Soham Deshmukh

LinkedIn: sdeshmuk
Website: soham97.github.io

Interest

Audio Processing, Multimodal Learning, Generative AI

Current position

Senior Applied Scientist at Microsoft Speech team
Research used in Microsoft products – Microsoft Teams, Azure Edge, Video Translation, Outlook Copilot and more

Education

Carnegie Mellon University — Pittsburgh, USA
Ph.D. in Electrical and Computer Engineering — 2025 *(expected)*
- Research: Learning Audio Foundation Models for Reasoning
- Advisor: Prof. Bhiksha Raj

Carnegie Mellon University — Pittsburgh, USA
Masters in Electrical and Computer Engineering — 2019 - 2020
- Research: Self-supervised learning for sound event detection
- Advisor: Prof. Bhiksha Raj

Veermata Jijabai Technological Institute — Mumbai, India
Bachelors in Technology, Electronics Engineering — 2015 - 2019
- Research: Detecting harmful content in online conversations
- Advisor: Prof. Faruk Kazi

Experience

Senior Applied Scientist, Microsoft Speech team
Speech and audio processing — Aug 2024 - current

Applied Scientist 2, Microsoft Speech team
Audio understanding and ASR adaptation — Mar 2022 - Aug 2024

Applied Scientist, Microsoft NLP team
Task Oriented Dialogue Understanding — Jan 2021 - Mar 2022

Research Assistant, MLSP Group
Advisor: Rita Singh
Topic: Physics-based models for vocal fold parameter estimation — Aug 2020 - Dec 2020

Applied Scientist Intern, Microsoft Yammer
Feed Recommendation and Information Retrieval — May 2020 - Aug 2020

Research Assistant, MLSP Group
Advisor: Bhiksha Raj
Topic: Audio event classification and detection — Jan 2020 - May 2020

Undergraduate Research Assistant, CoE-CNDS Lab
Advisor: Faruk Kazi
Topic: Deepfake Detection — 2018 - 2019

Intern, Siemens R&D
Topic: Signal Processing for Predictive Maintenance — May 2018 - Aug 2018

Undergraduate Research Assistant, CoE-CNDS Lab
Advisor: Faruk Kazi
Topic: Detecting harmful content in online conversations — 2017 - 2018

Publications

Complete list of publications available at *Google Scholar*

1. ADIFF: Explaining audio difference using natural language
   Soham Deshmukh, Shuo Han, Rita Singh, Bhiksha Raj
   *International Conference on Learning Representations (ICLR) 2025 (**spotlight**)*

2. MACE: Leveraging Audio for Evaluating Audio Captioning Systems
   Satvik Dixit, **Soham Deshmukh**, Bhiksha Raj
   IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Speech and Audio Language Models (SALMA) workshop, 2025

3. Audio entailment: Assessing deductive reasoning for audio understanding
   Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, Bhiksha Raj
   *Association for the Advancement of Artificial Intelligence (AAAI) 2025 (**Oral**)*

4. Domain Adaptation for Contrastive Audio-Language Models
   Soham Deshmukh, Rita Singh, Bhiksha Raj
   Annual Conference of the International Speech Communication Association (INTERSPEECH) 2024

5. PAM: Prompting Audio-Language Models for Audio Quality Assessment
   Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, Huaming Wang
   Annual Conference of the International Speech Communication Association (INTERSPEECH) 2024

6. SELM: Enhancing Speech Emotion Recognition for Out-of-Domain Scenarios
   Hazim Bukhari, **Soham Deshmukh**, Hira Dhamyal, Bhiksha Raj, and Rita Singh.
   Annual Conference of the International Speech Communication Association (INTERSPEECH) 2024

7. Training Audio Captioning Models without Audio
   Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang
   IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2024

8. Multi-modal Language Models in Bioacoustics with Zero-shot Transfer: A Case Study
   Zhongqi Miao, Benjamin Elizalde, **Soham Deshmukh**, Justin Kitzes, Huaming Wang, Rahul Dodhia, Juan Lavista Ferres
   Scientific Report, Nature Portfolio

9. Natural Language Supervision for General-Purpose Audio Representations
   Benjamin Elizalde*, **Soham Deshmukh***, Huaming Wang.
   IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2024

10. Prompting Audios Using Acoustic Properties For Emotion Representation
    Hira Dhamyal, Benjamin Elizalde, **Soham Deshmukh**, Huaming Wang, Bhiksha Raj, Rita Singh
    IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2024

11. Pengi: An Audio Language Model for Audio Tasks
    Soham Deshmukh, Benjamin Elizalde, Rita Singh, Huaming Wang
    Conference on Neural Information Processing Systems (NeurIPS) 2023

12. Audio Retrieval with WavText5K and CLAP Training
    Soham Deshmukh, Benjamin Elizalde, Mahmoud Al Ismail, Huaming Wang
    Annual Conference of the International Speech Communication Association (INTERSPEECH) 2023

