognition accuracy not only for phoneme-level tasks but also in word-level classification. To systematically validate this capability, this section utilizes the full Speech Commands benchmark dataset[94], a dedicated word-level corpus explicitly designed with approximately balanced frequency distributions across all lexical entries, for comprehensive evaluation.

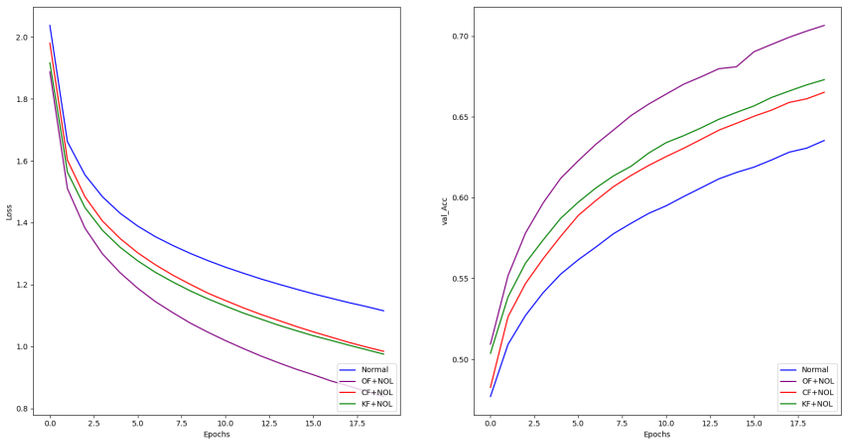


Figure 7-3 Comparisons of Loss and Accuracy on TIMIT without Selection.

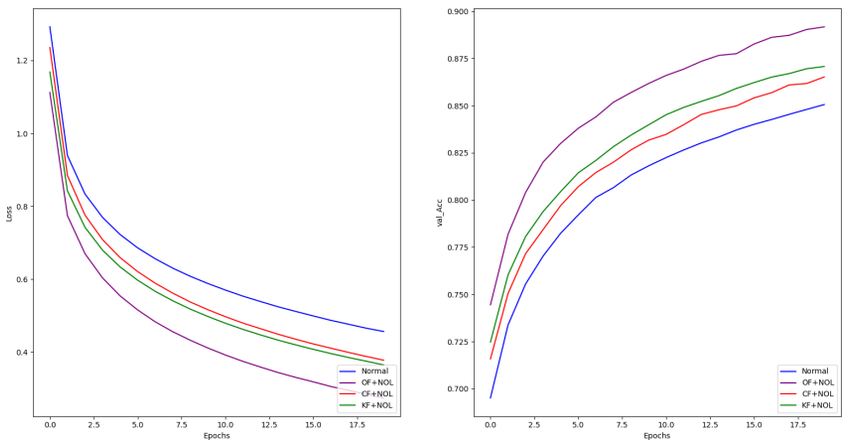


Figure 7-4 Comparisons of Loss and Accuracy on LJSpeech without Selection.

Figure 7-5 demonstrates that our neural network model exhibits robust adaptability to word-level tasks, further validating its versatility across lexical processing challenges.

7.3 Applications to Images

Applying these findings retroactively to image processing tasks demonstrates performance metrics comparable to those achieved with Klein bottle configurations, validating the cross-domain adaptability of method. We selected the CIFAR10 dataset for its higher complexity relative to MNIST, providing a more challenging benchmark to evaluate model robustness in handling intricate feature representations (see Figure 7-6).

Figure 7-6 demonstrates that our model achieves superior performance over conven-

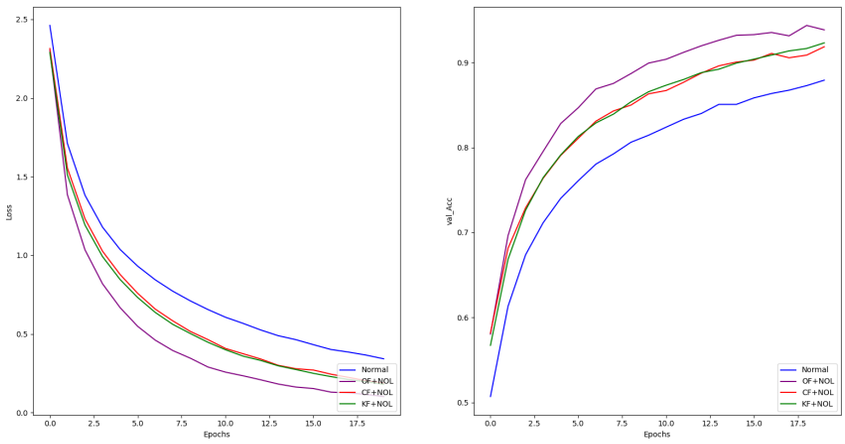


Figure 7-5 Comparisons of Loss and Accuracy on SpeechCommands.

