

Akash Mahajan

Site: akashmjn.me; Redwood City, CA

[@akashmjn](https://twitter.com/akashmjn) [LinkedIn/akashmjn](https://www.linkedin.com/in/akashmjn)

EXPERIENCE

Member of Technical Staff / Tech Lead | [Contextual AI](#) 2024-2025

- Product development (0→1): RAG platform for knowledge agents
 - Built core [multimodal document understanding](#) powering context ingestion for retrieval
 - Critical in landing company's first multi-million \$ enterprise [contract with Qualcomm](#)
- Applied research: synthesis of long complex documents, eval design
 - Combining segmentation models, VLMs and parsers for high-fidelity OCR with bbox provenance
 - Token-efficient synthesis of million+ token context via ingest-time compute (à la [llm-wiki](#))
[\[Demo: Chat with 250 page PDF in Cursor\]](#) [\[Blog: Agentic alternative to GraphRAG\]](#)
- SWE things: workflow/agent framework architecture, testing, observability and scalability
- Tech Lead Manager: DRI cross-company; Mentored team of 3, interviewed candidates

Senior Applied Scientist | [Microsoft Azure Speech](#) 2018-2023

- Model development: state-of-art transcription designed for scale [$X * 1e7$ hrs/mo]
 - Shipped both batch and streaming models to [Azure Batch](#), [Microsoft Word](#), and [Microsoft Teams](#)
 - In particular: Optimized Conformer batch model ([Whisper-comparable](#)) at 50x realtime
- Applied research: diarized multi-speaker multi-mic transcription
 - Shipped diarized in-conference room transcription device covered by [The Verge](#)
 - Lead contributor: ASR training recipes, evaluation metrics, cross-system error analysis
- Research engineering: data pipelines, optimizing distributed training and inference
 - Speeding up $O(1e20)$ FLOP training on low-cost V100 GPUs
 - Leveraged NVIDIA/ONNX profiling tools to fix bottlenecks in inference throughput

PROJECTS

[tinydiarize: Lightweight extension of Whisper for diarization](#) | [\[github\]](#) 2023

- Built an extension of OpenAI's Whisper ASR model for speaker diarization with special tokens
- Released with [integration in whisper.cpp](#) (50k+ stars) — runnable on MacBooks/iPhones

[OrcaHello: AI-assisted 24x7 hydrophone monitoring](#) | [\[about\]](#) 2019-2022

- Co-founded hackathon project awarded a [\\$30k grant by Microsoft](#); real-time alert system listening for endangered orca calls 24/7 at underwater "hydrophones"; [covered by Mongabay News](#)

[Attention I'm trying to speak: Text to speech synthesis](#) | [Stanford NLP](#) | [\[github\]](#) 2018

- Built low-cost convolution-attention TTS trainable with \$75 of compute; awarded [best poster](#) in [CS224n](#)

EDUCATION

[Stanford University, M.S. Management Science & Engineering](#) 2016-2018

Deep Learning/Digital Signal Processing, Databases/Computer Systems, Marketing/Strategy/Design
CA (course assistant) for Machine Learning (CS229) & Deep Learning (CS230)

[Indian Institute of Technology \(IIT\), Madras, B.Tech.](#) 2011-2015

Chemical Engineering, minor: Control Systems (linear algebra, stats, signal processing)

MISC

Patents

- [US11044287B1](#): Network resilient real-time voice communication leveraging on-device speech models (granted during pandemic) Microsoft, 2021
- [WO2018020475A1](#): Electric vehicle drivetrain predictive health monitoring Ather Energy, 2018

Tech Stacks

- Contextual AI: Python/pydantic, HF Transformers, Gradio, vLLM, FastAPI/asyncio/[Temporal](#)
- Microsoft: Pytorch distributed, Azure ML/blob pipelines, ONNX/C++

Other

- Peer reviewer for ICASSP, TMLR, and NeurIPS, CVPR workshops 2025/26
 - Wrote [case studies](#) on the music streaming industry for strategy coursework at Stanford 2016/17
 - Organized [Chennai's largest EDM gig](#) (5k+ attendees) at IIT Madras 2014
-