Topological Deep Learning for Speech Recognition

Zhiwang Yu[1,†] Pingyao Feng[2] Qingrui Qu[1] Haiyu Zhang[1] Yifei Zhu[1,*]

ABSTRACT

Topological data analysis (TDA) offers mathematical tools for deep learning and insights for its interpretability. Inspired by Carlsson and his collaborators' seminal work on topological convolutional neural networks with image and video data, in this study we design topology-aware convolution kernels which significantly improve speech recognition networks. Theoretically, by investigating orthogonal group actions on kernels, we establish a fiber-bundle decomposition of matrix spaces, enabling new filter generation methods. In practice, our proposed Orthogonal Filters layer achieves superior performance in phoneme recognition, particularly in low-noise scenarios, while demonstrating cross-domain adaptability for other audio and visual recognition tasks. This work reveals TDA's potential in neural network optimization, opening new avenues for interdisciplinary studies with topological methods and machine learning.

Index Terms: Topological data analysis, convolutional neural network, speech recognition, group action, orthogonal filters.

1 INTRODUCTION

1.1 Background

This study aims to integrate topological data analysis with deep neural networks, focusing on the application of topological convolution kernels in speech recognition tasks, particularly for phoneme identification and word classification. By incorporating topological feature extraction, we seek to enhance a network's ability to capture key topological characteristics in speech signals, thereby improving recognition performance.

The development of deep neural networks has undergone several critical phases. Early fully connected networks were constrained by computational limitations and theoretical understanding until the emergence of convolutional neural networks (CNNs), which marked the golden age of deep learning. CNNs significantly reduced parameter complexity through local connectivity and weight sharing while preserving essential spatial hierarchical features. However, traditional CNNs exhibit inherent limitations in comprehending the global topological properties of data. In this context, topological neural networks emerged as a promising solution. In 2004, de Silva and Carlsson identified a topological structure of three rings in image data using persistent homology [13]. Based on this, in [4], together with Ishkhanov and Zomorodian they further detected a distribution of such data as over the Klein bottle, a complex topological manifold. Around the same time, Carlsson proposed a theoretical framework explaining the significance of topological features in data analysis [2]. A decade later, since around 2018, Carlsson and his collaborators have extended topological analysis to the study of convolutional neural network weight distributions [3] and demonstrated significantly improved performance by directly encoding topological features in the design of convolution kernels [9]. These groundbreaking discoveries provided a starting point for our research, particularly in the integration of topological feature extraction methods into speech-specific convolution kernel designs. Of course, besides architectures, there are other aspects of topological deep learning, such as representations, which go beyond the scope of this article but serve as an overarching context of research and applications (see, e.g., [16]).

Traditional speech processing methods often overlook the rich topological structures embedded in speech signals. The complex patterns exhibited by speech signals in the time–frequency domain contain topological features that play a vital role in phoneme discrimination and word recognition [5, 14]. Using TDA tools, we can capture these structural characteristics more effectively. Notably, the local extrema and connectivity relationships formed by speech signals in a mel spectrogram constitute specific topological configurations, which exhibit systematic differences across phonemes. The topological convolution kernel proposed in this study is specifically designed to model these features, enabling simultaneous extraction of conventional spectral features while explicitly representing the topological properties of speech signals. This approach not only improves recognition accuracy, but also enhances the model's robustness to noise and variations, offering a new technical pathway for speech recognition systems in complex environments.

1.2 Statement of Results

The main focus of this article is the application of convolution kernels, constructed using the newly defined Orthogonal Filters (OF) layer, to phoneme recognition. Moreover, the approach is generalized to word recognition and image recognition, demonstrating its versatility and adaptability across multiple domains.

Firstly, in Sec. 3, geared towards speech data instead of image data, we consider the matrix space $M_{3\times3}(\mathbb{R})$, which is the most common space of convolution kernels, as $\{[\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3] \mid \boldsymbol{v}_i \in \mathbb{R}^3\}$. Without loss of generality, define the subspace $M = \{[\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3] \mid \|\boldsymbol{v}_1\|^2+\|\boldsymbol{v}_2\|^2+\|\boldsymbol{v}_3\|^2 = 1, \boldsymbol{v}_1+\boldsymbol{v}_2+\boldsymbol{v}_3 = \boldsymbol{0}\}$. Next, we define a group action on M by

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

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, π has the structure of a stratified fiber bundle. The fiber is $\mathrm{SO}(3)/(\mathrm{SO}(2)\rtimes\mathbb{Z}_2) \cong \mathbb{RP}^2$ when $\boldsymbol{v}_1+\boldsymbol{v}_3 = \boldsymbol{0}$, $\mathrm{SO}(3)/\mathrm{SO}(2) \cong S^2$ when $\boldsymbol{v}_1$ and $\boldsymbol{v}_3$ are collinear, $\mathrm{SO}(3)/\mathbb{Z}_2 \cong L(4,1)$ when $\boldsymbol{v}_1$ and $\boldsymbol{v}_3$ are of equal magnitudes, and $\mathrm{SO}(3)$ otherwise. The above provides a representation of M by special orthogonal group action.

Secondly, in Sec. 4, we define our OF layer by selecting elements in B and $\mathrm{SO}(3)$. Then, we compare neural networks constructed using OF convolution kernels to traditional neural networks and the networks proposed by Love et al. [9] 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 (Klein Filters) emerging as a superior approach.

Finally, in Sec. 5, the applicability of OF convolution kernels is further explored by extending their use to word datasets and image datasets. Results demonstrate consistent generalization properties that showcase the versatility and robustness of the proposed methodology.

In this article, we treat weight vectors and convolution kernels interchangeably, without differentiating the two concepts.

[1]Department of Mathematics, Southern University of Science and Technology

[†]Corresponding author. Email: 12131239@mail.sustech.edu.cn

[2]Department of Mathematics, North Carolina State University

[*]Corresponding author. Email: zhuyf@sustech.edu.cn

Updated: June 30, 2025

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, the source codes of all experiments are available at https://github.com/ZhiwangYu/TDLforSpeechRecognition.

