Taras Sereda. ML Researcher & Engineer

Machine Learning Researcher and
Engineer with 10+ years of expertise
in audio, vision, text processing.
Developed state-of-the-art ASR, TTS,
and computer vision systems.
Passionate about pushing the
boundaries of human knowledge.

Research focus: Machine Learning for systems.

Contacts and links

taras.y.sereda@proton.me

https://taras-sereda.github.io

github.com/taras-sereda

Publications and patents

• <u>Speeding up PyTorch inference by 87% on Apple devices with AI-generated Metal kernels</u>. 2025

Taras Sereda, Natalie Serrino, Zain Asgar

- <u>Pheme: Efficient and Conversational Speech Generation</u>. 2024 Budzianowski Paweł, Sereda Taras, Cichy Tomasz, and Vulić Ivan
- <u>Transcribe</u>, <u>Align and Segment: Creating speech datasets for low-resource languages</u>. 2024
 Sereda Taras
- <u>System and method for simultaneous multilingual dubbing of video-audio programs</u>. Patent number: 11942093
 Aleksandr Dubinsky, Taras Sereda. 2024

Talks

- <u>How Good is AI at Generating AI Kernels?</u> Talk at AI Benchmark Club, 2025. San Francisco, CA.
- <u>VocalicsAI. Artificial intelligence removes language barriers</u> Talk at FW days, 2021. Kyiv, Ukraine
- <u>Waveglow. Generative modeling for audio synthesis</u> Talk at FW days, 2019. Kyiv, Ukraine

Technical skills

- Programming: Python, Rust, CUDA C, Java
- ML/DL Frameworks: PyTorch, NVIDIA NeMo, Hugging Face, TensorFlow

- ML Techniques: self-supervised learning, CNNs, RNNs, Transformers, GANs
- Data processing & visualisation: numpy, scikit-learn, matplotlib, seaborn, plotly
- MLOps: Docker, Kubernetes, CI/CD, MLflow, Weights & Biases
- Databases: PostgreSQL, MongoDB
- Cloud computing: Amazon AWS, Google GCP

Leadership skills

resilience, strategic thinking and vision

Professional Experience

ML Researcher

<u>Gimlet Labs</u> (Nov 2024 - present)

- Leading research on LLM-powered kernel generation for GPU/accelerator optimization.
- Developing k-forge, an AI agent that autonomously generates and optimizes low-level kernels directly from PyTorch. Focusing on accelerations of both training and inference workloads across backends CUDA, ROCm, and Metal.

ML Consultant

Self-employed (Sep 2022 - present)

- Developed a real-time ASR system for low SNR environments (-10 to 0 dB), outperforming open-source SOTA models on domain-specific data
- Collaborated with PolyAI to create a zero-shot TTS system for call center automation, achieving faster-than-real-time performance on A100 GPUs

Visiting Scholar

Ukrainian Catholic University (Mar 2022 - present)

- Deliver lectures on speech and audio processing, focusing on TTS, ASR, and source separation(SS)
- Mentor and advise students on academic challenges in ML and audio processing
- Developed course material: github.com/taras-sereda/deep-learning-for-audio

Co-founder, Director of Research

Vocalics.ai Kyiv, Ukraine - (Feb 2019 - Sep 2022)

• Led R&D for a novel multi-lingual speech synthesis system preserving speaker identity and style

- Managed a team of 4 researchers, driving innovation in SOTA speech synthesis technologies
- Implemented and customized multiple SOTA papers in speech synthesis, including autoregressive and parallel TTS approaches
- Developed methodology for quality evaluation of generated speech in dimensions of intelligibility, speaker and prosody similarity
- Conducted customer interviews and market research to validate product-market fit

ML Engineer

Whisper.ai San Francisco, CA, USA - (Feb 2018 - Jul 2018)

- Developed advanced audio source separation models (PIT, Chimera network) deployed on ARM devices
- Improved listening experience for those with mild hearing loss in noisy environments

ML Researcher

Ring Labs Kyiv, Ukraine - (Sep 2016 - Apr 2017)

- Established and led the ML department as the first employee in the Ukrainian R&D office
- Developed object detection algorithms based on R-CNN and YOLO with customized in-house implementations

ML Engineer (Part time)

IPGraphy Kyiv, Ukraine - (Oct 2015 - Jun 2016)

- Developed core algorithm for visual object similarity search applied to trademark images
- Created an IP rights NN-based tool to enhance attorney productivity

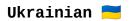
ML Engineer

```
DepositPhotos Kyiv, Ukraine - (Dec 2014 - Dec 2015)
```

Developed core algorithms for a virtual stylist app, including:

- Recommender system for clothing combination suggestions
- Neural networks for automatic clothes classification and wardrobe categorization

Languages



Native speaker



Proficient speaker

Education

Masters's Degree in Mathematical Modelling

2011 - 2012

KNEU - Kyiv, Ukraine

Bachelor's Degree in Cybernetics

2007 - 2011

KNEU - Kyiv, Ukraine