MAI-Voice-1 Model Card

Date: April 2, 2026

Model summary

MAI-Voice-1 is a text-to-speech(TTS) model that generates high-fidelity, natural, and expressive speech. It captures human-like intonation, rhythm, and emotional nuance, enabling engaging and lifelike conversational experiences. It strictly follows the provided transcript and supports per-turn emotion control

About this model

There are two ways to set the voice for your project.

Curated voice library:

Licensed voices designed to work straight out of the box.

Voice prompting:

Provide a few secs long audio clip with your request and the model matches it instantly.

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-1:

1. *Apply for gated access* — Submit a request via the Azure AI Custom Neural Voice and Custom Avatar Limited Access Review.
2. *Access personal voice APIs* — Once approved, use the APIs available at cognitive-services-speech-sdk/samples/custom-voice
3. 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.
4. Synthesize text using the created voice

Key capabilities

Key model capabilities

1. **High fidelity Natural Voice Synthesis -**Produces voice with the intonation, rhythm, and emotional range of a real speaker.
2. **State–of-the-Art Voice Prompting -** Provide few seconds of an audio clip(up to 120secs) 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.
3. **Fine grained control -** Shape delivery at the turn/sentence level by controlling the emotion and tone of the output.
4. **Long-form content-** Built for extended content covering audiobooks, lectures, podcasts, training materials, and long-form narration.

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

Use cases

Text to speech offers a variety of features catering to a wide range of intended uses across industries and domains. All text to speech features are subject to the terms and conditions applicable to customers' Azure subscription, including the Azure Acceptable Use Policy and the Code of conduct for Azure AI Speech text to speech.

Key use cases

- **Media: Entertainment** - Give characters a voice. Generate expressive, lifelike audio for games, films, podcasts, audiobooks, and immersive AR/VR experiences.
- **Virtual Assistants and Chatbots** - Make your assistant sound like it belongs in your product. Power conversational agents across apps, vehicles, appliances, and customer service with a branded voice.
- **Accessibility Features** - Build products that more people can use. Add audio narration for visually impaired users and voice support for individuals with speech impairments.
- **Educational and Interactive Learning** - Build character and brand voices for online courses, interactive lessons, simulations, and guided tours.
- **Media: Marketing and Advertising** - Develop a consistent, recognizable voice across product launches, campaigns, and ads.

- **Self-authored Content** - Voice talent can bring blogs, books, social media content, and personal stories to life using a custom voice built from their own.
- **Interactive Voice Response (IVR) Systems** - Build dynamic, natural and expressive voices for call centers and automated phone interactions.
- **Public Service and Informational Announcements** - Deliver clear and engaging voice messages for public venues, traffic updates, weather alerts, event information, and schedules.

Out of scope use cases

Usage will be restricted to use the service in any way that is inconsistent with the Code of Conduct

Pricing

Amongst HD voices, MAI-Voice-1 is available at a competitive rate of $22.00/1M chars.

Input formats

Plain text or Speech Synthesis Markup Language (SSML), which supports emotion control.

Supported language

English (soon expanding to 10+ languages).

Supported Azure regions

For now available in Central US, Japan West and Sweden Central. Expanding to more regions soon.

Sample JSON response

Endpoint	Request Type	Response Format
POST /cognitiveservices/v1	SSML + headers	Binary audio file (MP3 / WAV / Opus / etc.)
Speech SDK SpeakTextAsync	Text or SSML	SDK stream + result metadata
Batch synthesis API	Long-form SSML/Text	Asynchronous job → downloadable audio file

Distribution

MAI-Voice-1 is available through the following methods to support a wide range of integration scenarios:

- **Speech SDK** Integrate TTS capabilities directly into applications using Azure's Speech SDK, available for platforms including .NET, Python, Java, JavaScript, and C++.
- **REST API** Access TTS functionality via a public, subscription-based API for flexible integration into web services, mobile apps, and backend systems.

More information

Learn more in the full Azure AI Speech Service documentation.