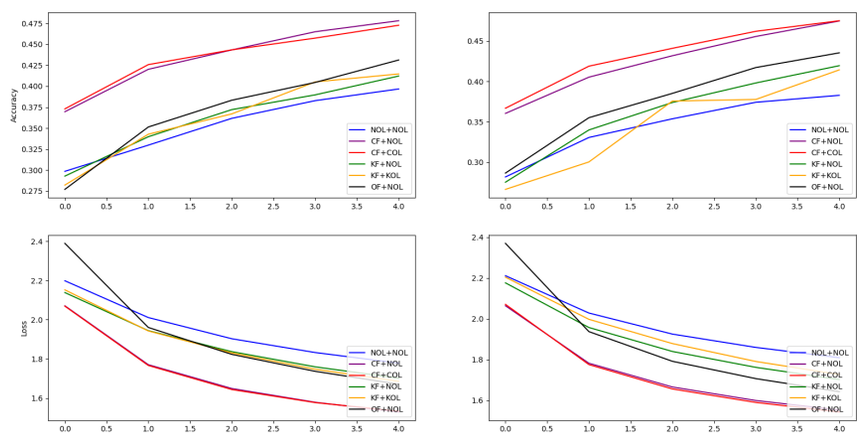


Figure 7-6 Comparisons of Loss and Accuracy on CIFAR10.

tional neural networks on image-based tasks, while maintaining parity with architectures utilizing Klein features, underscoring its cross-modal versatility.

7.4 Riemannian Geometric Theoretical Framework for Kernel Space Analysis

In Chapters 5 and 6, we examined the topological properties of the spectrogram convolution kernel space and proposed novel kernel constructions. In this section, we extend the discussion by integrating a Riemannian geometric perspective. This enriched framework not only reinforces our previous analyses but also introduces new tools to quantify kernel variations, assess noise robustness, and inspire advanced optimization strategies in

neural network architectures.

7.4.1 Differential Geometric Interpretation of Kernel Contrast

Building on the metric tensor g established in Theorem 2.9, we now provide a differential geometric perspective on the contrast dynamics in the kernel space.

Definition 7.1 (Contrast Form): For any kernel $\boldsymbol{A} \in M$, define the **contrast 1-form** $\omega_{\text{con}} \in \Omega^1(M)$ as:

$$\omega_{\text{con}}|_{\boldsymbol{A}} = \frac{1}{\|\boldsymbol{A}\|} \sum_{i=1}^{2} (\boldsymbol{v}_i - \boldsymbol{v}_{i+1}) \otimes d\boldsymbol{v}_i,$$

where $d\boldsymbol{v}_i$ denotes the exterior derivative of the column vector $\boldsymbol{v}_i$, and $\|\cdot\|$ is an appropriate norm on M.

Proposition 7.1 (O(3)-Invariance): The contrast form defined above is invariant under the action of the orthogonal group:

$$Q^* \omega_{\text{con}} = \omega_{\text{con}}, \quad \forall Q \in \text{O}(3),$$

where Q^* denotes the pullback via the group action $\theta(Q, \cdot)$.

Proof: Let $Q\boldsymbol{A} = [Q\boldsymbol{v}_1, Q\boldsymbol{v}_2, Q\boldsymbol{v}_3]$. Using the O(3)-invariance of the norm $\|\cdot\|$, we have:

$$\begin{aligned} Q^* \omega_{\text{con}}|_{\boldsymbol{A}} &= \frac{1}{\|Q\boldsymbol{A}\|} \sum_{i=1}^{2} (Q\boldsymbol{v}_i - Q\boldsymbol{v}_{i+1}) \otimes d\,(Q\boldsymbol{v}_i)\,, \\ &= \frac{1}{\|\boldsymbol{A}\|} \sum_{i=1}^{2} Q\,(\boldsymbol{v}_i - \boldsymbol{v}_{i+1}) \otimes Q\,(d\boldsymbol{v}_i)\,, \quad \text{(by group action properties)} \\ &= \frac{1}{\|\boldsymbol{A}\|} \sum_{i=1}^{2} (\boldsymbol{v}_i - \boldsymbol{v}_{i+1}) \otimes d\boldsymbol{v}_i \quad (\text{since } Q^T Q = I) \\ &= \omega_{\text{con}}|_{\boldsymbol{A}}. \end{aligned}$$

■

7.4.2 Differential Geometric Analysis of Noise Robustness

In practical scenarios, convolution kernels encounter noise perturbations that can affect performance. To incorporate noise into our framework, consider a small perturbation $\delta\boldsymbol{A}$ applied to a kernel $\boldsymbol{A} \in M$. The resulting variation in the contrast form is expressed via the Lie derivative:

$$\delta\omega_{\text{con}} = \mathcal{L}_{\delta\boldsymbol{A}} \omega_{\text{con}},$$

where $\mathcal{L}_{\delta \boldsymbol{A}}$ denotes the Lie derivative along the vector field generated by $\delta \boldsymbol{A}$. This formulation captures the sensitivity of the kernel's contrast dynamics to perturbations.

