+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on 03/11/2026

Sign up

Training Moshi & S2S Training: Real-time Voice AI

2 days (14 hours)

Overview

Moshi is an open-source native speech-to-speech (S2S) generator that represents a technological breakthrough in conversational AI. This modern approach is based on an end-to-end paradigm for designing fluid, fast voice assistants capable of managing emotions in real time.

Our Moshi & S2S training will help you understand the Native Speech-to-Speech architecture, master the Mimi engine, optimize inference with Rust, and deploy your projects via WebSockets or gRPC streams. You will also learn to manage full-duplex interactivity, reduce critical latency, and ensure the security of audio streams.

By the end of the training, you will be able to create, deploy, and maintain high-performance voice AI systems, understand their advantages over traditional cascading (ASR+LLM+TTS), and industrialize your workflows for professional projects.

Like all our training courses, this one is based on the latest stable version of Moshi and emphasizes a practical, hands-on approach.

Like all our training courses, this one will introduce you to **the latest stable version** of the technology and its new features.

Objectives

- Understand the Native S2S philosophy and the architecture of the Mimi engine.
- Deploy and evaluate Moshi models for local inference.
- Master interrupt handling and full-duplex flow management.
- Optimize performance (RTF) and VRAM consumption.

- Integrate a complete S2S stack into a web interface.

Target Audience

- AI / Machine Learning Developers
- Backend and Systems Engineers
- Technical project managers

Prerequisites

- Proficiency in Python and a basic understanding of PyTorch
- Knowledge of Deep Learning (Transformers)
- Familiarity with Linux and the command line

Software Requirements

- At least 16 GB of RAM, NVIDIA GPU with 12 GB of VRAM
- Linux (Ubuntu recommended) or Windows with WSL2
- A terminal with GPU support (NVIDIA Container Toolkit)
- A code editor and Python 3.10+

Moshi & S2S Training Program

[Day 1 - Morning]

The Native Speech-to-Speech Architecture

- The Moshi Revolution: Why the "End-to-End" Paradigm Is Replacing Cascading (ASR+LLM+TTS)
- The Mimi Engine: How the neural codec and audio codebooks work
- Latency and Streams: Understanding the Joint Prediction of User and Model Streams
- Environment Setup: Installing PyTorch, Rust/Candle, and Quantized Models
- Hands-on Workshop: Local Deployment and First Voice Exchange with VRAM Monitoring.

[Day 1 - Afternoon]

Mastering Interactivity and Full-Duplex

- Interruption Management: Concurrent Speech Detection Algorithms Without Traditional VAD
- Audio Token Streaming: Manipulating Output Streams in Real Time
- Emotions and Prosody: Controlling voice style and audio sampling rate
- Stream Stability: Buffer Management to Prevent Glitches

- Hands-on workshop: Developing an assistant that can be interrupted cleanly in the middle of a sentence.

[Day 2 - Morning]

Inference Optimization (Performance-Critical)

- Moshi-backend (Rust): Using the high-performance implementation for production
- Quantization and Precision: Trade-offs between 4-bit, 8-bit, and bf16 depending on the hardware
- KV-Caching Audio: Optimizing memory for long conversations
- Latency Profiling: Identifying Bottlenecks (Mimi Decoding vs. LLM Inference)
- Hands-on Workshop: Comparative Benchmarking and Optimization of RTF (Real-Time Factor).

[Day 2 - Afternoon]

Industrialization and System Integration

- Server Architecture: Setting up a WebSocket/gRPC tunnel for raw audio
- Web/Client Interface: Microphone capture and PCM speaker playback via browser
- S2S Observability: Metrics for jitter, packet loss, and perceived quality
- Security and Ethics: Stream encryption and audio privacy management
- Hands-on Workshop: Building a complete stack (S2S Backend + Interactive Web Frontend).

Target Audience

This training is intended for both individuals and companies, large or small, wishing to train their teams in new advanced IT technologies or to acquire specific professional knowledge or modern methods.

Entry-level assessment

The pre-training assessment complies with Qualiopi quality standards. Upon final registration, the learner receives a self-assessment questionnaire that allows us to evaluate their estimated proficiency in various types of technologies, as well as their expectations and personal goals regarding the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could pose challenges for monitoring and ensuring the smooth running of the training session.

Teaching Methods

Practical Training: 60% practical, 40% theoretical. Training materials will be distributed in digital format to all participants.

Organization

The course alternates between theoretical input from the instructor, supported by examples and

reflection sessions and group work.

Assessment

At the end of the session, a multiple-choice questionnaire verifies that the skills have been properly acquired.

Certification

A certificate will be issued to each trainee who has completed the entire training program.