Karim Ali

Machine Learning Engineer
kareem.ali.ali.99@gmail.com | +381 61 6295 274 | Belgrade, Serbia | linkedin.com/in/karim-alia0a21414b | github.com/karimkalimu

Skills

Audio ML: Speech Synthesis, Automatic Speech Recognition, Voice Cloning, Voice Conversion, Source Separation, Vocal Separation, Target Speech Extraction, Music Transcription, CTC Alignment, Model Evaluation
ML Engineering: Python, PyTorch, CUDA, ONNX, TensorRT, TFLite, CoreML, Hugging Face
Backend and Infrastructure: FastAPI, Flask, Redis, RabbitMQ, PostgreSQL, Docker, Object Storage, Data Engineering, GPU Inference Pipelines, Async Worker Orchestration, Queue Based Processing, Model Serving
Real Time and Product: LiveKit, Azure Speech Services, Speech API Integration, LLM Integration, Swift, Flutter, Electron
Audio and Media Tooling: FFmpeg, Librosa, SoundFile, Pydub, C++

Experience

Machine Learning Engineer - NetEase Youdao - Remote
January 2025 - Present
- Engineered real time voice agent services on LiveKit, covering token generation, Python worker orchestration, role based assistant behavior, and provider integration.
- Integrated speech and LLM services across synthesis, recognition, translation, and conversational workflows using internal services, Azure Speech, and custom platform extensions.
- Contributed backend services, framework packaging, and release integration for shipped consumer audio apps, including SingUp, Clear AI Audio Filter, and AI Remix.

Machine Learning Engineer - NetEase Youdao - Remote
December 2021 - December 2024
- Built and operated the GPU backend for SingUp custom voice and AI cover generation, including training workers, inference workers, model storage, and callbacks.
- Deployed Python, PyTorch, FastAPI, Redis, FFmpeg, and object storage integrations across multi GPU server environments.
- Designed GPU serving workflows for high load voice conversion and cover generation, including multi worker utilization, queue routing, and memory aware scheduling.
- Trained and adapted voice models from user audio using vocal cleaning, source separation, augmentation, and fine tuning from pretrained checkpoints.
- Implemented backend workflows for queue driven processing, worker orchestration, status tracking, uploads, callbacks, and production runbooks.
- Created evaluation tools and test environments for separation, target speech extraction, speech synthesis, and music generation experiments.

Algorithm Engineer - Kuai Peilian (快陪练) - Beijing, China
March 2020 - September 2021
- Developed algorithm workflows for automatic piano performance assessment using music transcription, finger tracking research, and audio alignment.
- Trained and evaluated models using preprocessed datasets and notebook based experimentation workflows.
- Implemented spectrogram alignment with dynamic programming for comparing performance audio.
- Implemented a C++ SDK around a TFLite model for real time piano transcription on iOS and Android.

Algorithm Engineer - Kuai Peilian (快陪练) - Beijing, China
June 2019 - January 2020
- Supported 2+ internal algorithm prototypes for music education and piano practice technology, covering transcription, scoring logic, and performance analysis experiments.

Algorithm Engineer - Music++ - Beijing, China
January 2019 - May 2019
- Built and evaluated 2+ early algorithm prototypes for music technology applications, focusing on data preprocessing, training experiments, and inference testing.

Projects

Real Time Voice Assistant Platform - LiveKit, Python, Flask, Azure Speech, Speech APIs, LLM Services

- Implemented server and worker components for real time voice chat, translation, token generation, and assistant orchestration.
- Customized LiveKit integrations and unified multiple speech and LLM providers behind one real time workflow.

Production Voice Conversion and Separation Stack - Python, PyTorch, FastAPI, Redis, PostgreSQL, RabbitMQ, FFmpeg
- Developed service layers for voice conversion, voice cloning, vocal separation, source separation, job queues, GPU workers, and object storage uploads.
- Maintained deployment workflows, worker status flows, API contracts, and operational tooling for production audio services.

Dubbing and Prompt TTS Data Pipeline - FastAPI, PostgreSQL, CTC Alignment, Silero VAD, Object Storage
- Designed media processing pipelines for ingestion, separation, diarization, subtitle extraction, translation, speech synthesis, timestamp refinement, and silence reduction.
- Implemented pipeline services with async workers, database coordination, concurrency limits, and artifact handling flows.

SingUp GPU Voice Backend and Consumer Audio Releases - Python, PyTorch, FastAPI, Redis, RabbitMQ, GPU Inference, Voice Training, Object Storage
SingUp | Clear AI Audio Filter | Easy Chord | AI Remix
- Built GPU training and inference support for SingUp, a public App Store AI music app with 15K ratings, covering custom voice training, AI cover generation, model download, and result callbacks.
- Implemented Redis and RabbitMQ worker pipelines for audio download, vocal cleaning, separation, RVC training, generated audio upload, status tracking, and multi GPU deployment.

Independent App Store Releases
Igramo Jamb | Hora
- Published independent iOS apps on the App Store, including Igramo Jamb and Hora.

Model Conversion and Fine Tuning Workflows - PyTorch, ONNX, TensorRT, ZipVoice, CosyVoice, Model Optimization
- Evaluated and adapted speech synthesis, source separation, voice conversion, and multilingual fine tuning workflows for Arabic, Serbian, and audio product experiments.
- Created demos, data preparation scripts, export paths, and evaluation helpers for testing research models in practical environments.

Education

Master's degree, Computer science - University of Science and Technology Beijing - Beijing, China
Aug 2018 - Dec 2020

HSK, Chinese Language - University of Science and Technology Beijing - Beijing, China
Aug 2017 - May 2018

Bachelor of Science - BS, Computer Engineering - Lebanese International University
2012 - 2016