Proposition 7.2 (Noise Robustness Criterion): If for each admissible noise perturbation $\delta \boldsymbol{A}$, it holds that

$$\|\mathcal{L}_{\delta \boldsymbol{A}} \omega_{\text{con}}\| \leq \epsilon,$$

for some small constant $\epsilon > 0$, then the kernel space M demonstrates inherent noise robustness, maintaining stable contrast properties under such perturbations.

Proof: By linearizing ω_{con} around $\boldsymbol{A}$ via a Taylor expansion and invoking the O(3)-invariance, we achieve a bound on the first-order variation of the contrast form. This ensures that directional perturbations induced by $\delta \boldsymbol{A}$ remain controlled, thereby implying robustness of the kernel configuration. ■

7.4.3 Sectional Curvature and its Implications for Regularization

A key geometric quantity in assessing local stability is the sectional curvature. Given a 2-dimensional subspace of the tangent space $T_{\boldsymbol{A}}M$ spanned by vectors X, Y, the sectional curvature is defined as:

$$K(X,Y) = \frac{\langle R(X,Y)Y, X\rangle}{\|X\|^2\|Y\|^2 - \langle X,Y\rangle^2},$$

where $R(\cdot,\cdot)$ is the Riemannian curvature tensor. Regions in M with low sectional curvature indicate a locally "flat" geometry, often associated with improved stability and robustness in the optimization landscape.

Regularization via Curvature Control

High curvature regions may signal sensitive or unstable kernel configurations. Therefore, one may augment the training loss with a regularization term penalizing high curvature:

$$\mathcal{L}_{\text{reg}} = \lambda \int_M \phi(K(X,Y))\, d\mu,$$

where ϕ is an appropriate penalty function, $\lambda > 0$ a regularization parameter, and $d\mu$ the measure on M. This approach promotes smoother variations in the kernel space, aligning with the qualitative insights from Chapters 5 and 6.

7.4.4 Advanced Riemannian Optimization Perspectives

The geometric framework naturally motivates the use of Riemannian optimization techniques. In contrast to standard gradient descent, Riemannian gradient descent respects the manifold structure of M. The update rule is given by:

$$k_{n+1} = \exp_{k_n}(-\eta \operatorname{grad}_M f(k_n)),$$

where $\exp_{k_n}$ denotes the Riemannian exponential map at k_n, η is the learning rate, and $\operatorname{grad}_M f(k_n)$ represents the gradient of the objective function on the Riemannian manifold, specifically evaluated at the point k_n, as discussed in[86]. This procedure moves along geodesic paths, inherently incorporating both the metric and curvature information, and potentially yielding more stable convergence behavior.

7.4.5 Unified Theoretical Insights and Future Directions

To summarize, the integration of Riemannian geometric tools into the analysis of the kernel space achieves the following:

• **Quantitative Contrast Analysis:** The contrast form provides a differential geometric measure of kernel variation, extending the topological descriptions from Chapter 5.

• **Noise Robustness:** The Lie derivative-based noise analysis establishes rigorous criteria for the stability of kernel contrast under perturbations, reinforcing the discussions of Chapter 6.

• **Curvature-Aware Regularization:** Sectional curvature insights offer a basis for developing regularization strategies that penalize unstable, highly curved regions in the kernel space.

• **Advanced Optimization:** The use of Riemannian gradient descent and related techniques leverages the manifold structure of M, opening avenues for more efficient and stable training algorithms.

These unified insights not only enrich the theoretical framework presented in earlier chapters but also suggest several promising directions for future research:

（1）Developing hybrid topological-geometric regularizers to improve kernel stability.

（2）Empirically validating Riemannian optimization techniques in neural network training.

（3）Extending the noise analysis to encompass adversarial perturbations and more general noise models.

（4）Investigating the interplay between sectional curvature and generalization performance in deep learning.

Such endeavors could lead to a more robust, theoretically grounded approach to neural network optimization and design.

7.5 Limitations

Despite the promising results demonstrated in this dissertation, certain limitations constrain its current scope and highlight directions for further improvement. These limitations are categorized below:

• **Scope Constraint in Multimodal Analysis:** While the research framework aimed to incorporate video data for multimodal learning, technical constraints such as real-time video processing challenges and insufficient alignment between audio and visual data pipelines restricted the implementation. Addressing these challenges will require deeper exploration of synchronized audiovisual models.

