

Mis à jour le 10/03/2026

S'inscrire

Formation Formation Moshi & S2S : Speech-to-Speech haute performance

Présentation

Moshi est un générateur de parole natif (Speech-to-Speech) open source qui s'inscrit dans une rupture technologique de l'IA conversationnelle. Cette approche moderne repose sur un paradigme End-to-End pour concevoir des assistants vocaux fluides, rapides et capables de gérer les émotions en temps réel.

Notre formation Moshi & S2S vous permettra de comprendre l'architecture Native Speech-to-Speech, de maîtriser le moteur Mimi, d'optimiser l'inférence avec Rust et de déployer vos projets via des flux WebSockets ou gRPC. Vous apprendrez également à gérer l'interactivité Full-Duplex, à réduire la latence critique et à assurer la sécurité des flux audio.

À l'issue de la formation, vous serez en mesure de créer, déployer et maintenir des systèmes d'IA vocale performants, de comprendre leurs avantages face au cascading traditionnel (ASR+LLM+TTS) et d'industrialiser vos workflows pour des projets professionnels.

Comme toutes nos formations, celle-ci s'appuie sur la dernière version stable de [Moshi](#) et privilégie une approche pratique et opérationnelle.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Comprendre la philosophie du Native S2S et l'architecture du moteur Mimi.
- Déployer et quantifier des modèles Moshi pour l'inférence locale.
- Maîtriser la gestion des interruptions et du flux Full-Duplex.
- Optimiser les performances (RTF) et la consommation VRAM.
- Intégrer une stack complète S2S dans une interface Web.

Public visé

- Développeurs IA / Machine Learning
- Ingénieurs Backend et Systèmes
- Chefs de projet technique

Pré-requis

- Maîtrise de Python et bases de PyTorch
- Connaissances en Deep Learning (Transformers)
- Familiarité avec Linux et la ligne de commande

Pré-requis logiciels

- 16 Go de RAM au minimum, GPU NVIDIA avec 12 Go de VRAM
- Linux (Ubuntu recommandé) ou Windows avec WSL2
- Un terminal avec support GPU (NVIDIA Container Toolkit)
- Un éditeur de code et Python 3.10+

Formation Moshi & S2S : Déployer l'IA Conversationnelle Temps Réel

[Jour 1 - Matin]

L'architecture Native Speech-to-Speech

- Révolution Moshi : Pourquoi le paradigme "End-to-End" supprime le cascading (ASR+LLM+TTS)
- Le moteur Mimi : Fonctionnement du codec neural et des codebooks audio
- Latence et flux : Comprendre la prédiction conjointe des flux utilisateur et modèle
- Setup Environnement : Installation PyTorch, Rust/Candle et modèles quantifiés
- Atelier pratique : Déploiement local et premier échange vocal avec monitoring VRAM.

[Jour 1 - Après-midi]

Maîtrise de l'Interactivité et du Full-Duplex

- Gestion des interruptions : Algorithmes de détection de parole concurrente sans VAD classique
- Streaming de tokens audio : Manipulation des flux de sortie en temps réel
- Émotions et Prosodie : Contrôler le style de voix et la température de sampling audio
- Stabilité du flux : Gestion des buffers pour éviter les craquements (glitches)
- Atelier pratique : Développement d'un assistant capable d'être interrompu proprement en pleine phrase.

[Jour 2 - Matin]

Optimisation de l'Inférence (Performance Critique)

- Moshi-backend (Rust) : Utiliser l'implémentation performante pour la production
- Quantification et Précision : Arbitrage entre 4-bit, 8-bit et bf16 selon le hardware
- KV-Caching Audio : Optimiser la mémoire pour les conversations longues
- Profiling de latence : Identifier les goulots (décodage Mimi vs Inférence LLM)
- Atelier pratique : Benchmarking comparatif et optimisation du RTF (Real-Time Factor).

[Jour 2 - Après-midi]

Industrialisation et Intégration Système

- Architecture Serveur : Mise en place d'un tunnel WebSocket / gRPC pour l'audio brut
- Interface Web/Client : Capture micro et restitution haut-parleur en PCM via navigateur
- Observabilité S2S : Métriques de gigue (jitter), perte de paquets et qualité perçue
- Sécurité et Éthique : Chiffrement des flux et gestion de la confidentialité audio
- Atelier pratique : Création d'une stack complète (Backend S2S + Frontend Web interactif).

Sociétés concernées

Cette formation s'adresse à la fois aux particuliers ainsi qu'aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte

des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.