

Seunghyun Oh

Ph.D. Student · University of Washington · Seattle, WA

✉ shoh10@cs.washington.edu 🏠 ooshyun.github.io 📄 [Google Scholar](#) 🎧 [ooshyun](#) 🔗 [LinkedIn](#)

Summary

Ph.D. student who builds and evaluates proactive multimodal agents, with 5+ years of prior industry experience in on-device ML. Co-first author of Jarvis, a real-time agent that follows a group discussion against a shared document and speaks up only when the group misses or misstates a fact, evaluated on the CHI-180-proactive benchmark and in a live study with 23 participants. First author of Aurchestra (MobiSys 2026, Best Artifact Award), real-time soundscape control for hearables. I aim to extend this to agents that learn when and for whom to help over days of audio and video.

Research Interests

Proactive conversational agents, multi-party spoken interaction, evaluation of interactive agents, real-time and on-device audio AI

Education

University of Washington Ph.D. in Computer Science & Engineering, Paul G. Allen School (Advisor: Prof. Shyamnath Gollakota) · Expected graduation: Jun 2030	Sep 2025 – Present Seattle, WA
Hanyang University Master of Engineering in Electronics and Computer Engineering, GPA 3.8/4.0	Mar 2018 – Feb 2020 Seoul, Korea
Inha University Bachelor of Engineering in Information and Communication Engineering, GPA 3.3/4.0	Mar 2012 – Feb 2018 Incheon, Korea

Papers

Jarvis: A Proactive Speech Agent for Multi-Party Conversations

Seunghyun Oh*, Hirotaka Hiraki*, Shuyue Stella Li, Yulia Tsvetkov, Shyamnath Gollakota. arXiv preprint [arXiv:2610.07506](https://arxiv.org/abs/2610.07506), 2026. (*Equal contribution)

Fine-grained Soundscape Control for Augmented Hearing (Aurchestra)

Seunghyun Oh, Malek Itani, Aseem Gauri, Shyamnath Gollakota. The 24th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2026). **Best Artifact Award**. DOI: [10.1145/3745756.3809210](https://doi.org/10.1145/3745756.3809210)

Trigger Sound Suppression for Misophonia

Vaishnavi Vidyasagar, Jasmine Zhang, Mahima Uliyar, Seunghyun Oh, Emily Catherine Gates, Mark Zachary Rosenthal, Shyamnath Gollakota. arXiv preprint [arXiv:2609.36351](https://arxiv.org/abs/2609.36351), 2026.

Selected Projects

Jarvis: Proactive Speech Agent for Multi-Party Conversations [\[site\]](#)

Jan 2026 – Sep 2026

Summary — First real-time proactive speech agent that audibly joins multi-party conversations and speaks up only when a document-grounded fact is missed or misstated

- Results: live study with 23 participants, 163 of 164 interventions correct or partially correct; on CHI-180-proactive, 95–97% correct or partially correct and silent on 97% of group self-corrections
- Built CHI-180-proactive (180 CHI papers, synthesized group discussions, 4,320 planted events), which scores both when to speak (missed or misstated facts) and when to stay silent (group self-corrections)
- Floor-taking: waits for an opening, handles barge-in, raise-hand mode, shows cited evidence
- Self-hosted backbone: open-weight Qwen3.6-35B-A3B (3B-active MoE) with deterministic checks, served with vLLM on 1–8 GPUs (Slurm), 1–2 s decision latency in live sessions

Summary — First system enabling real-time, multi-class soundscape control on resource-constrained hearables, letting users independently mix sounds in their environment

- User studies: listening (n = 17), smartphone UI (n = 7), preference survey (Vercel, Supabase)
- Designed multi-output target sound extraction (TSE) network with dual-path time–frequency backbone and FiLM class conditioning, extracting up to 5 overlapping targets
- Fine-tuned Audio Spectrogram Transformer (AST) for 20-class sound event detection
- Hardware-optimized networks run in real time: 4.47–5.23 ms per 6 ms audio chunk on Raspberry Pi 4B, Orange Pi 5B and GAP9

Work Experience

Cochl (Sound AI Startup)

Jul 2023 – Aug 2025

Backend Engineer, SDK

Seoul, Korea

- Developed Sound AI SDK for cross-platform application: x86, ARM (armv7, aarch64), Android NDK
- Built acceleration backends for AI inference: TensorRT (NVIDIA Jetson), Qualcomm NPU, Edge TPU
- Ported the SDK to MCU/NPU clients: STM32 (U5, U3), Arm Ethos-U55, Syntiant NDP120, ESP32-S3
- Dockerized dev environment (x86, aarch64, armv7hf, armv7hl; C++, Python, Android apps) and built CI/CD pipeline for build, on-device testing, benchmark and SDK release
- Tools: C++, C, Python, TensorFlow, TensorFlow Lite, ONNX, TensorRT, QNN, SNPE, Shell, Jenkins, GitOps, AWS S3, AWS ECR, Docker

Freelance

Apr 2023 – Jul 2023

Embedded Machine Learning Engineer

Seoul, Korea

- Designed and trained deep learning model for speech enhancement for STM32F746VE

Olive Union (Hearing aid)

Apr 2020 – Apr 2023

Embedded DSP Engineer

Seoul, Korea

- Optimized digital filter implementation using SIMD on Tensilica HiFi DSP, reducing cycles by 72% and memory by 85%, improving battery life by 35 minutes
- Built a fixed-point virtual hardware platform (CMSIS-DSP) to evaluate DSP algorithms
- Designed the MCU–DSP communication protocol

Skills

ML & LLM Serving: PyTorch, vLLM, SGLang (Qwen3 serving), TensorFlow, TensorFlow Lite, ONNX, TensorRT, Qualcomm QNN

Programming Languages: Python, C, C++, CMake, Bash, LaTeX

Systems: Docker, AWS (EC2, S3, ECR), GCP, Jenkins, Git

Embedded: NVIDIA Jetson, Edge TPU, STM32, ESP32-S3, Arm Ethos-U55, Syntiant NDP120, HiFi DSP

Language: Korean (Native), English (Proficient)

Honors & Self-Directed Study

Best Artifact Award, MobiSys 2026

Jun 2026

Top-5 Ranking, ICASSP 2023 SP Clarity Challenge

Jan 2023 – Feb 2023

TinyML Study

Sep 2023 – May 2024

- Studied and organized AI model compression from MIT TinyML and Efficient Deep Learning Computing
- **Outcome:** [TinyML KOR Study Archive](#)