• **Incomplete Spatial Characterization in Speech Processing:** Although optimized convolutional kernels demonstrated improvements for speech tasks, the framework lacks tools to fully characterize spatial dynamics in speech patterns, such as incorporating 3D vocal tract modeling or airflow dynamics. Such limitations reduce the precision of phoneme distribution mapping, particularly in high-dimensional acoustic spaces.

• **Restricted Topological Applicability in Kernel Construction:** The study primarily focused on leveraging SO(3)-informed kernels and persistence-based topological tools. However, the integration of advanced topological frameworks, such as graph persistent homology, remains underexplored. Expanding these methods could enhance kernel versatility for non-Euclidean data domains.

• **Limited Robustness Under Adversarial Noise:** While the proposed methods showed moderate resilience against white Gaussian noise, their performance under adversarial perturbations has not been systematically evaluated. This restricts the generalizability of the kernels in highly noisy environments and adversarial settings.

• **Computational Resource Dependency:** The mathematical complexity of kernel optimization and the high-dimensional manifold structures necessitate significant computational resources. This reliance may hinder scalability for large-scale applications or lower-resource settings, limiting real-world deployability.

Future work addressing these limitations will focus on expanding the current method-

ologies to encompass multimodal learning, advanced topological features, and robustness measures while improving computational efficiency and scalability.

CONCLUSION

Main Results

This dissertation integrates topological methods into neural network architectures, with a specific focus on convolutional kernels, achieving the following main results:

（1）**Reproduction of Core Experimental Results:** This study successfully replicates key experimental results from Love and Carlsson's work on MNIST and CIFAR10 datasets, validating their proposed methodologies and demonstrating their robustness. The systematic replication process not only confirms the effectiveness of topologically-informed convolutional kernels but also provides a baseline for extending these methods to new data modalities.

（2）**Exploration of Topological Characteristics in Speech Signals:** By leveraging persistent homology and principal component analysis, this research pioneers the extraction of topological structures from weight vectors in speech signal datasets. The integration of such topological insights marks an initial step towards bridging the gap between speech recognition models and topological data analysis, fostering innovative approaches for phoneme-based feature extraction.

（3）**SO(3)-Inspired Convolution Kernels:** By leveraging the group action of the special orthogonal group $SO(3)$, the study introduces a structured framework for convolutional kernels tailored to spectrogram analysis. These kernels effectively capture symmetry and hierarchical data properties, demonstrating utility across tasks like speech recognition.

（4）**Topological Representations in Neural Networks:** The work extends existing methodologies by embedding neural weight vectors into a topological framework, highlighting the interpretive potential of persistence diagrams and adjacency complexes in analyzing weight distributions.

（5）**Theoretical Basis for Manifold Analysis:** A theoretical framework is established to bridge geometric representations and spectral optimization, laying a foundation for integrating manifold-based methods into feature extraction and optimization processes.

（6）**Supplemental Exploration of Riemannian Geometry:** The supplementary

integration of Riemannian geometry tools into kernel analysis paves the way for potential advancements in geometric regularization and kernel optimization. These initial explorations suggest promising directions for future work on stability and robustness in kernel-based neural architectures.

Innovation points

（1）**Leveraging topological information for speech signal recognition:** Inspired by the successful application of convolutional kernels in image analysis by Love et al.[53], this work introduces the spectrogram as a crucial analytical tool for speech data. By treating speech data as two-dimensional spectrograms, this approach bridges methodologies from image processing to the domain of speech.

（2）**Theoretical Definition of Orthogonal Feature Layer (OF) for Speech:** Based on the unique properties of speech data, this research formulates representations for speech contrast and develops principal bundle representations of speech convolutional kernels. The resulting filters, termed **Orthogonal Feature Layer(OF)**, form a novel class of convolutional kernels designed specifically for speech data.

（3）**Higher Performance on Phoneme Data:** Neural networks constructed using OF convolutional kernels are rigorously compared to traditional neural networks and the networks proposed by Love et al. on phoneme datasets. The results indicate that OF achieves the highest accuracy under low noise conditions. However, in high noise environments, OF's performance declines, with KF (kernel filters) emerging as the superior approach.

（4）**Extension to Word and Image Data:** The applicability of OF convolutional kernels is further explored by extending their use to word datasets and image datasets. Results demonstrate consistent generalization properties, showcasing the versatility and robustness of the proposed methodology.

（5）**Theoretical Extensions to Riemannian Geometry:** To enhance the theoretical foundations, this research attempts to generalize the convolutional kernel theory within the framework of Riemannian geometry. This extension provides deeper insights and opens avenues for further exploration and application.

