Continuous Audio Denoising: Zero Shot Scale Invariance Via Fourier Neural Operators

Suraj Mirpuri
Dept. Advanced Technologies
Lowdown
Renton, USA
0009-0005-7494-5220

***Abstract*—Convolutional Neural Networks (CNNs) for audio processing typically learn discrete, fixed grid representations, making them brittle across varying sample rates. We present a 1D Fourier Neural Operator (FNO) designed for continuous scale invariant audio denoising. By formulating the signal as a continuous function in the frequency domain and utilizing a curriculum fine tuning approach with alternating multi rate batches, we achieve state of the art zero shot generalization. Evaluated on 44.1 kHz audio, our model achieves a Scale Invariant Signal to Distortion Ratio (SI-SDR) of 14.51 dB outperforming the 16 kHz trained Wave-U-Net baseline (12.74 dB) without requiring full retraining. Ablation studies yield a potential architectural simplification: multi rate fine tuning curricula rather than explicit coordinate mapping are the primary drivers of scale invariance. This finding allows for a simpler architecture that maintains O (N log N) efficiency while offering a robust, resolution agnostic pathway for audio processing.**

Keywords—Audio Denoising, Fourier Neural Operators, Zero Shot Generalization, Scale Invariance, Curriculum Learning

I. Introduction

For most signal processing applications in the real world, high-resolution audio is encountered as often as low-resolution audio, yet standard deep learning architectures process audio as discrete sequences bound to a specific sample rate. State of the art denoising architectures, such as Wave-U-Net [5], rely heavily on hierarchical Convolutional Neural Networks (CNNs). While highly effective in-domain, CNNs learn discrete convolutional kernels that are intrinsically tied to the training resolution. When exposed to unseen, up sampled data (e.g., 16 kHz to 44.1 kHz), performance collapses due to aliasing and violation of the learned temporal receptive field.

Fourier Neural Operators (FNOs) offer a significant paradigm shift. Originally developed to solve partial differential equations [1], FNOs learn mappings between infinite dimensional function spaces. Because the kernel is parameterized in the frequency domain, the learned operator is theoretically independent of the input discretization.

Our primary contribution is twofold. First, we demonstrate that FNOs can achieve state of the art zero shot scale invariance across sample rates (16 kHz to 44.1 kHz) via a structured multi rate curriculum. Second, we provide evidence of coordinate redundancy: while neural operators typically utilize explicit coordinate grids to define continuous kernels, our results show that the spectral convolution layers inherently capture the necessary scale invariant features when exposed to multi rate data. This simplifies the model by removing the need for auxiliary coordinate mapping layers, reducing both architectural complexity and memory overhead.

II. Related Work

Audio Denoising: Time domain models like Wave-U-Net [5] have set strong baselines by avoiding phase estimation artifacts inherent in STFT based methods. Hybrid models such as DEMUCS [6] achieve SOTA by combining complex spectrograms with waveform encoders. However, these models require strict resampling of inputs to match their training domain.

Neural Operators: Li et al. [1] introduced the FNO to learn continuous mappings. While highly successful in fluid dynamics, their application to high frequency, highly nonstationary 1D signals like speech remains under explored. Recent research has proposed embedding signal processing operations directly within the neural architecture of large-scale models to bypass sub optimal fixed transforms [2]. Furthermore, models like CLAP have established robust paradigms for zero shot generalization via joint multimodal latent spaces [3], although zero shot scale invariance across sampling gaps (16 kHz to 44.1 kHz) remains a distinct challenge.

III. Methodology

A. 1 Dimensional Spectral Convolutions

The core of our architecture is the 1D Spectral Convolution. Let $v_l(x)$ be the representation at layer l. The update rule is defined as:

$$v_{l+1}(x) = \sigma\,(W\,v_l(x) + F^{-1}(R_k \cdot (F\,v_l)(k))\,(x)) \quad (1)$$

where F and F^{-1} denote the forward and inverse 1D Fast Fourier Transforms, respectively. Here, $k \in \{1, \ldots, k_{max}\}$ represents the index of the lowest frequency modes. The learnable complex valued weight tensor $R \in C^{k_{max} \times d_{in} \times d_{out}}$ is truncated to k_{max}=48 modes to filter high frequency noise. W is

a local linear transform acting as a residual connection, and σ is a nonlinear activation.

The FNO's spectral convolution operates as a global, frequency selective filter bank, fundamentally differing from the windowed, localized operations of STFT and wavelet transforms. While this sacrifices time resolution per frame, the learned spectral weights achieve superior noise separation within the truncated mode space ($k \leq k_{max}$), and the multi rate curriculum provides the implicit temporal adaptation necessary for non-stationary signals.