1.3 Outline

This article integrates theory, methodology, and applications of topological deep learning through 5 systematically organized sections. Sec. 2 introduces the fundamental objects of study, CNNs, topological CNNs, and phonemes. Sec. 3 formalizes the spectrogram convolution kernel space through geometric and topological constraints, including contrast maximization and group actions, providing a structured framework for speech processing. This leads to Sec. 4, where we introduce novel kernel designs, demonstrating improved phoneme recognition accuracy and robustness against noise. Sec. 5 discusses broader implications, including model performance on unfiltered phonemes, as well as generalization to word and image tasks.

2 Topological Convolutional Neural Networks and Phonetic Data

2.1 Convolutional Neural Networks

Definition 2.1 (Algebraic Formalism of Convolutional Neural Networks)**.** *A* convolutional neural network *is a feedforward system* $\mathcal{N} = (V, E, \Lambda)$ *where*

1. *the vertex set* $V = \bigsqcup_{k=0}^{L} V_k$ *decomposes into layers with* V_0 *representing the input and* V_L *representing the output,*
2. *the directed edges* $E \subset \bigcup_{k=0}^{L-1}(V_k \times V_{k+1})$ *respect layer ordering, and*
3. *the weight parameters* $\Lambda = \{\lambda_e \in \mathbb{R}\}_{e\in E}$ *exhibit translational symmetry.*

The CNN dynamics are governed by the following 2 fundamental constraints.

- **Spatial Locality**: For convolutional layers $V_k = \chi_k \times \mathbb{Z}^d$, each $(v,w) = \big((\kappa,\boldsymbol{x}),(\kappa',\boldsymbol{x}')\big) \in E$ satisfies $\|\boldsymbol{x}-\boldsymbol{x}'|rVert_\infty \le r_k$ for some receptive field radius r_k.
- **Parameter Sharing**: Weight values $\lambda_{(\kappa,\boldsymbol{x}),(\kappa',\boldsymbol{x}')}$ depend solely on κ, κ' and the displacement $\boldsymbol{x}-\boldsymbol{x}'$.

Definition 2.2 (Forward Propagation)**.** *The activation* a_w *at node* $w \in V_{k+1}$ *is computed as*

$$a_w = \sigma\left(\sum_{\substack{v\in V_k \\ (v,w)\in E}} \lambda_{(v,w)} a_v + b_w\right)$$

where σ *denotes the ReLU activation and* b_w *denotes the bias term.*

CNNs implement multiscale processing through the following interleaved operations.

- **Convolutional Blocks**: Combine spatial filtering (via learned kernels) with pointwise nonlinearities. Each block transforms feature maps $F_k : \mathbb{Z}^d \to \mathbb{R}^{c_k}$ to $F_{k+1} : \mathbb{Z}^d \to \mathbb{R}^{c_{k+1}}$ (where $c_i = \dim(F_i)$ is the channel dimension at layer i) through

$$F_{k+1}(\boldsymbol{x}) = \sigma\left(\sum_{\|\boldsymbol{y}\|\le r} K(\boldsymbol{y}) F_k(\boldsymbol{x}+\boldsymbol{y}) + \boldsymbol{b}\right)$$

  where $\mathbf{x} \in \mathbb{Z}^d$ is spatial position in output feature map, $\mathbf{y} \in \{-r,...,r\}^d$ is offset within a convolution kernel, $K(\mathbf{y}) \in \mathbb{R}$ is kernel weight at offset $\mathbf{y}$, $F_k(\mathbf{x}+\mathbf{y})$ is input feature at position $\mathbf{x}+\mathbf{y}$, $b \in \mathbb{R}$ is bias term, σ is nonlinear activation (e.g., ReLU), and r is kernel radius (e.g., $r = 1$ for 3×3 kernels).

- **Downsampling**: Pooling layers induce spatial compression by local aggregation, typically via max or average operations over $s \times s$ windows.

Remark 2.3. *The architectural constraints of CNNs – locality, weight sharing, and hierarchical composition – encode an implicit prior favoring translation-equivariant feature detection while maintaining parametric efficiency.*

2.2 Topological Convolutional Neural Networks

To contextualize our analytical framework (cf. Carlsson [2]), we adopt theoretical proof of the existence of Klein bottle in the space of local image patches. 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 $\mathscr{F} \subset C(D,\mathbb{R})$ that satisfy the approximation property

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

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

Let $\mathscr{Q}$ denote the space of bivariate quadratic polynomials, explicitly parametrized as

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

This constitutes a 6-dimensional real vector space. Our analysis focuses on the constrained subspace $\mathscr{P} \subset \mathscr{Q}$ defined by the conditions

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

The linear constraint alone reduces $\mathscr{Q}$ to a 5-dimensional affine subspace, while the quadratic normalization further restricts $\mathscr{P}$ to a 4-dimensional ellipsoid embedded within this subspace.

We subsequently characterize the submanifold $\mathscr{P}_0 \subset \mathscr{P}$ 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 $\mathscr{Q}$ is 4-dimensional: 3 parameters define q, and (λ,μ) lies on the unit circle, which is 1-dimensional. Incorporating the two additional constraints reduces this to a 2-dimensional complex $\mathscr{P}_0$.