Future Work

Building on the contributions of this dissertation, future research could explore the following directions:

（1）**Multimodal Learning:** Investigate the application of topological methods in multimodal tasks, such as synchronized audio-visual recognition or sensor data integration, aiming to evaluate the adaptability of topology-enhanced architectures.

（2）**Advanced Adversarial Robustness:** Explore the role of topological kernels in defending against adversarial attacks, focusing on their capacity to preserve model integrity under perturbations.

（3）**Extended Topological Applications:** Extend the use of persistent homology and other invariants to non-Euclidean data, such as graph and point cloud structures, further validating their versatility.

（4）**Optimization in High-Dimensional Spaces:** Employ advanced Riemannian optimization strategies to refine kernel parameterizations, leveraging geometric constraints to improve convergence and generalization.

（5）**Topological-Geometric Feature Extraction:** Combine persistent homology and geometric data analysis more systematically to uncover latent features across diverse data types (e.g., dynamic networks, manifolds), adapting methodologies to their intrinsic topological and geometric properties. In fact, in other work by our research group[20], the results of persistent homology have been used as topological features input into machine learning and neural networks.

REFERENCES

[1] ABDEL-HAMID O, MOHAMED A R, JIANG H, et al. Convolutional Neural Networks for Speech Recognition[C]//IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014: 1533-1545.

[2] ADAMS H, EMERSON T, KIRBY M, et al. Persistence images: A stable vector representation of persistent homology[J]. Journal of Machine Learning Research, 2017, 18(8): 1-35.

[3] AMODEI D, ANANTHANARAYANAN S, ANUBHAI R, et al. Deep speech 2: End-to-end speech recognition in english and mandarin[C]//International conference on machine learning. 2016: 173-182.

[4] BAAS N A, CARLSSON G E, QUICK G, et al. Topological data analysis[M]. Springer, 2020.

[5] BOTTOU L, CORTES C, DENKER J S, et al. Comparison of classifier methods: a case study in handwritten digit recognition[C]//Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5): vol. 2. 1994: 77-82.

[6] BROWN K A, KNUDSON K P. Nonlinear statistics of human speech data[J]. International Journal of Bifurcation and Chaos, 2009, 19(07): 2307-2319.

[7] CARLSSON G. Topology and data[J]. Bulletin of the American Mathematical Society, 2009, 46(2): 255-308.

[8] CARLSSON G, GABRIELSSON R B. Topological Approaches to Deep Learning[C]// BAAS N A, CARLSSON G E, QUICK G, et al. Topological Data Analysis. Cham: Springer International Publishing, 2020: 119-146.

[9] CARLSSON G, ISHKHANOV T, DE SILVA V, et al. On the local behavior of spaces of natural images[J]. International journal of computer vision, 2008, 76: 1-12.

[10] CARRIÈRE M, RABADÁN R. Topological Data Analysis of Single-Cell Hi-C Contact Maps [C]//BAAS N A, CARLSSON G E, QUICK G, et al. Topological Data Analysis. Cham: Springer International Publishing, 2020: 147-162.

[11] CHAZAL F, SILVA V, GLISSE M, et al. The Structure and Stability of Persistence Modules [M]. 2012.

[12] COHEN T, WELLING M. Group equivariant convolutional networks[C]//International conference on machine learning. 2016: 2990-2999.

[13] COHEN-STEINER D, EDELSBRUNNER H, HARER J. Stability of persistence diagrams [C]//Proceedings of the twenty-first annual symposium on Computational geometry. 2005: 263-271.

[14] CURRY J M. Sheaves, cosheaves and applications[M]. University of Pennsylvania, 2014.

[15] DE SILVA V, GHRIST R. Homological Sensor Networks[J]. Notices of the AmericanMathematical Society, 2007, 54(1).

[16] DINDIN M, UMEDA Y, CHAZAL F. Topological Data Analysis for Arrhythmia Detection through Modular Neural Networks[C]//GOUTTE C, ZHU X. Advances in Artificial Intelligence. Cham: Springer International Publishing, 2020: 177-188.

[17] EDELSBRUNNER H, HARER J, et al. Persistent homology-a survey[J]. Contemporary mathematics, 2008, 453(26): 257-282.

[18] EDELSBRUNNER H, MUCKE E P. Three-Dimensional Alpha Shapes[J]. ACM Transactions on Graphics, 13:1, 1994: 43-72.

[19] Edelsbrunner, Letscher, Zomorodian. Topological persistence and simplification[J]. Discrete & computational geometry, 2002, 28: 511-533.

[20] FENG P, QU Q, ZHANG H, et al. Topology-enhanced machine learning for consonant recognition[Z]. Preprint. 2025.

[21] FORMAN R. A user's guide to discrete morse theory.[J]. Séminaire Lotharingien de Combinatoire [electronic only], 2002, 48: B48c-35.

