SATVIK DIXIT

MS Student in Electrical and Computer Engineering, Carnegie Mellon University
Email: satvikdixit@cmu.edu | Website: https://satvik-dixit.github.io/ | LinkedIn | Google Scholar | Github

EDUCATION

Carnegie Mellon University Pittsburgh, PA
Master of Science in Electrical and Computer Engineering Aug 2023 - Dec 2024 (Expected)

- **Research areas:** Audio and Speech Processing, Generative Models | **GPA:** 3.95/4.0
- **Advisors**: Dr Bhiksha Raj, Dr. Chris Donahue

Indian Institute of Technology (IIT) Delhi New Delhi, India
Bachelor of Technology in Electrical Engineering Aug 2019 - Aug 2023

- **Research areas:** ML, Signal Processing | **GPA:** 8.6/10.0

RESEARCH EXPERIENCE

Audio Language Models for Novel Sound Understanding

Research Assistant | Advisor: Professor Bhiksha Raj, CMU October 2024 - Present

- Developing methods to label unfamiliar sounds using fine-tuned audio language models (ALMs).
- Fine-tuning ALMs on large-scale audio datasets to extract binary characteristics (such as is it human generated? Is it rhythmic? etc.) for a given audio
- Extracting these characteristics for unfamiliar audio data and leveraging an external LLM (such as GPT-4o) to identify the most probable label using world knowledge

Text to Audio Morph Generation

Research Assistant |Advisor: Professor Chris Donahue, CMU July 2024 - Present

- Developing methods to combine multiple sound categories to generate novel hybrid sounds
- Extending text-to-audio models to support blending text and audio prompts using user-defined weights and temporal envelopes for enhanced user control
- Conducting a comprehensive audit of 15+ widely-used audio-text datasets to assess quality
- Evaluating dataset captions based on key factors, including accuracy, coverage, specificity, and clarity

Leveraging Audio To Evaluate Audio Captioning Systems

Research Assistant |Advisor: Professor Bhiksha Raj, CMU Sep 2024 - Oct 2024

- Developed MACE (Multimodal Audio Caption Evaluation) , the first metric to integrate both audio and reference captions for comprehensive audio caption evaluation
- Achieved state-of-the-art performance, with 3.28% and 4.36% relative improvements in accuracy over the widely-used FENSE metric on Clotho-Eval and AudioCaps-Eval benchmarks
- Submitted findings to ICASSP Speech and Audio Language Models Workshop 2025 [1][PDF][Code]
- Currently extending this approach to the domain of audio question answering

Evaluating Visual Language Models on Audio Spectrogram Classification

Research Assistant |Advisor: Professor Chris Donahue, CMU July 2024 - Sep 2024

- Developed a novel task VSC (Visual Spectrogram Classification) to evaluate the ability of Vision Language Models (such as GPT-4o, Claude, and Gemini) to classify audio using spectrogram images
- Benchmarked zero-shot and few-shot performance of state-of-the-art VLMs on the VSC task and performed ablation studies to optimise the spectrogram hyperparameters
- Conducted human studies to show VLMs display human-expert-level performance on the VSC task
- Accepted at Neurips Audio Imagination Workshop 2025 [2][PDF]

Improving Speaker Representations Using Contrastive Losses on Multi-scale Features

Research intern | Advisor: Professor Bhiksha Raj, CMU May 2024 - Sept 2024

- Designed the MFCon (Multi-scale Feature Contrastive) loss for speaker verification, achieving a 9.05% improvement in Equal Error Rate (EER) on the VoxCeleb-1O benchmark
- Demonstrated that explicitly enhancing the speaker separability of the intermediate feature maps by using contrastive losses, improves the discriminative ability of the final speaker embedding
- Conducted extensive ablation studies to identify optimal configurations and hyperparameters
- Submitted findings to ICASSP 2025 [3][PDF][Code]

Automatic Speech Recognition For Low Resource Languages

Research Assistant |Advisor: Professor Shinji Watanabe, CMU Jan 2024 - Mar 2024

- Added the ASR recipe for a Luganda (African dialect) dataset to the ESPNet toolkit (7k+ stars)
- Improved Word Error Rate (WER) by 28.3% over the baseline using a CTC-attention-based architecture with SpecAugment and speed perturbation techniques [PR]

Explaining Embeddings for Speech Emotion Recognition by Predicting Acoustic Features

Research Intern | Advisor: Dr. Satrajit Ghosh May 2022 - Aug 2023

- Developed a probing-based framework to explain pre-trained audio embeddings (e.g., WavLM) through acoustic features (e.g., eGeMAPS), evaluated on RAVDESS and SAVEE datasets.
- Created a novel metric, Information Increase, to quantify the contribution of specific acoustic features in the embedding and identified the key feature categories for each emotion
- Submitted findings to ICASSP 2025 [4][PDF][Code]

Room Acoustics Simulation

Research intern |Advisor: Dr. Robin Scheibler, EPFL June 2021 - Aug 2021

- Worked on developing Pyroomacoustics: an open-source package for room acoustics simulation
- Improved RIR simulation accuracy by adding a 'directivity' functionality to mics and sources [demo]

PUBLICATIONS & PREPRINTS

[1] Satvik Dixit, Soham Deshmukh, Bhiksha Raj. "MACE: Leveraging Audio for Evaluating Audio Captioning Systems." (under review at **ICASSP SALMA** Workshop 2025) [PDF]

[2] Satvik Dixit, Laurie Heller, Chris Donahue. "Vision Language Models Are Few-Shot Audio Spectrogram Classifiers." **NeurIPS Audio Imagination** Workshop 2024 [PDF]

[3] Satvik Dixit, Massa Baali, Rita Singh, and Bhiksha Raj. "Improving Speaker Representations Using Contrastive Losses on Multi-scale Features." (under review at **ICASSP** 2025) [PDF]

[4] Satvik Dixit, Daniel Low, Gasser, Fabio, Satrajit Ghosh. "Explaining Deep Learning Embeddings for Speech Emotion Recognition by Predicting Interpretable Acoustic Features." (under review at **ICASSP** 2025) [PDF]

TEACHING EXPERIENCE & REVIEWING

Teaching Assistant: Signals and Systems (18290) for Fall 2024 at Electrical and Computer Engineering, CMU

Teaching Assistant: Signals and Systems (18290) for Spring 2024 at Electrical and Computer Engineering, CMU

Reviewer: ICASSP Speech and Audio Language Models (SALMA) Workshop 2025

SKILLS

Programming Languages: Python, Java, LaTeX, Linux, MATLAB

Frameworks and Tools: PyTorch, Hugging Face, GCP, AWS, Git, CUDA, SpeechBrain, ESPNet

CMU Coursework: Speech Recognition and Understanding, Deep Generative Modeling, Advanced Natural Language Processing, Machine Learning (ML), Deep Learning, ML for Signal Processing