Theorem 2.4 (Carlsson [2])**.** $\mathscr{P}_0$ *is homeomorphic to the Klein bottle* $\mathscr{K}$*.*

Proof. The function space $\mathscr{P}_0$ consists of 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

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

Let us construct $\mathscr{K}$ by quotient maps as follows. The original space $\mathscr{Q}$ 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

$$\theta_2(q) = q_2$$

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

Define an involution $f : \mathscr{Q} \to \mathscr{Q}$ 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

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

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 $\mathscr{Q}$. 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:

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

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

2.3 Phonemes

2.3.1 Phonetic Building Blocks

Phonemes, systematically categorized into vowels and consonants based on articulatory properties, serve as the atomic units of speech. The systemic coordination between these units forms the structural basis of the spoken language. This hierarchical organization drives research emphasis toward suprasegmental analysis (words/sentences), where expanded contextual dependencies enable more reliable pattern identification. Most speech systems employ a three-tiered processing hierarchy:

- **Phoneme Level**: 40–60 basic units (English: 44 phonemes) with 50–200 ms duration, subject to coarticulatory variation
- **Syllable Level**: Combinations constrained by phonotactic rules, yielding approximately 10^2 to 10^5 licit structures possible through phoneme concatenation. The number of licit syllables varies greatly across different languages, depending on their syllable structure rules.
- **Prosodic Level**: Supra-segmental features, such as pitch contours and stress patterns, encode pragmatic and syntactic information, these features play a crucial role in conveying semantic information

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

2.3.2 Phonetic Classification via IPA Standards

The International Phonetic Alphabet (IPA) is a system of phonetic notation designed to represent the sounds of spoken language. Each symbol in the IPA corresponds to a specific phoneme. This includes consonants, vowels, and suprasegmental features like stress and intonation. Our analysis focuses exclusively on pulmonic consonants and vowels, as non-pulmonic consonants exhibit negligible prevalence in English.

Consonants are classified through three articulatory dimensions. The **place** of articulation refers to where the airflow is obstructed, such as bilabials [p][b], labiodentals [f][v], and alveolars [t][d]. The **manner** of articulation describes how the airflow passes through the oral cavity, including plosives [p][t][k], fricatives [s][z] [f], affricates [ts][tʃ], nasals [m][n][ŋ], and approximants [j] [w]. **Voicing** indicates whether the vocal cords vibrate; for example, [p] is voiceless, while [b] is voiced. This three-dimensional classification system comprehensively describes the phonetic characteristics of consonants.

Vowels are systematically mapped in the IPA based on **tongue height**, **backness**, and **lip rounding** (Fig. 1). Tongue height is divided into high, mid, and low, with [i] as a high vowel, [e] as a mid-high vowel, and [a] as a low vowel. The frontness or backness refers to the position of the tongue in the mouth, with [i] as a front vowel, [u] as a back vowel, and [ə] as a mid vowel. Lip rounding indicates whether the lips are rounded during articulation; for instance, [i] is an unrounded vowel, while [u] is a rounded vowel. This three-dimensional classification method allows for precise identification of various vowels, such as [i] (front, high, unrounded) and [u] (back, high, rounded) in English.

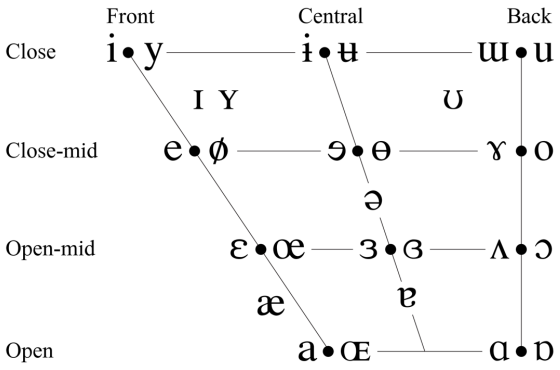

Figure 1: Positioning of vowels in oral cavity

To streamline English phonetic notation, ARPABET was developed as a practical alternative, encoding the 39 phonemes of General American English into ASCII-based representations. Created by the Advanced Research Projects Agency (ARPA) during the 1970s as part of the Speech Understanding Research project, ARPABET provides a systematic mapping of phonemes and allophones using distinct ASCII character sequences. Two encoding schemes were initially proposed: a single-character system (with alternating uppercase and lowercase letters) and a more flexible one- or two-character case-insensitive system. The latter gained broader adoption due to its practicality. In this study, we exclusively employ the two-letter coding scheme for phonetic representation. For a detailed comparison between ARPABET and the IPA, please refer to the mapping table in Tab. 1.

Both Fig. 1 and Tab. 1 are adapted from Wikipedia articles: https://en.wikipedia.org/wiki/International_Phonetic_Alphabet and https://en.wikipedia.org/wiki/ARPABET.

2-Letter Codes	IPA	Examples	2-Letter Codes	IPA	Examples	2-Letter Codes	IPA	Examples
AA	ɑ∼ɒ	b**a**lm, b**o**t	UW	u	b**oo**t	N	n	**n**ight
AE	æ	b**a**t	UX	ʉ	d**u**de	NX or NG	ŋ	si**ng**
AH	ʌ	b**u**tt	B	b	**b**uy	NX	ɾ̃	wi**nn**er
AO	ɔ	c**augh**t, st**o**ry	CH	tʃ	**Ch**ina	P	p	**p**ie
AW	aʊ	b**ou**t	D	d	**d**ie	Q	ʔ	uh-oh
AX	ə	comm**a**	DH	ð	**th**y	R	ɹ	**r**ye
AXR	ɚ	lett**er**, forw**ar**d	DX	ɾ	bu**tt**er	S	s	**s**igh
AY	aɪ	b**i**te	EL	l̩	bott**le**	SH	ʃ	**sh**y
EH	ɛ	b**e**t	EM	m̩	rhyth**m**	T	t	**t**ie
ER	ɝ	b**ir**d, fore**wor**d	EN	n̩	butt**on**	TH	θ	**th**igh
EY	eɪ	b**ai**t	F	f	**f**ight	V	v	**v**ie
IH	ɪ	b**i**t	G	g	**g**uy	W	w	**w**ise
IX	ɨ	ros**e**s, rabb**i**t	HH or H	h	**h**igh	WH	ʍ	**wh**y (without wine–whine merger)
IY	i	b**ea**t	JH	dʒ	**j**ive			
OW	oʊ	b**oa**t	K	k	**k**ite	Y	j	**y**acht
OY	ɔɪ	b**oy**	L	l	**l**ie	Z	z	**z**oo
UH	ʊ	b**oo**k	M	m	**m**y	ZH	ʒ	plea**s**ure

Table 1: ARPABET–IPA phonetic notation system mapping table

2.3.3 STFT and Spectrograms

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 [11].