[22] FUKUSHIMA K. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

[23] GABRIELSSON R B, CARLSSON G. Exposition and interpretation of the topology of neural networks[C]//2019 18th ieee international conference on machine learning and applications (icmla). 2019: 1069-1076.

[24] GAKHAR H, PEREA J A. Sliding window persistence of quasiperiodic functions[J]. Journal of Applied and Computational Topology, 2024, 8(1): 55-92.

[25] GEIRHOS R, RUBISCH P, MICHAELIS C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[C]//International conference on learning representations. 2018.

[26] GIUSTI C, PASTALKOVA E, CURTO C, et al. Clique topology reveals intrinsic geometric structure in neural correlations[J]. Proceedings of the National Academy of Sciences, 2015, 112(44): 13455-13460.

[27] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as Space-Time Shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253.

[28] GRIGOR'YAN A, LIN Y, MURANOV Y, et al. Homologies of path complexes and digraphs [J]. arXiv preprint arXiv:1207.2834, 2012.

[29] GUO Y, LIU Y, OERLEMANS A, et al. Deep learning for visual understanding: A review[J]. Neurocomputing, 2016, 187: 27-48.

[30] GYULASSY A G. Combinatorial construction of Morse-Smale complexes for data analysis and visualization[M]. University of California, Davis, 2008.

[31] HANSEN J, GEBHART T. Sheaf neural networks[J]. arXiv preprint arXiv:2012.06333, 2020.

[32] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[33] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[34] HOFER C, KWITT R, NIETHAMMER M, et al. Deep learning with topological signatures [J]. Advances in neural information processing systems, 2017, 30.

[35] HORAK D, MALETIĆ S, RAJKOVIĆ M. Persistent homology of complex networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009(03): P03034.

[36] HUBEL , WIESEL . Receptive Fields and Functional Architecture of Monkey Striate Cortex [J]. The Journal of physiology, 1968, 195(1): 215-243.

[37] ITO K, JOHNSON L. The LJ Speech Dataset[Z]. https://keithito.com/LJ-Speech-Dataset/. 2017.

[38] KENNEL M B, BROWN R, ABARBANEL H D I. Determining embedding dimension for phase-space reconstruction using a geometrical construction[J]. Physical Review A, 1992, 45(6): 3403-3411.

[39] KHASAWNEH F A, MUNCH E. Chatter detection in turning using persistent homology[J]. Mechanical Systems and Signal Processing, 2016, 70: 527-541.

[40] KNUDSON K, WANG B. Discrete Stratified Morse Theory: Algorithms and A User's Guide [J]. Discrete & Computational Geometry, 2022, 67(4): 1023-1052.

[41] KRAMÁR M, GOULLET A, KONDIC L, et al. Persistence of force networks in compressed granular media[J]. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 2013, 87(4): 042207.

[42] KRIZHEVSKY A, HINTON G, et al. Learning multiple layers of features from tiny images [J]. 2009.

[43] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[44] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[45] LEE A B, PEDERSEN K S, MUMFORD D. The nonlinear statistics of high-contrast patches in natural images[J]. International Journal of Computer Vision, 2003, 54: 83-103.

[46] LEE J M, LEE J M. Smooth manifolds[M]. Springer, 2003.

[47] LEE Y, BARTHEL S D, DŁOTKO P, et al. Quantifying Similarity of Pore-Geometry in Nanoporous Materials[J]. Nature Communications, 2017, 8(1): 15396.

[48] LI J. Recent Advances in End-to-End Automatic Speech Recognition[J]. APSIPA Transactions on Signal and Information Processing, 2022, 11(1).

[49] LI M Z, RYERSON M S, BALAKRISHNAN H. Topological Data Analysis for Aviation Applications[J]. Transportation Research Part E: Logistics and Transportation Review, 2019, 128: 149-174.

[50] LIU J Y, JENG S K, YANG Y H. Applying topological persistence in convolutional neural network for music audio signals[J]. arXiv preprint arXiv:1608.07373, 2016.

[51] LIU X, FENG H, WU J, et al. Persistent Spectral Hypergraph Based Machine Learning (PSH-ML) for Protein-Ligand Binding Affinity Prediction[J]. Briefings in Bioinformatics, 2021, 22(5): bbab127.

[52] LIU X, WANG X, WU J, et al. Hypergraph-Based Persistent Cohomology (HPC) for Molecular Representations in Drug Design[J]. Briefings in Bioinformatics, 2021, 22(5): bbaa411.

[53] LOVE E R, FILIPPENKO B, MAROULAS V, et al. Topological convolutional layers for deep learning[J]. Journal of Machine Learning Research, 2023, 24(59): 1-35.