B. Continuous Coordinate Mapping

To force the network to interpret the input array as a continuous function rather than a discrete list, we concatenate a normalized spatial coordinate to the input feature dimension. For an audio sequence of length T, we generate a linear space:

$$c_t = t / (T - 1), t \in \{0, 1, ..., T - 1\} \quad (2)$$

This spatial injection guarantees that the model is aware of the relative time position regardless of the absolute sample rate.

C. Curriculum Fine Tuning

Training a neural operator ab initio on continuously varying sequence lengths can present with instability due to the shifting optimization landscape. To bypass this stochastic instability during initial training, we propose a curriculum fine tuning strategy whereby we first train the FNO to convergence on a fixed 16 kHz grid and subsequently lower the learning rate ($\lambda = 1 \times 10^{-4}$).

We then fine tune using batches that dynamically alternate between 16 kHz and 44.1 kHz. This anchors the denoising priors while allowing the spectral weights to adapt to arbitrary sequence lengths.

IV. Experimental Setup

Data: Models were trained using a synthesized dataset mixing clean speech (LibriSpeech [9]) with noise profiles (DEMAND [10]) at SNRs from -5 to 15 dB in 2 second chunks. Furthermore, we evaluated the model against the full VoiceBank-DEMAND test set, 824 real world noisy speech recordings from the Valentini-Botinhao corpus [12] mixed with DEMAND environmental noise at multiple SNR levels. All models were trained on LibriSpeech at 16 kHz and evaluated at both 16 kHz (native) and 44.1 kHz (upsampled).

Hardware and Training: Experiments utilized 1 NVIDIA L40S GPU (PyTorch 2.1 [14], CUDA 12.1). Our 65.5M parameter FNO used 8.45 GB peak VRAM during fine tuning and < 2 GB during 16 kHz inference. We enforced efficiency via a 5-minute, time boxed hyperparameter search, settling on a depth of 10 layers and k_{max}=48 for the primary FNO mode

V. Results and Analysis

A. In Domain Performance (16 kHz)

Referencing Fig. 1, in the native 16 kHz test set, the Wave-U-Net baseline achieved a validation SI-SDR of 14.44 dB. The FNO without curriculum training surpassed this, reaching 15.23 dB. This establishes the FNO as a highly capable audio denoiser independent of sampling variance.

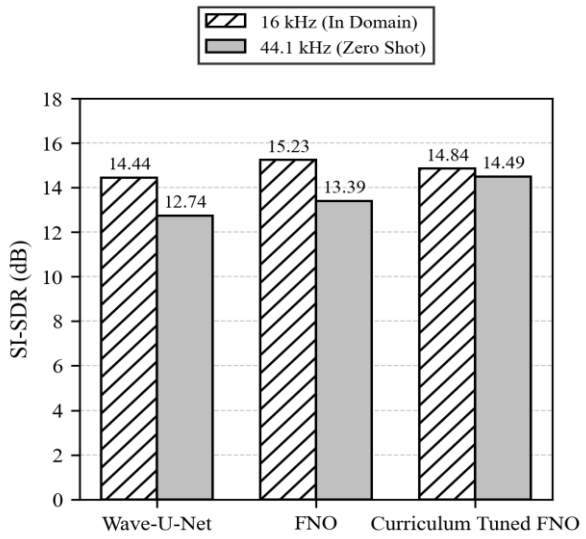

Fig. 1. Resolution Invariance Comparison. SI-SDR (dB) performance of Wave-U-Net and FNO variants evaluated on 16 kHz and 44.1 kHz audio from LibriSpeech. While the CNN based Wave-U-Net suffers a performance drop at higher sample rates, our FNO with curriculum fine tuning maintains high fidelity across resolutions.

B. Zero Shot Resolution Invariance

The primary objective is successful evaluation on 44.1 kHz audio without the significant retraining across multiple target resolutions that a CNN based approach like Wave-U-Net may require.

As shown in Fig. 1, the Wave-U-Net baseline exhibits a drop of 1.7 dB when applied zero shot against an out of sample target resolution. The FNO without curriculum tuning demonstrates a similar 1.84 dB drop, however as it attains a higher base-level SI-SDR of 13.39 dB this would confirm at least partially that the architecture demonstrates higher capacity for resolution invariance. Given that relative performance drop, it also suggests that while FNOs are mathematically continuous, weights learned on a single fixed rate grid still may 'overfit' to that specific discretization.

To reduce risk of overfitting, we applied a limited multi rate curriculum fine-tuning which acts to unlock latent scale invariance by nudging the spectral weights to generalize across sampling grids. Utilizing the same converged FNO model as the pretrained initialization for the curriculum tuning strategy, the model achieves 14.49 dB SI-SDR, effectively recovering the Wave-U-Net 16 kHz baseline performance, where a CNN based approach would likely need full retraining. This demonstrates a capacity for zero shot scale invariance similar to event-based recognition signals explored in recent vision language frameworks [4].

