MAI-Voice-2 Model Card

Date: June 2, 2026

Model summary

The next iteration in our speech synthesis family, MAI-Voice-2 is a prompted text-to-speech (TTS) model that generates high-fidelity, natural, and expressive speech across 15 languages. It captures human-like intonation, rhythm, and emotional nuance, enabling engaging and lifelike conversational experiences.

About this model

Voice can be configured using: - Curated voice library (licensed voices)
- Voice prompting via short audio clips (5-60 seconds)

Voice prompting safeguards

To ensure voice prompting is compliant, we adhere to Microsoft's Responsible AI policies and apply the following steps to ensure voice cloning is conducted lawfully.

Steps to Access Personal Voice (Voice Cloning) with MAI-Voice-2: - Apply for gated access — Submit a request via the Azure AI Custom Neural Voice and Custom Avatar Limited Access Review. - Access personal voice APIs — Once approved, use the APIs available at cognitive-services-speech-sdk/samples/custom-voice - Upload audio consent and prompt — Provide a recorded audio consent statement from the voice talent along with the required prompt to create a personal voice profile. For more information, read about Azure Speech consent statement.
- Synthesize text using the created voice

Key capabilities

Key model capabilities

- High fidelity natural voice synthesis
- Voice prompting with improved quality and stability. Provide few seconds of an audio clip (5-60 seconds) and the model clones it instantly. No fine-tuning required allowing you to onboard a

consented voice of your choice easily. Access requires Microsoft approval and guardrails are in place to avoid misuse.
- Fine-grained control of tone and delivery: Shape delivery at the turn/sentence level by controlling the emotion and tone of the output.
- Long-form speech generation: Built for extended content covering audiobooks, lectures, podcasts, training materials, and long-form narration.
- Multilingual support across 15 languages and 18 locales.
- Multi-speaker generation support

Together, these capabilities give developers the building blocks to ship voice at scale, across accessibility, virtual assistants, media narration, and customer service

---

Use cases

Key use cases

- **Media: Entertainment** – Generate expressive voices for games, films, podcasts, and immersive experiences.
- **Virtual Assistants and Chatbots** – Power conversational agents across apps and devices with natural voices.
- **Accessibility Features** – Provide narration for visually impaired users and assistive voice technologies.
- **Educational Experiences** – Build interactive learning content with expressive narration.
- **Marketing and Advertising** – Deliver consistent voice experiences across campaigns.
- **Self-authored Content** – Turn written content into spoken audio using custom voice characteristics.
- **IVR Systems** – Enable natural, expressive call center interactions.
- **Public Announcements** – Deliver clear, engaging voice output for public information systems.

Out of scope use cases

Model prioritizes naturalness and expressivity over latency-critical scenarios.

Pricing

$22 per 1M characters

Input formats

Transcript text + speaker prompt audio + optional annotations.

Supported languages

English(US), English (Australia), Hindi, French, German, Italian, Portuguese (Brazil), Hindi, Spanish (Spain), Spanish(Mexico), Korean (South Korea), Mandarin, Russian, Thai, Dutch, Turkish, Romanian, Hungarian.

Output

24kHz mono audio.

Sample JSON response

N/A

Distribution

Available via Azure Speech SDK and REST APIs.

MAI-Voice-2 can be accessed globabally. The model is currently served from three Azure regions

- East US (EUS)
- Sweden Central (SEC)
- SEA=Southeast Asia (SEA)

to which the requests are routed.

More information

Learn more in Azure AI Speech documentation.