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

$$X(f,t) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-2\pi i f \tau} \, \mathrm{d}\tau$$

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 STFT. This process captures temporal–frequency domain features, providing a foundation for subsequent audio analysis. Fig. 2 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.

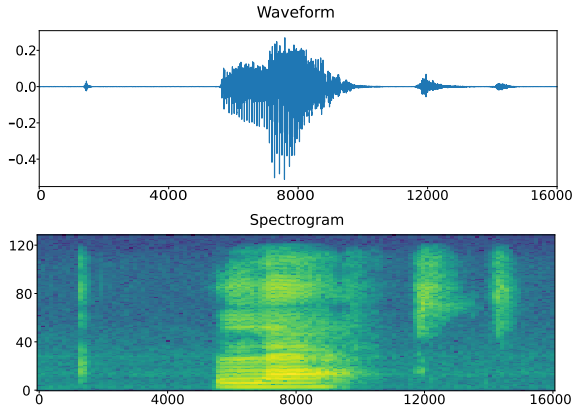

Figure 2: Waveform of the word “left” and its corresponding spectrogram on mini speech commands

3 The Space of Spectrogram Convolution Kernels

In this section, we consider the group action of the third-order special orthogonal group $\mathrm{SO}(3)$ on the space of 3×3 real matrices. By leveraging the invariance properties of the group action, we first reduce the dimension of the matrix space to 5. Subsequently, a new representation of the matrix space is introduced through orbit spaces and the special orthogonal group.

3.1 The Space of High-Contrast Spectrogram Convolution Kernels

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

Definition 3.1 (Norm of Convolution Kernels)**.** *Let*

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

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*

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

Definition 3.2 (Contrast of Convolution Kernels)**.** *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 3.3. *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 normalized and optimized for high contrast.

Definition 3.4 (Normalized Convolution Kernels)**.** *We consider the subspace of* $M_{3\times 3}(\mathbb{R})$ *consisting of convolution kernels* $\boldsymbol{A} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$ *satisfying the unit norm condition*

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

Definition 3.5 (Contrast-Maximizing Constraint)**.** *In order to maximize contrast, we further impose the constraint that the kernels belong to the orthogonal complement of the zero-contrast subspace. Concretely, we require*

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

Definition 3.6 (The Kernel Space M)**.** *Let M denote the set of all normalized 3×3 convolution kernels satisfying the contrast-maximizing constraint, i.e.,*

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

Theorem 3.7. *The space M is homeomorphic to the 5-dimensional sphere S^5.*

Sketch of Proof. The constraints $\|\boldsymbol{A}\| = 1$ and $\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 = \mathbf{0}$ define a smooth submanifold of $M_{3\times 3}(\mathbb{R})$. One may show via dimension counting and the implicit function theorem that this submanifold has dimension $9 - 1 - 3 = 5$, since $M_{3\times 3}(\mathbb{R}) \cong \mathbb{R}^9$ 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. □

3.2 Group Actions and Quotient Spaces

Observe that the group of orthogonal transformations acts on M as follows.

Definition 3.8 (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 3.9 (Contrast Projection for General Matrices)**.** *Given any matrix $\boldsymbol{A} \in M_{3\times 3}(\mathbb{R})$ whose 3 column vectors are not identical (otherwise 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, i.e.,*

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

2. ***Centering**: Subtract the mean value from each component and obtain*

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

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

Remark 3.10. *This projection satisfies*

- *invariance under orthogonal transformations, i.e., $\|\boldsymbol{Q}\boldsymbol{v}\| = \|\boldsymbol{v}\|$, and*
- *translation invariance, i.e., $\boldsymbol{v}_i \mapsto \boldsymbol{v}_i + \boldsymbol{c}$ cancels in* (1)*.*

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. In particular, for any $\boldsymbol{m} \in M$ and 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 3.11 (Quotient Space under Orthogonal Group Action)**.** *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:

$$x + y + z = \frac{1}{2} \quad \text{and} \quad z^2 \le xy. \tag{2}$$

Proposition 3.12. *The quotient space $B = M/\mathrm{SO}(3)$ is homeomorphic to the closed disk D^2.*

Sketch of Proof. Note that the relations (2) are equivalent to

$$x + y + z = \frac{1}{2} \quad \text{and} \quad 9\left(x + y - \frac{2}{3}\right)^2 + 3(x-y)^2 \le 1.$$

From this, one can show that the set of equivalence classes is continuously parametrized by two independent parameters satisfying an inequality that defines a closed 2-dimensional disk. A detailed study of the invariants associated with the $\mathrm{SO}(3)$-action yields the claim that B is homeomorphic to D^2. □

In particular, the boundary of B, denoted by ∂B, corresponds to when the equality $z^2 = xy$ holds in (2).

Remark 3.13. *For the equivalence class (orbit) containing $\frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ in B, its preimage set in M along the quotient map 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 2, not $3 = 5 - 2$, which shows that $M \to B$ is not a fiber bundle.

Indeed, we obtain a *stratified* fiber bundle (see, e.g., [12]) over B as follows.

- On the boundary of B, denoted ∂B, the fiber is $\mathrm{SO}(3)/\mathrm{SO}(2) \cong S^2$, each modulo rotations around a fixed axis.
- On the region where $x = y$, the fiber is $\mathrm{SO}(3)/\mathbb{Z}_2 \cong L(4,1)$, each modulo rotation by 180° about a fixed axis. Here $L(4,1)$ is a lens space.
- On the intersection of the two aforementioned cases, i.e. $\boldsymbol{v}_1 + \boldsymbol{v}_3 = \mathbf{0}$, the fiber is given by $\mathrm{SO}(3)/(\mathrm{SO}(2) \rtimes \mathbb{Z}_2)$, which is isomorphic to $\mathbb{RP}^2$ (the real projective plane).
- On the remaining portion, we have a principal $\mathrm{SO}(3)$-bundle.

