1st Latin American Music Information Retrieval Workshop
December 9 - 11, 2024 – Rio de Janeiro

AEROMamba: An efficient architecture for audio super-resolution using generative adversarial networks and state space models

Wallace Abreu*, Luiz W. P. Biscainho
Signals, Multimedia and Telecommunications Lab, Federal University of Rio de Janeiro, Brazil
*wallace.abreu@smt.ufrj.br

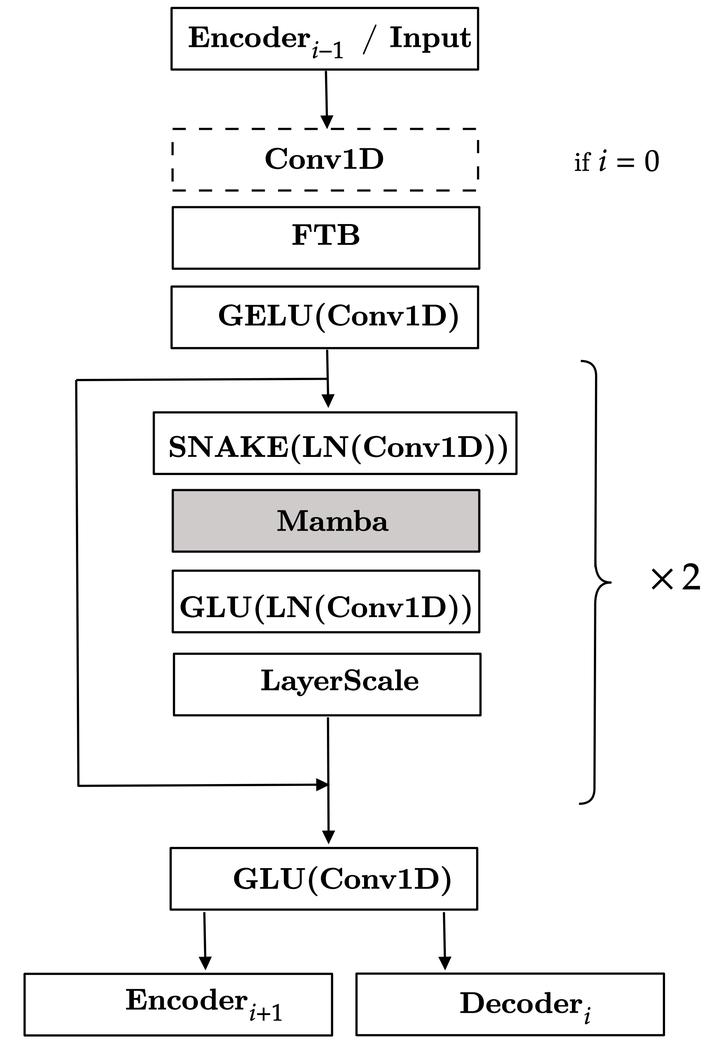
FAPERJ CAPES CNPq

1. MOTIVATION

- Background – Restoration of analog audio:
  - Historical recordings
    - Media degradation
    - Technology limitations
- Super-resolution
  - Digital audio applications
    - Lossy operations to reduce storage usually involve bandwidth reduction
    - e.g. Decimation and compression
- Generative adversarial networks:
  - Sampling is faster than Markov-chain-based architectures (e.g. Diffusion Models)
  - AERO as the base model
- Proposed Modification:
  - Replacement of recurrent and attention layers by Mamba

2. PROPOSED MODEL

- AEROMamba
  - AERO = Audio-Super Resolution Model
  - Mamba = Efficient State Space Model
- AERO
  - GAN architecture:
    - Hybrid Demucs-based Generator
    - MelGAN Discriminator
- Mamba
  - Emulates selective mechanism of attention
  - Exploits GPU structure



3. RESULTS

- **Experiments**
  - Super-resolution of popular music and piano pieces from 11.025 to 44.1 kHz
    - PianoEval (private collection dataset)
    - MUSDB18-HQ (open-source dataset)
- **Evaluation**
  - GPU usage and Inference speed
  - Objective metrics: ViSQOL and LSD
  - Subjective metrics: listening tests with 20 subjects, targeting audio similarity
- **Conclusions**
  - AEROMamba is 15x faster than AERO in inference using 5x-9x less GPU
  - AEROMamba achieves higher perceptual quality than AERO
  - For a fixed batch size, AEROMamba needs 2x-4x less GPU to train

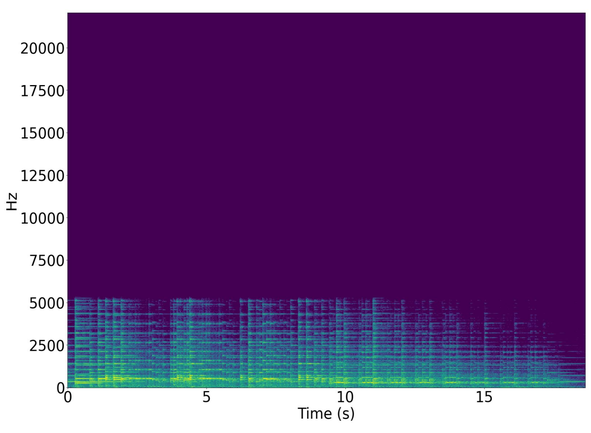
Computational Performance

Method	NVIDIA RTX 3090		NVIDIA RTX 2080 Ti		Parameters
	GPU Usage (MB)	Time (s)	GPU Usage (MB)	Time (s)	
AERO	17091	1.246	16420*	–	19,432,958
AEROMamba	3000	0.087	1914	0.063	20,964,190

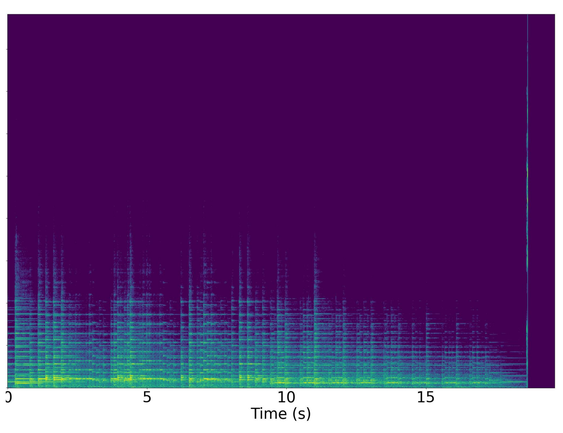
Objective and Subjective Metrics

Model	MUSDB18			PianoEval		
	ViSQOL ↑	LSD ↓	Score ↑	ViSQOL ↑	LSD ↓	Score ↑
Low-Resolution	1.82	3.98	38.22	4.36	1.09	72.92
AERO	2.90	1.34	60.03	**4.38**	**0.99**	76.89
AERO-HQ				4.34	1.04	-
AEROMamba	**2.93**	**1.23**	**66.47**	4.43	0.98	-
AEROMamba-HQ				**4.38**	1.00	**84.41**

Low Res - 11.025 kHz



AERO - 44.1 kHz



AEROMamba - 44.1 kHz