NVIDIA PersonaPlex: One Speech Model That Listens, Talks, and Clones Any Voice

Kabui, Charles

2026-02-19

NVIDIA released PersonaPlex, a 7B-parameter speech-to-speech model that operates in full-duplex mode, meaning it listens and speaks at the same time. Built on Kyutai's Moshi architecture, it uses a dual-stream configuration where the listening stream encodes incoming user audio at 24kHz while the speaking stream generates outgoing speech simultaneously. Two prompts configure each conversation: a voice prompt (a short audio clip that sets the vocal

identity) and a text prompt (defining persona attributes like role and scenario). The model was fine-tuned on 1,217 hours of real conversations from the Fisher English corpus plus roughly 2,250 hours of synthetic dialogues generated with Qwen3-32B and Chatterbox TTS. On the FullDuplexBench benchmark, PersonaPlex achieves a 95% takeover rate on user interruptions with 240ms latency, and 170ms latency on smooth turn-taking.

The practical impact is in customer service, voice assistants, and real-time translation. Current voice AI systems are mostly half-duplex: they wait for you to finish, then respond. PersonaPlex handles overlapping speech, backchannels ("uh-huh"), and interruptions natively. The voice cloning from a short prompt means businesses can deploy consistent branded voices without recording extensive datasets. The model runs on A100 or H100 GPUs and is released under NVIDIA's Open Model License, which permits commercial use.

Sources:

- PersonaPlex on Hugging Face
- PersonaPlex Paper (arXiv:2602.06053)
- NVIDIA PersonaPlex Project Page
- PersonaPlex GitHub