Example 3.14. *More explicitly, given $(x,y) \in B$, we can choose a representative in its preimage in M to be $[\boldsymbol{v}_1, -\boldsymbol{v}_1 - \boldsymbol{v}_3, \boldsymbol{v}_3]$ with*

$$\boldsymbol{v}_1 = \sqrt{\frac{x}{3}}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{v}_3 = \sqrt{\frac{y}{3}}\cos\phi \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \sqrt{\frac{y}{6}}\sin\phi \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

where $\sin\phi = \sqrt{1 - \cos^2\phi}$ and $\cos\phi = \frac{\frac{1}{2} - x - y}{\sqrt{xy}}$ for $xy \neq 0$. This choice extends continuously to $(x,y) = (0, \frac{1}{2})$ and $(x,y) = (\frac{1}{2}, 0)$.

3.3 Summary of Theoretical Framework

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 a topological foundation for further analysis and applications in spectrogram-based speech processing.

4 New Spectrogram Convolution Filters

As shown in Lee et al. [8], there exist 8 basic vectors in the image patch. Here, up to constant factors, they will be reduced to just 2, since 2 of them are of zero contrast and the remaining can be reduced to 2 vectors through group actions.

Let us consider the orbits of these 2 vectors under group actions as convolution kernels, namely,

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

with $\boldsymbol{Q} \in \mathrm{SO}(3)$.

Additionally, this section focuses exclusively on phoneme-level recognition. Regarding the dataset, we cannot directly obtain phoneme-level annotations but instead employ segmentation tools. The Montreal Forced Aligner (MFA) [10] is utilized for this purpose. 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 overrepresented phonemes create prediction biases. Therefore, all experiments in this section employ a balanced subset of 500 samples per phoneme class for classification tasks. Finally, the primary datasets used in this section are derived from the SpeechBox corpus [1], TIMIT [17], and LJSpeech [6] with specific implementation details provided in the experimental section. We selected only half of the LJSpeech dataset for computability.

The general procedure for all experiments in this section goes as in Fig. 3. First, segment the audio signals from the dataset into phonemes through MFA. Subsequently, convert the audio signal corresponding to each phoneme into a spectrogram via STFT (see Sec. 2.3.3). These spectrograms are then fed into a CNN for training, where the network architecture contains two convolutional layers with 64 filters each, ultimately yielding the classification accuracy.

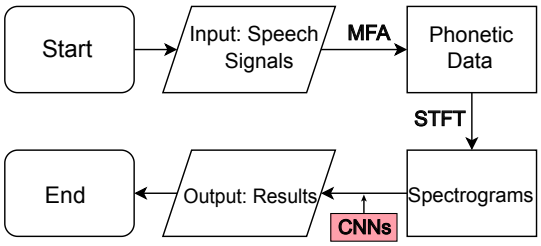

Figure 3: Workflow of topological CNNs for speech recognition, where topological enhancement takes place in prescribing the CNN kernels.

4.1 Theoretical Construction of Orthogonal Filters Layer

Given the 2 initial matrices $\boldsymbol{A}_1, \boldsymbol{A}_2 \in M_{3\times 3}(\mathbb{R})$, the layer is constructed through the following mathematical operations.

4.1.1 Matrix Augmentation

Extend the matrix set to ensure algebraic closure under inversion:

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

4.1.2 SO(3)-Informed Kernel Generation

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

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

Stochastic kernel generation proceeds as follows.

1. Sample $\theta_x, \theta_y, \theta_z \sim \mathscr{N}(0, \sigma^2)$ independently.
2. Construct Lie algebra element
$$\boldsymbol{\theta} = \sum_{i=x,y,z} \theta_i \boldsymbol{L}_i \in \mathfrak{so}(3).$$
3. Apply the exponential map and obtain
$$\boldsymbol{R} = \exp(\boldsymbol{\theta}) \in \mathrm{SO}(3)$$
where
$$\exp(\boldsymbol{\theta}) = \boldsymbol{I} + \frac{\sin\|\boldsymbol{\theta}\|}{\|\boldsymbol{\theta}\|}\boldsymbol{\theta} + \frac{1-\cos\|\boldsymbol{\theta}\|}{\|\boldsymbol{\theta}\|^2}\boldsymbol{\theta}^2$$
with $\|\boldsymbol{\theta}\| = \sqrt{\sum_{i=x,y,z}\theta_i^2}$.

4.1.3 Definition of Orthogonal Filters Layer

Definition 4.1 (Orthogonal Filters (OF) Layer)**.** *Given the kernel space M, each convolution kernel of untrained* Orthogonal Filters layer *is defined by*

$$\boldsymbol{W}_k = \alpha \cdot \boldsymbol{R}_k \boldsymbol{M}_k$$

where $\boldsymbol{R}_k \in \mathrm{SO}(3)$, $\boldsymbol{M}_k \in \mathscr{M}$, and $\alpha \in \mathbb{R}^+$ is an adjustable scaling factor.

4.2 Empirical Evaluation of Orthogonal Filters Layer in Phoneme Classification Tasks

In this section, we conduct experiments on phoneme classification and compare the newly proposed OF layers with multiple other convolutional neural network architectures, such as CF (Circle Filters) and KF (Klein Filters) from [9]. We will classify approximately 40 phoneme categories (with prosodic stress markers merged) across distinct datasets and experimental conditions to evaluate model robustness and generalizability.

4.2.1 Experiments on OF Kernels

The comparative results shown in Fig. 4 reveal 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.

