Fish Audio S2: Open-Source Text-to-Speech Beats Google and OpenAI in Blind Listening Tests

Kabui, Charles

2026-06-01

Fish Audio released S2, a text-to-speech system that reads free-form stage directions written inline with the words. Drop `[whisper in small voice]` or `[professional broadcast`

`tone]` next to any phrase and the model steers prosody and emotion to match. Its dual-autoregressive design pairs a 4-billion-parameter model along the time axis with a 400-million-parameter model that fills in acoustic detail at each step. Trained on over 10 million hours of audio in around 50 languages, S2 scored 0.515 on the Audio Turing Test, where 0.5 means listeners label synthetic speech as human about half the time. It also posted the lowest word error rate on the Seed-TTS Eval benchmark in both Chinese (0.54%) and English (0.99%), beating Seed-TTS, MiniMax Speech-02, and Qwen3-TTS.

High-quality emotional TTS has been a paid API, billed per character. S2 ships model weights, fine-tuning code, and an SGLang inference engine together, reaching a real-time factor of 0.195 with around 100 milliseconds to first audio on a single NVIDIA H200. A podcaster or game studio can self-host the same quality. The catch: weights are under a Fish Audio research license, not Apache or MIT, so commercial deployment needs a separate agreement.

An open release now leads a public, blind listening benchmark against closed systems from Google and OpenAI. On EmergentTTS-Eval, S2 wins 81.88% of comparisons against a gpt-4o-mini-tts baseline, including 91.61% on paralinguistics. Self-hosted TTS used to win on cost. It is now winning on quality.

Read More: Supertonic 3: 99M-Parameter On-Device TTS Across 31 Languages, No GPU Required

Sources:

- Fish Audio Open-Sources S2: Fine-Grained Control Meets Production Streaming (Fish Audio)
- Fish Audio S2 Technical Report (arXiv)
- Fish Audio S2 Pro model weights (Hugging Face)
- Fish Speech repository (GitHub)
- Audio Turing Test paper (arXiv)