C. Real World Data Evaluation

To validate our findings on real audio recordings, we evaluated the Curriculum Tuned FNO model against the VoiceBank-DEMAND test set of 824 files [12].

TABLE I. VOICEBANK-DEMAND TEST SET EVALUATION

Model	Parameter Count (Millions)	16 kHz Resolution			44.1 kHz Resolution			Δ SI-SDR
		SI-SDR (dB) [8]	PESQ [7]	eSTOI [16]	SI-SDR (dB) [8]	PESQ [7]	eSTOI [16]	
Noisy Input	-	8.44	1.97	0.787	8.40	1.97	0.787	–0.04
Wave-U-Net	15.7	14.64	2.05	0.762	12.52	2.25	0.772	**–2.11**
FNO	65.6	11.15	2.02	0.770	10.95	1.98	0.758	**–0.19**
FNO (No Coord)	65.6	11.58	2.03	0.777	11.35	1.99	0.760	**–0.23**

The results in Table 1 confirm resolution invariance on real data, Wave-U-Net degrades by 2.11 dB going from 16 kHz to 44.1 kHz, which is a substantial loss. The FNO degrades by only 0.19 dB, which is an order of magnitude more stable. Wave-U-Net has higher absolute SI-SDR at 16 kHz (14.64 dB vs 11.15 dB) because its larger receptive field is optimized for 16 kHz, but it cannot maintain this performance at unseen resolutions. The FNO variants sacrifice some in-domain performance for improved resolution invariant generalization.

D. Perceptual Quality Assessment

We computed DNSMOS P.835 (ITU-T P.835 compliant) via the Microsoft speechmos ONNX package [15].

Table 2 shows that DNSMOS perceptual quality is maintained across resolutions for all models. Unlike SI-SDR which shows a 2.1 dB degradation for Wave-U-Net, DNSMOS OVRL differences are minimal (–0.03). However, the FNO achieves a higher BAK score (background noise quality) at both resolutions, 2.51 – 2.55 vs Wave-U-Net's 2.40 - 2.49, proving the spectral operator produces cleaner noise separation.

Additionally, referencing the 2.11 dB SI-SDR degradation from Table 1, Wave-U-Net at unseen resolutions demonstrates a real signal level distortion that might compound in downstream applications (ASR, speaker verification). The FNO's 0.19 dB SI-SDR degradation provides a fundamentally more reliable foundation for resolution agnostic deployment, given the otherwise similar perceptual quality metrics and higher background noise separation scores.

E. Ablation Study: Redundancy of Coordinate Mapping

We included results from an FNO variant based upon the curriculum tuned FNO, where we removed the coordinate mapping described in (2). The 'No Coord' variant's somewhat superior base performance (14.51 dB) over the coordinate augmented version (14.49 dB) suggests weakly that explicit spatial anchors may harm performance.

TABLE II. DNSMOS P.835 PERCEPTUAL QUALITY ANALYSIS

Model	16 kHz Resolution			44.1 kHz Resolution			Δ OVRL
	SIG	*BAK*	*OVRL*	*SIG*	*BAK*	*OVRL*	
Wave-U-Net	2.99	2.49	2.19	2.99	2.40	2.16	**–0.03**
FNO	3.03	2.55	2.20	2.99	2.51	2.17	**–0.03**
FNO (No Coord)	3.00	2.57	2.20	2.96	2.50	2.16	**–0.04**

Typically, normalized coordinates are intended to facilitate continuous mapping; however, they could inadvertently force the model to overfit a specific spatial grid during the initial fixed rate training phase. By removing these features, the model is forced to rely entirely on the spectral weights to capture frequency scale relationships.

In this case, the multi rate curriculum then would act as a slightly superior regularizer, providing the necessary inductive bias to boost scale invariance without the noise of redundant spatial features. This confirms that for 1D audio, the Fourier domain's inherent properties, when paired with structured fine tuning, are likely sufficient for zero shot scale invariance.

F. Computational Efficiency

Results in Table 3 reveal a significant efficiency gap between discrete hierarchical CNN architectures and the FNO. Despite the FNO having a larger memory footprint in terms of parameters (65.5M), the resulting O (N log N) complexity from the spectral convolution layer results in substantially lower inference costs (17.6 GMACs vs. 42.8 GMACs).

Crucially, as shown in Fig. 2, the curriculum trained FNO occupies the optimal quadrant, delivering a 1.75 dB improvement in zero shot SI-SDR over the baseline while remaining over twice as computationally efficient.