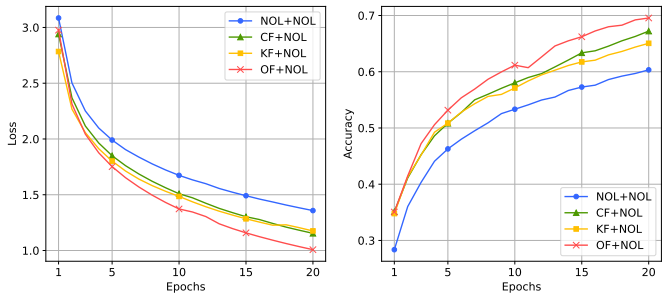

Figure 4: Comparisons of loss and accuracy on SpeechBox

4.2.2 Experiments on Canonical OF Kernels

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

$$\boldsymbol{Q}\begin{bmatrix}1 & 0 & -1\\ 1 & 0 & -1\\ 1 & 0 & -1\end{bmatrix}/\sqrt{6} \text{ and } \boldsymbol{Q}\begin{bmatrix}1 & 0 & 1\\ 1 & 0 & 1\\ 1 & 0 & 1\end{bmatrix}/\sqrt{6},\ \boldsymbol{Q}\in \mathrm{SO}(3). \quad (3)$$

In essence, this set of convolution kernels corresponds to vertical stripe detectors with the middle column set to be zero, structured as $[\boldsymbol{v}_1, \boldsymbol{0}, \pm\boldsymbol{v}_1]$, which is homeomorphic to the 2-sphere. For simplicity, the neural network architectures constructed using this set of convolution kernels will retain the name of OF convolutional layers.

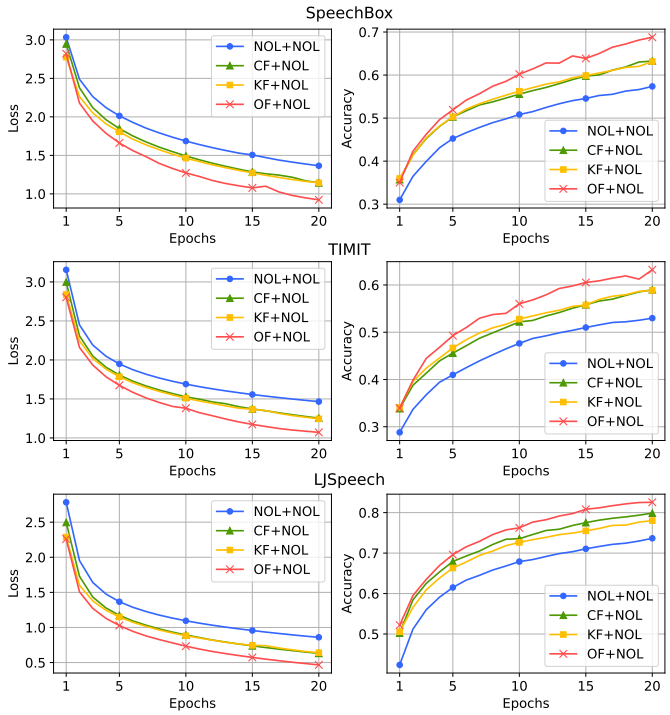

Figure 5: Comparisons of loss and accuracy on SpeechBox, TIMIT, and LJSpeech (with canonical OF kernels)

First, let us analyze the performance of these convolution kernels on the SpeechBox dataset (see Fig. 5). Here, we observe that the accuracy has approached 70%, outperforming both the previous orthogonal counterparts and other comparative models. Experimental results on the two additional datasets, TIMIT and LJSpeech, are also reported, yielding consistent findings.

4.2.3 Experiments with Noise

Analysis of the figures reveals that the datasets exhibit descending accuracy rankings: LJSpeech > SpeechBox > TIMIT, which is likely attributed to the variation in acoustic clarity across the datasets. This section investigates the impact of introducing additive white Gaussian noise (AWGN) on model performance.

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

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

with P_{signal} and P_{noise} representing the power of the original speech signal and the injected Gaussian noise, respectively. The implementation protocol comprises the following 3 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 shown in Fig. 6. The graphical comparison

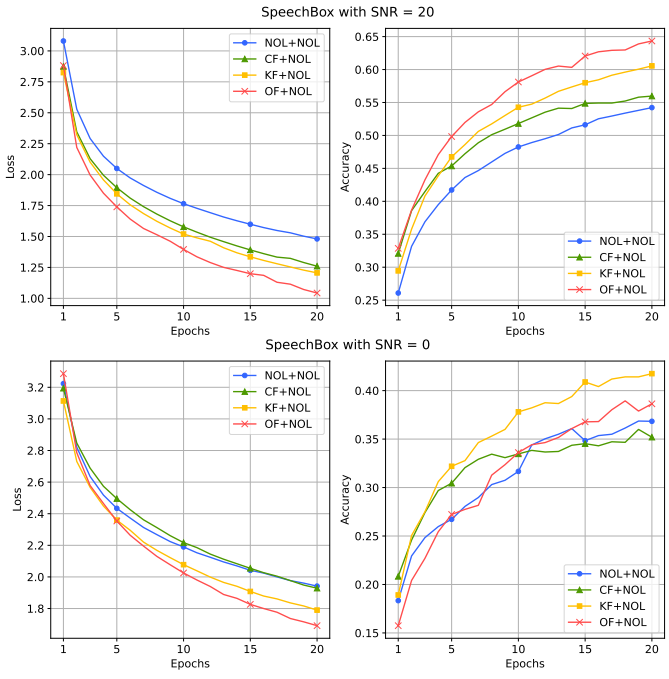

Figure 6: Comparisons of loss and accuracy on SpeechBox with noise (Upper: with SNR = 20; Lower: with SNR = 0)

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, and KF achieves the optimal performance.

This phenomenon might arise from the severe degradation of vertical stripe structures caused by additive noise, leading to reduced accuracy. Consequently, in anti-noise experiments, KF manifests enhanced stability, while OF maintains superior accuracy under low-noise scenarios.

As for the convolution 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.

5 Further Applications and Extensions of Theoretical Framework

This section focuses on addressing gaps and extending prior experimental findings. We begin by supplementing earlier experiments with an analysis of scenarios where no phoneme filtering is applied, providing insights into performance under realistic conditions. Subsequently, we examine how different convolutional neural network architectures perform in word and image classification tasks, showcasing OF's versatility and efficiency across domains.

