Akash Mahajan

Site: akashmjn.me; Redwood City, CA @akashmjn LinkedIn/akashmjn

EXPERIENCE

Applied Researcher / Tech Lead | Contextual AI 2024-2025

- Product development (0→1): RAG platform for knowledge agents
  - Built core multimodal document understanding (parsing/OCR, representation) system powering document ingestion and retrieval
  - Critical in landing company's first multi-million $ enterprise contract with Qualcomm
- Applied research: Vision models, VLM workflow/agent framework, document hierarchy/metadata for retrieval, annotation/eval process & tooling
- Tech Lead Manager: roadmap and release planning; DRI with Forward Deployed/Eng/PM/Marketing; Mentored engineers/interns, interviewed candidates
- Links: Introducing the Document Parser for RAG, Demo: llms.txt for Documents

Senior Applied Scientist | Microsoft Azure Speech 2018-2023

- Model development: state-of-art speech transcription designed for scale [O(1e7) hrs/month]
  - Conformer-S2S with Whisper-comparable accuracy at 50x realtime on Azure Batch; streaming LSTM-based models for Microsoft Teams
- Applied research: data & training recipes, evaluation metrics, error analysis
  - Multi-speaker, multi-mic-array diarized (who-said-what) meeting transcription; Scalability-focused architecture design
- Research engineering: data pipeline, distributed training, optimizing training & inference
  - 2.5x speedup for training large-scale (800M+ param) models in <1 week on V100 GPUs; 2.6x speedup in S2S decoder time/token, 50-80% increase in 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 (38k+ stars) — runnable on MacBooks/iPhones

OrcaHello: AI-assisted 24x7 hydrophone monitoring | Microsoft Global Hackathon 2019-2022

- Co-founded award-winning project ($30k AI for Earth grant); system for 24/7 acoustic monitoring of endangered Orcas at multiple "hydrophone" locations; in the wild for >4 years [listen here]

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: Leveraging on-device speech models for resilient calling in poor network conditions Microsoft, 2021
- WO2018020475A1: Remote EV drivetrain predictive health monitoring via motor current frequency analysis Ather, 2018

Tech Stacks

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

Other

- Wrote case studies on the music streaming industry for strategy coursework at Stanford 2016/17
- Organized Chennai's largest EDM festival (5k+ attendees) at IIT Madras 2014