Agora and Gradium Announce Strategic Partnership to Bring Ultra-Low Latency Text-to-Speech to Conversational AI Engine

July 8, 2026

Gradium's industry-leading text-to-speech technology will be available in Agora Conversational AI Engine later this month, giving developers another high-performance option for building natural, low-latency voice AI experiences.

SAN JOSE, CA, Agora, Inc. (NASDAQ: API), the leading platform for real-time engagement, and Gradium, today announced a strategic partnership to bring Gradium's Text-to-Speech (TTS) technology directly into Agora Conversational AI Engine.

Developers building voice AI agents on Conversational AI Engine will be able to enable Gradium's TTS with a single configuration—no separate accounts, no additional orchestration layer, no extra latency hops.

"Conversational AI Engine is where developers come to build production-grade voice agents, and the quality of the voice layer is everything. By integrating Gradium's ultra-low latency TTS with Agora's platform, we're giving developers another high-quality voice option that helps them deliver responsive, lifelike conversational experiences with less complexity."

— Patrick Ferriter, SVP Product, Agora

"Agora gives voice agent developers global real-time infrastructure. Pairing that with Gradium TTS inside Conversational AI Engine means teams get low-latency speech synthesis without building the transport layer themselves."

— Neil Zeghidour, CEO, Gradium

Expanding the Conversational AI Ecosystem

The upcoming integration with Gradium expands the growing ecosystem of AI providers available through Agora. Beginning later this month, developers will be able to combine Agora's real-time infrastructure with Gradium's low-latency text-to-speech technology to build more responsive conversational AI experiences while choosing the providers that best meet their performance and deployment requirements.

According to Gradium, its streaming TTS delivers a median time-to-first-audio (P50) of approximately 155 milliseconds under concurrent load. The integration will also provide access to voice cloning and deployment options spanning cloud, private cloud, and on-premises environments.

Agora's global real-time network, with 200+ points of presence across six continents, provides Gradium's TTS a distribution channel and infrastructure layer it would otherwise take years to build independently. Developers building conversational AI capabilities with Agora are exactly the audience Gradium was built for—teams shipping production voice agents who need every millisecond and won't tolerate variable latency under load.

What Developers Get

Developers will be able to enable Gradium TTS directly from their project settings. The integration will include:

- Low-latency TTS over WebSocket, multiplexed across conversation turns
- Voice cloning from a 10-second audio sample, available on all Gradium plans
- Five deployment modes: cloud, inference partners, dedicated, private cloud, and on-premises, all on the same API surface
- TTS and STT as well as voice cloning for English, French, German, Spanish, and Portuguese

About Agora

Agora (NASDAQ: API) is redefining Human-AI (H2AI) and Human-Human (H2H) interactions as the global leader in Real-Time Engagement (RTE). Agora provides developers with simple, flexible, and powerful APIs alongside no-code deployment tools to rapidly embed real-time conversational AI, voice, video, interactive live streaming, and chat into applications and IoT devices. Headquartered in Santa Clara, CA, Agora is trusted by over 5800 organizations worldwide to power best-in-class digital experiences across all industries. For more information, visit agora.io.

About Gradium

Gradium is a real-time voice AI platform built from the research of Kyutai, the AI research lab behind Moshi and Hibiki. Gradium provides ultra-low latency TTS, STT, and voice cloning APIs for developers building production voice agents, with deployment options from shared cloud to on-premises. Learn more at gradium.ai.

Media Contacts

Agora PR: press@agora.io

Gradium PR: press@gradium.ai