5.1 Phoneme Classification

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 selection) to assess performance under more realistic conditions.

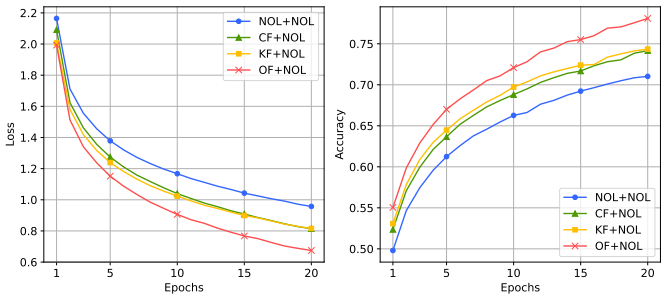

Figure 7: Comparisons of loss and accuracy on SpeechBox without selection

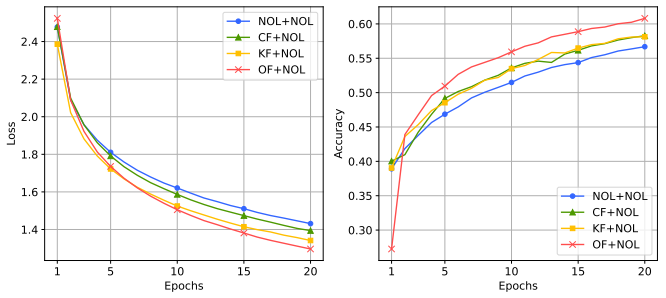

Figure 8: Comparisons of loss and accuracy on SpeechBox (SNR = 0) without selection

Here, the 4 figures illustrate the training performance of various neural network architectures across 4 datasets, SpeechBox (Fig. 7), SpeechBox with SNR = 0 (Fig. 8), TIMIT (Fig. 9), and LJSpeech (Fig. 10), under conditions where no phoneme-count filtering is applied.

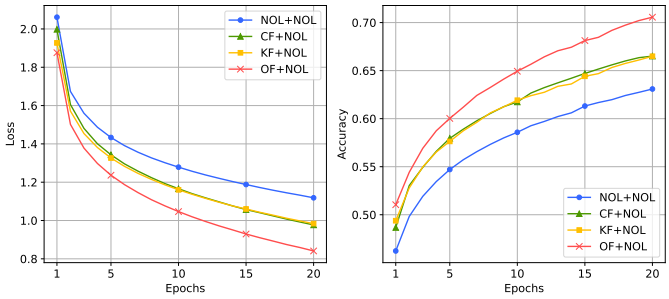

Figure 9: Comparisons of loss and accuracy on TIMIT without selection

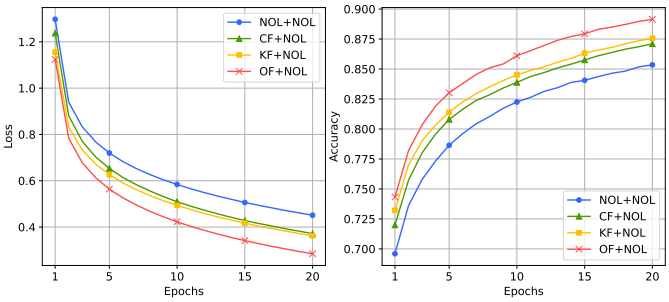

Figure 10: Comparisons of loss and accuracy on LJSpeech without selection

The experimental results align with expectations in that our proposed convolution kernel remains optimal, particularly under noise-free conditions. However, it is noteworthy that neural networks incorporating circle features and Klein features unexpectedly outperformed traditional architectures, despite prior assertions of their incompatibility with audio tasks. It is desirable to have a better understanding of the mechanism behind topological inputs enhancing neural networks.

5.2 Word Classification

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 [15], a dedicated word-level corpus explicitly designed with approximately balanced frequency distributions across all lexical entries, for comprehensive evaluation. Fig. 11 demonstrates that our neural network model exhibits robust adaptability to word-level tasks, further validating its versatility across lexical processing challenges.

5.3 Image Classification

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 our OF method. We selected the CIFAR10 dataset [7] for its higher complexity relative to MNIST, providing a more challenging benchmark to evaluate model robustness in handling intricate feature representations (see Fig. 12). The results demonstrate that

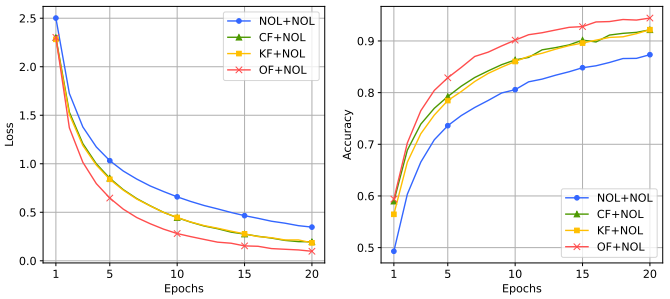

Figure 11: Comparisons of loss and accuracy on SpeechCommands

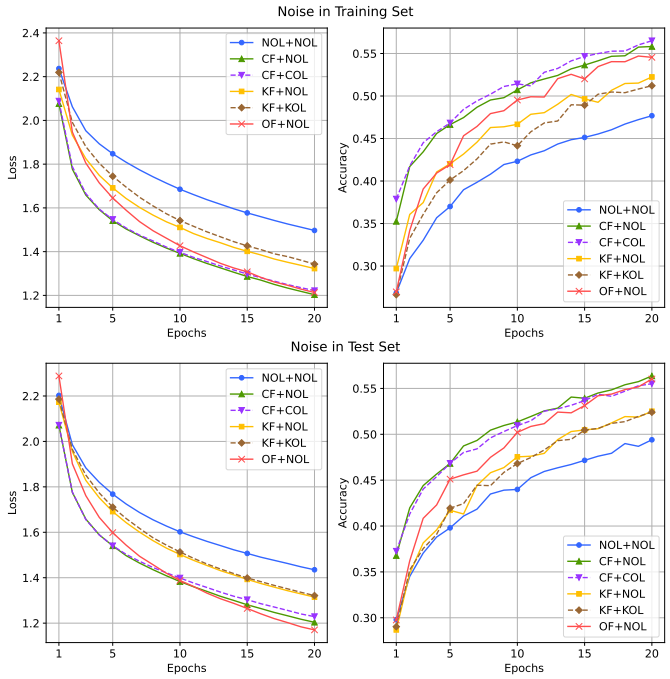

Figure 12: Comparisons of loss and accuracy on CIFAR10

our model achieves superior performance over conventional neural networks on image-based tasks, while maintaining parity with architectures utilizing Klein features, underscoring its cross-modal versatility.