[54] NAKAMURA T, HIRAOKA Y, HIRATA A, et al. Persistent homology and many-body atomic structure for medium-range order in the glass[J]. Nanotechnology, 2015, 26(30): 304001.

[55] OPPENHEIM A V. Discrete-time signal processing[M]. Pearson Education India, 1999.

[56] OUDOT S Y. Persistence theory: from quiver representations to data analysis: vol. 209[M]. American Mathematical Society Providence, 2015.

[57] PELLIOTT J, DAWBER P. Symmetry in Physics I, Principles and Simple Applications[Z]. 1990.

[58] PEREA J A, DECKARD A, HAASE S B, et al. SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data[J]. BMC bioinformatics, 2015, 16: 1-12.

[59] PEREA J A, HARER J. Sliding windows and persistence: An application of topological methods to signal analysis[J]. Foundations of computational mathematics, 2015, 15: 799-838.

[60] PIKE J A, KHAN A O, PALLINI C, et al. Topological data analysis quantifies biological nano-structure from single molecule localization microscopy[J]. Bioinformatics, 2020, 36(5): 1614-1621.

[61] QAISER T, TSANG Y W, TANIYAMA D, et al. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features[J]. Medical image analysis, 2019, 55: 1-14.

[62] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[63] RAWAT W, WANG Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review[J]. Neural Computation, 2017, 29(9): 2352-2449.

[64] RIECK B, YATES T, BOCK C, et al. Uncovering the Topology of Time-Varying fMRI Data Using Cubical Persistence[J]. Advances in Neural Information Processing Systems 33, 2020.

[65] ROBINSON M. Topological signal processing: vol. 81[M]. Springer, 2014.

[66] ROTE G, VEGTER G. Computational topology: An introduction[G]//Effective Computational Geometry for curves and surfaces. Springer, 2006: 277-312.

[67] RUMELHART D E, MCCLELLAND J L. Learning Internal Representations by Error Propagation[M]//Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. 1987: 318-362.

[68] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[69] SAINATH T N, KINGSBURY B, MOHAMED A R, et al. Improvements to deep convolutional neural networks for LVCSR[C]//2013 IEEE workshop on automatic speech recognition and understanding. 2013: 315-320.

[70] SAINATH T N, VINYALS O, SENIOR A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2015: 4580-4584.

[71] SALAS L S. The Three Gap Theorem in Persistent Homology[D]. Michigan State University, 2024.

[72] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing Human Actions: A Local SVM Approach[C]//Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Cambridge, UK: IEEE, 2004: 32-36 Vol.3.

[73] SEVERSKY L M, DAVIS S, BERGER M. On time-series topological data analysis: New data and opportunities[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2016: 59-67.

[74] SHEN L, FENG H, LI F, et al. Knot data analysis using multiscale Gauss link integral[J]. Proceedings of the National Academy of Sciences, 2024, 121(42): e2408431121.

[75] SILVA V D, CARLSSON G. Topological estimation using witness complexes[C]//GROSS M, PFISTER H, ALEXA M, et al. SPBG'04 Symposium on Point - Based Graphics 2004. The Eurographics Association, 2004.

[76] SIZEMORE A E, GIUSTI C, KAHN A, et al. Cliques and Cavities in the Human Connectome [J]. Journal of Computational Neuroscience, 2018, 44(1): 115-145.

[77] SKRABA P, OVSJANIKOV M, CHAZAL F, et al. Persistence-based segmentation of deformable shapes[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. 2010: 45-52.

[78] SMITH A D, DŁOTKO P, ZAVALA V M. Topological data analysis: concepts, computation, and applications in chemical engineering[J]. Computers & Chemical Engineering, 2021, 146: 107202.

[79] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[80] SpeechBox[Z]. Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/allsstar.

[81] SULTANA F, SUFIAN A, DUTTA P. A Review of Object Detection Models Based on Convolutional Neural Network[G]//MANDAL J K, BANERJEE S. Intelligent Computing: Image Processing Based Applications. Singapore: Springer Singapore, 2020: 1-16.

[82] SULTANA F, SUFIAN A, DUTTA P. Advancements in image classification using convolutional neural network[C]//2018 Fourth international conference on research in computational intelligence and communication networks (ICRCICN). 2018: 122-129.

[83] SUN S, ZHANG B, XIE L, et al. An unsupervised deep domain adaptation approach for robust speech recognition[J]. Neurocomputing, 2017, 257: 79-87.

[84] TAKENS F. Detecting Strange Attractors in Turbulence[C]//RAND D, YOUNG L S. Dynamical Systems and Turbulence, Warwick 1980. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981: 366-381.

[85] TAN X, CHEN J, LIU H, et al. *NaturalSpeech* : End-to-End Text-to-Speech Synthesis With Human-Level Quality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4234-4245.

