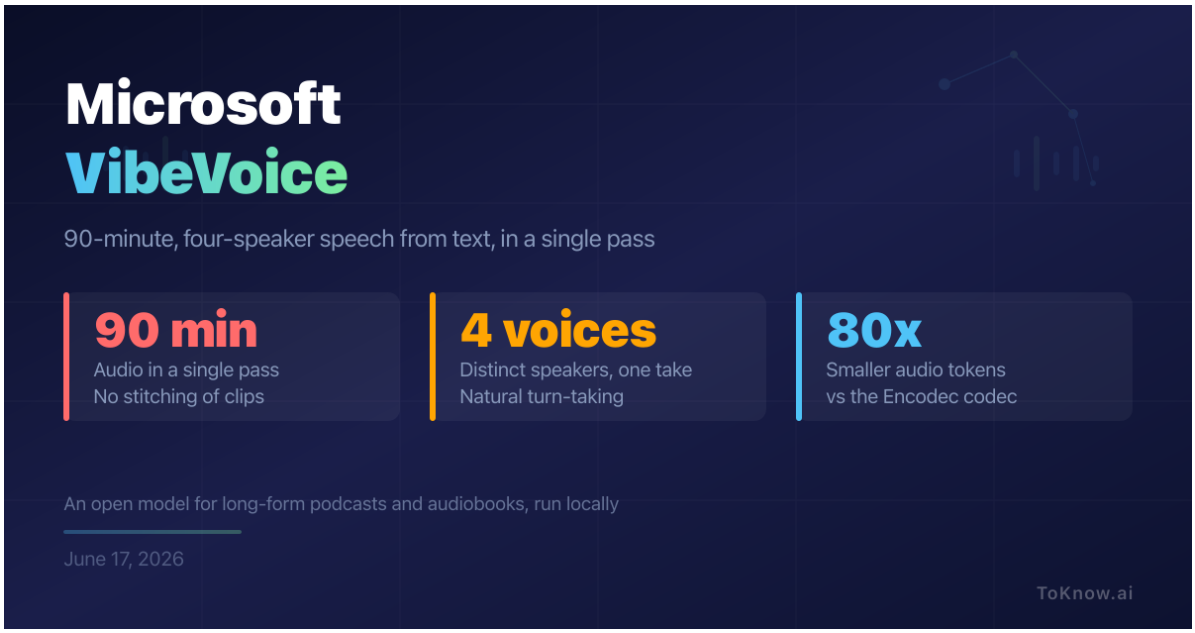
Microsoft VibeVoice: A Model That Turns Text Into 90-Minute, Four-Speaker Audio

Kabui, Charles

2026-06-17

Read at ToKnow.ai



Microsoft Research built VibeVoice, a text-to-speech model that turns a written script into up to 90 minutes of continuous audio with as many as four distinct speakers, in a single pass. Most open speech models manage one or two speakers for a couple of minutes; VibeVoice keeps each voice steady across a full podcast. The trick is a speech tokenizer running at just 7.5 frames per second, about 80 times more efficient than the common Encodec codec, which lets

a 64,000-token context hold an hour and a half of speech. A language model (Qwen2.5, in 1.5B and 7B sizes) tracks the dialogue while a small diffusion head fills in the sound one chunk at a time, an approach called next-token diffusion. In tests, the 7B version beat ElevenLabs and Google's Gemini text-to-speech on how realistic listeners found the audio.

This makes long-form audio cheap to produce. Hand VibeVoice a full podcast script and it returns an hour of natural back-and-forth between named speakers, instead of recording hosts or paying a commercial service per character. The model runs locally on a single GPU, with no subscription. For audiobooks, training material, and multi-host shows, a studio session becomes one generation step.

The enabling idea is compression. By squeezing each second of audio into just 7.5 tokens, VibeVoice fits 90 minutes into a context window most models spend on a few minutes. It points to a pattern in generative audio: better tokenizers, not just bigger models, unlock long-form output.

Read More: Google DeepMind Lyria 3 takes generative audio in a different direction, producing full music tracks from a text prompt.

Sources:

- VibeVoice Technical Report (arXiv)
- VibeVoice-1.5B model card (Hugging Face)
- VibeVoice on GitHub
- Multimodal Latent Language Modeling with Next-Token Diffusion (arXiv)