6 Conclusions

This study establishes two principal contributions to convolution kernel design. First, it rigorously bridges geometric feature representation with frequency-domain characteristics by systematically analyzing 3×3 kernels through dual frameworks, i.e., manifold theory and Fourier spectral decomposition. Second, it derives an optimized set of foundational kernels from first principles, demonstrating measurable improvements over conventional initialization methods.

Practically, our work leverages topological methods (inspired by Carlsson and his collaborators' pioneering work) to extract weight distribution features, enabling the construction of specialized kernels for phoneme recognition. A novel contrast metric, based on temporal audio variations, further guides targeted kernel selection. Experiments reveal that the proposed kernels significantly enhance speech recognition accuracy, particularly in low-noise environments, while maintaining robustness against moderate noise. Notably, the kernels' performance extends beyond their original domain: though designed for audio tasks, they achieve competitive accuracy in traditional image classification, underscoring the versatility of our topological approach.

While the kernels exhibit modest gains under high noise, their cross-domain efficacy—matching or surpassing methods such as Love et al.'s in both speech and image tasks—highlights the broader potential of mathematically grounded kernel optimization. Future work could refine noise resilience, but the current results affirm that unifying geometric and spectral principles yields kernels with generalized representational power.

Finally, there are two questions which we do not address in this article. First, a direct comparison between the OF-topological deep learning with spectrograms as presented here and state-of-the-art neural networks for speech recognition, such as Gated Recurrent Unit, or how the former approach may enhance the latter. Indeed, topological characteristics for audio and speech signals and their integration with machine learning (in terms of both feature representations and neural network architectures) have yet gained wide recognition and utilization in the field of speech processing. They are currently more of theoretical interest and practical potential to the industry. Nevertheless, in consultation with experts in the field, our research group has carried out experiments in this direction and obtained informative results that confirm the potential of topological deep learning (see [5, esp. Secs. 2.2, 2.1.2, and 2.1.3]). The other question concerns intrinsic topological distributions of speech or audio signals, analogous to the Klein bottle model for high-contrast local natural image data. The spectrograms we work with here serve as a mediator between the two types of data, image and speech, though our proposed OF kernels apply and extend to visual and time series data with certain rotational asymmetry as well.

Acknowledgments

This article grew out of the first author's doctoral thesis, and he thanks Prof. Fuquan Fang for his support and guidance. The authors extend their sincere gratitude to Prof. Gunnar Carlsson for his insightful discussion on topological methods in image analysis. This work was partly supported by the National Natural Science Foundation of China grant 12371069 and a grant from the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science.

References

[1] SpeechBox. 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.

[2] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, Apr. 2009. doi: 10.1090/S0273-0979-09-01249-X

[3] G. Carlsson and R. B. Gabrielsson. Topological approaches to deep learning. In *Topological Data Analysis*, pp. 119–146. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-43408-3_5

[4] G. Carlsson, T. Ishkhanov, V. De Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76:1–12, Jan. 2008. doi: 10.1007/s11263-007-0056-x

[5] P. Feng, Q. Qu, H. Zhang, S. Yi, Z. Yu, Z. Ding, and Y. Zhu. Topology-enhanced machine learning for consonant recognition. 2025. arXiv:2311.15210. Updated at https://yifeizhu.github.io/tail.pdf.

[6] K. Ito and L. Johnson. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[7] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf, 2009.

[8] A. B. Lee, K. S. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54:83–103, Aug. 2003. doi: 10.1023/A:1023705401078

[9] E. R. Love, B. Filippenko, V. Maroulas, and G. Carlsson. Topological convolutional layers for deep learning. *Journal of Machine Learning Research*, 24(59):1–35, 2023.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech 2017*, pp. 498–502. doi: 10.21437/Interspeech.2017-1386

[11] A. V. Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[12] E. Ross. Stratified vector bundles: Examples and constructions. *J. Geom. Phys.*, 198:Paper No. 105114, 25, 2024. doi: 10.1016/j.geomphys.2024.105114

[13] V. d. Silva and G. Carlsson. Topological estimation using witness complexes. In M. Gross, H. Pfister, M. Alexa, and S. Rusinkiewicz, eds., *SPBG'04 Symposium on Point-Based Graphics 2004*. The Eurographics Association, 2004. doi: 10.2312/SPBG/SPBG04/157-166

[14] E. Tulchinskii, K. Kuznetsov, L. Kushnareva, D. Cherniavskii, S. Barannikov, I. Piontkovskaya, S. Nikolenko, and E. Burnaev. Topological data analysis for speech processing. In *Proc. Interspeech 2023*, pp. 311–315. doi: 10.21437/Interspeech.2023-1861

[15] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. arXiv:1804.03209.

[16] A. Zia, A. Khamis, J. Nichols, U. B. Tayab, Z. Hayder, V. Rolland, E. Stone, and L. Petersson. Topological deep learning: A review of an emerging paradigm. *Artificial Intelligence Review*, 57(4):77, Feb. 2024. doi: 10.1007/s10462-024-10710-9

[17] V. Zue, S. Seneff, and J. Glass. Speech database development at MIT: Timit and beyond. *Speech Communication*, 9(4):351–356, 1990. doi: 10.1016/0167-6393(90)90010-7