[86] TANG F, FENG H, TINO P, et al. Probabilistic learning vector quantization on manifold of symmetric positive definite matrices[J]. Neural Networks, 2021, 142: 105-118.

[87] TINARRAGE R. Computing persistent Stiefel–Whitney classes of line bundles[J]. Journal of Applied and Computational Topology, 2022, 6(1): 65-125.

[88] TRALIE C J. Early MFCC and HPCP fusion for robust cover song identification[J]. arXiv preprint arXiv:1707.04680, 2017.

[89] TRALIE C J, PEREA J A. (Quasi) periodicity quantification in video data, using topology[J]. SIAM Journal on Imaging Sciences, 2018, 11(2): 1049-1077.

[90] TURNER K, MUKHERJEE S, BOYER D M. Persistent Homology Transform for Modeling Shapes and Surfaces[J]. Information and Inference, 2014, 3(4): 310-344.

[91] UMEDA Y. Time Series Classification via Topological Data Analysis[J]. Transactions of the Japanese Society for Artificial Intelligence, 2017, 32(3): 1-12.

[92] VAN HATEREN J H, VAN DER SCHAAF A. Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex[J]. Proceedings of the Royal Society of London. Series B: Biological Sciences, 1998, 265(1394): 359-366.

[93] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[94] WARDEN P. Speech commands: A dataset for limited-vocabulary speech recognition[J]. arXiv preprint arXiv:1804.03209, 2018.

[95] YAO Y, SUN J, HUANG X, et al. Topological methods for exploring low-density states in biomolecular folding pathways[J]. The Journal of chemical physics, 2009, 130(14).

[96] ZAITOUN N M, AQEL M J. Survey on Image Segmentation Techniques[J]. Procedia Computer Science, 2015, 65: 797-806.

[97] ZHANG H, LIN X, WEI Y, et al. Validation of deep learning-based DFCNN in extremely large-scale virtual screening and application in trypsin I protease inhibitor discovery[J]. Frontiers in molecular biosciences, 2022, 9: 872086.

[98] ZHENG Q, YANG M, YANG J, et al. Improvement of Generalization Ability of Deep CNN via Implicit Regularization in Two-Stage Training Process[J]. IEEE Access, 2018, 6: 15844-15869.

[99] ZHENG R C, AI Y, LING Z H. Incorporating Ultrasound Tongue Images for Audio-Visual Speech Enhancement[J]. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 2024, 32: 1430-1444.

[100] ZOMORODIAN A, CARLSSON G. Computing Persistent Homology[J]. Discrete Comput. Geom, 2005, 33: 249-274.

[101] ZUE V, SENEFF S, GLASS J. Speech database development at MIT: TIMIT and beyond[J]. Speech communication, 1990, 9(4): 351-356.

[102] 卓沛生, 何梓彤, 蔺宏伟. PHTNet: A Shape Recognition Network Based on Multi-Perspective Topological Features[J]. Journal of Computer-Aided Design & Computer Graphics, 2024.

ACKNOWLEDGEMENTS

I am deeply grateful to my mentors and colleagues for their support in completing this research.

First and foremost, I sincerely thank Prof. Fang Fuquan for his guidance. He introduced me to topological data analysis and encouraged me to pursue this research direction. His suggestions on applying topological methods to data science problems were particularly insightful.

I am especially grateful to Prof. Zhu Yifei for his patient supervision. His advice on topological convolutional neural networks (TCNN), especially regarding algorithm optimization, helped me overcome many technical challenges. His encouragement during difficult moments was invaluable.

I thank all members of Prof. Zhu Yifei's research group for their support. Special thanks to Mr. Qu Qingrui for his coding assistance. The discussions with office members, whether about mathematics or research challenges, were very helpful.

Finally, I thank all who supported me during this research.

RESUME AND ACADEMIC ACHIEVEMENTS

Resume

The auther was born in July 1996, in Chaoyang, Liaoning, China.

In September 2014, he was admitted to Tianjin University(TJU). In July 2018, he obtained a bachelor's degree in science from the School of Mathematics, TJU.

In September 2018, he began his graduate study in the School of Mathematics, TJU, and got a master of science degree in Mathematics, in June 2021.

Since August 2021, he has started to pursue his doctor's degree of science in the Department of Mathematics in the College of Science, Southern University of Science and Technology(SUSTech).

Academic Achievements during the Study for an Academic Degree

[1] Pingyao Feng, Siheng Yi, Qingrui Qu, Zhiwang Yu, Yifei Zhu. Topology combined machine learning for consonant recognition. arXiv:2311.15210 [cs.LG], 2023.

[2] Zhiwang Yu. Topological Deep Learning for Speech Data. arXiv:2505.21173[cs.ML], 2025.