Publications

13. Multi-View Learning for Speech Emotion Recognition With Categorical Emotion, Categorical Sentiment, & Dimensional Scores.
    Daniel Tompkins, Dimitra Emmanouilidou, **Soham Deshmukh**, Benjamin Elizalde
    IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023
14. CLAP: Learning Audio Concepts from Natural Language Supervision
    Benjamin Elizalde, **Soham Deshmukh**, Mahmoud Al Ismail, Huaming Wang
    IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023
15. Improving weakly supervised sound event detection with self-supervised auxiliary tasks
    Soham Deshmukh, Bhiksha Raj, Rita Singh.
    Annual Conference of the International Speech Communication Association (INTERSPEECH) 2021
16. Interpreting glottal flow dynamics for detecting COVID-19 from voice
    Soham Deshmukh, Mahmoud Al Ismail, Rita Singh
    IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021
17. Detection of COVID-19 through the analysis of vocal fold oscillations
    Mahmoud Al Ismail, **Soham Deshmukh**, Rita Singh
    IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021
18. Temporal and Stochastic Modelling of Attacker Behaviour
    Rahul Rade, **Soham Deshmukh**, Ruturaj Nene, Amey Wadekar, Ajay Unny
    International Conference on Intelligent Information Technologies (ICIIT), 2019
19. Tackling Toxic Online Communication with Recurrent Capsule Networks
    Soham Deshmukh, Rahul Rade
    Conference on Information and Communication Technology (CICT), 2018

Patents

1. Training framework for automated tasks involving multiple machine learning models
   Charles Yin-che Lee, Ruijie Zhou, Neha Nishikant, **Soham Deshmukh**, Jeremiah D Greer
   US Patent, US-17/516940, 2023

Teaching

Teaching Assistant, Carnegie Mellon University
Course: Graph Signal Processing by José Moura 2024.08 - 2024.12

Teaching Assistant, Carnegie Mellon University
Course: Machine Learning for Signal Processing by Bhiksha Raj 2023.08 - 2023.12

Teaching Assistant, Carnegie Mellon University
Course: Introduction to Machine Learning by Gauri Joshi 2020.01 - 2020.05

Invited Talks

- Audio Foundation Models, Sphinx lunch, hosted by Shinji Watanabe [slides] 2024
- Towards zero-shot audio models, Robust MLSP, Carnegie Mellon University 2023
- Learning audio concepts from natural language supervision, Microsoft Research, Audio Group 2022
- Weakly and semi-supervised learning with its applications in audio and speech, Spoken Language Systems group (SLS), CSAIL, MIT 2020
- Attacker behaviour profiling and modelling framework for honeypot data: CoE-CNDS, ICICI bank, Cyber Peace Foundation 2018

Academic service

- **Organizer:**

  Workshop Speech and Audio Language Models (SALMA) at ICASSP 2025

  Special Session on Synergy between human and machine approaches to sound/scene recognition and processing at ICASSP 2023

- **Reviewer:**

  International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

  Conference of the International Speech Communication Association (INTERSPEECH)

  Conference on Neural Information Processing Systems (NeurIPS)

  International Conference on Learning Representations (ICLR)

  Detection and Classification of Acoustic Scenes and Events (DCASE)

  IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)

Research Advising

Interns advised at Microsoft

- Hira Dhamyal, PhD student, Carnegie Mellon University (co-advised with Benjamin Elizalde) $\overset{\text{next}}{\Longrightarrow}$ Siri, Apple
- Ruijie Zhuo, PhD student, University of California, Berkeley $\overset{\text{next}}{\Longrightarrow}$ Applied Scientist, Microsoft
- Neha Nishikant, Bachelors, Carnegie Mellon University $\overset{\text{next}}{\Longrightarrow}$ Data scientist, Palantir

Students mentored

- Satvik Dixit, Masters student, Carnegie Mellon University
- Shuo Han, Masters student, Carnegie Mellon University
- Hazim Bukhari, Masters student, Carnegie Mellon University

Open-source

1. **Models:** CLAP (500+ stars), Pengi (300+ stars), PAM (50+ stars)
2. **Datasets:** Audio Difference, Audio Entailment, Style transfer for Audio Captioning, WavText5K

Press coverage

- **Microsoft Unlocked** Audio AI used for bioacoustics in Amazon Rainforest
- **Microsoft Research Blog** Research on Automated Audio Captioning featured in Microsoft Research Blog
- **Analytics India Magazine 2023:** Microsoft launches Pengi, an Audio Language Model for Open-ended Tasks
- **Business Insider 2020:** Do I sound sick to you? Researchers are building AI that would diagnose COVID-19 by listening to people talk
- **Pittsburgh News 2020:** Coronavirus detected by voice? Carnegie Mellon researchers Develop app to 'listen' for signs of COVID-19
- **Forbes 2020:** AI and medical diagnostics: can a smartphone app detect COVID-19 from speech or cough?
- **Indiatimes 2020:** News coverage of Deepfake efforts in VJTI CoE-CNDS
- **CoE-CNDS 2019:** 4.49 Crore funding from MHA for AI Deepfake work and detection in the wild
- **DNIF newsletter 2019:** Modelling attacker behavioral patterns using statistical machine learning algorithms