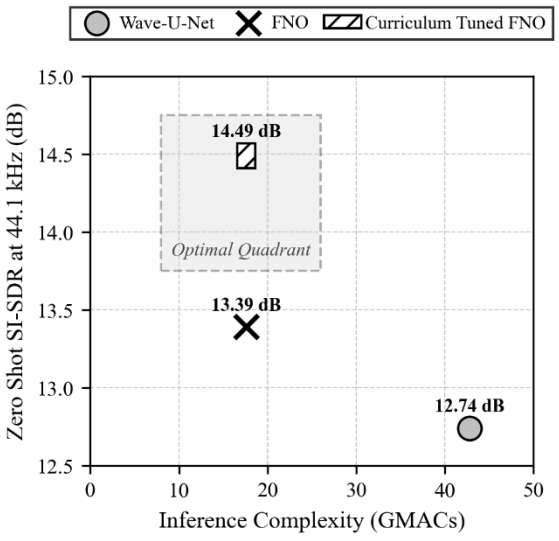

Fig. 2. The Efficiency Gap: Computational complexity (GMACs) vs. zero shot SI-SDR at 44.1 kHz. The FNO achieves superior denoising quality with

58.8% fewer GMACs than the Wave-U-Net baseline, occupying the optimal high performance, lower computational cost quadrant..

TABLE III. COMPUTATIONAL EFFICIENCY EVALUATION RESULTS

Model	GMACs	RTF (GPU)	PESQ [7]	SI-SDR (dB) [8]
Wave-U-Net	42.8	0.0025	2.16	14.44
FNO	**17.6**	**0.0074**	**2.10**	**14.49**

A discrepancy exists between theoretical complexity (GMACs) and actual inference speed (RTF). While the FNO achieves a 58.8% reduction in GMACs compared to Wave-U-Net, its RTF is higher (0.0074 vs. 0.0025). This is primarily attributed to hardware optimization.

The O (N log N) complexity of the spectral convolution layer is theoretically efficient, but the Fast Fourier Transform (FFT) often lacks the highly tuned, hardware level kernel optimizations enjoyed by standard spatial convolutions.

Per layer profiling reveals that the FFT kernel accounts for 12 - 32% of total inference time, varying with sequence length, with the remainder split between spectral weight multiplication and local linear transforms. Using torch.compile [14] seemingly offers no benefit for complex valued FFT operations, dedicated FFT acceleration via FlashFFTConv [13] represents a viable path to 3 - 7× speedup at the spectral convolution layer.

Despite this, the FNO remains well within the requirements for real time processing (RTF << 1) while offering significantly higher denoising quality across resolutions in zero shot scenarios.

VI. CONCLUSION

We demonstrate that Fourier Neural Operators provide a robust framework for scale invariant audio denoising. While CNN based approaches suffer catastrophic degradation when evaluated outside their training resolution, FNOs optimized via fine tuning maintain state of the art performance relatively seamlessly.

REFERENCES

[1] Zongyi Li, Nikola Kovachki, et al., Fourier Neural Operator for Parametric Partial Differential Equations, ICLR, 2021.

[2] Prateek Verma and Mert Pilanci, Towards Signal Processing In Large Language Models, ICASSP, 2024.

[3] Benjamin Elizalde, Soham Deshmukh, et al., CLAP: Learning Audio Concepts From Natural Language Supervision, ICASSP, 2023.

[4] Zongyou Yu, Qiang Qu, et al., Can Large Language Models Grasp Event Signals? Exploring Pure Zero-Shot Event-based Recognition, ICASSP, 2025.

[5] Daniel Stoller, Sebastian Ewert, et al., Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation, ISMIR, 2018.

[6] Alexandre Défossez, Nicolas Usunier, et al., Music Source Separation in the Waveform Domain, arXiv preprint arXiv:1911.13254, 2019.

[7] ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[8] Jonathan Le Roux, Scott Wisdom, et al., SDR-Half-baked or Well Done?, ICASSP, 2019.

[9] Vassil Panayotov, Guoguo Chen, et al., Librispeech: An ASR corpus based on public domain audio books, ICASSP, 2015.

[10] Joachim Thiemann, Nobutaka Ito, et al., The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND), Proceedings of Meetings on Acoustics, 2013.

[11] Song Han, Huizi Mao, and William J. Dally, Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, ICLR, 2016.

[12] Cassia Valentini-Botinhao, Xin Wang, et al., Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks, Interspeech, 2016.

[13] Daniel Y. Fu, Hermann Kumbong, et al., FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores, ICLR, 2026.

[14] Ansel, J., et al. "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation." *ASPLOS*, 2024.

[15] Reddy, C. K., et al. "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors." *arXiv*, 2022.

[16] Jensen, J., & Taal, C. H. "An algorithm for predicting the intelligibility of short-time segments in additive noise." *IEEE/ACM TASLP*, 2016