Audio Stack — one-page brief

Fully local voice cloning, TTS, lip-sync avatars, music, and auto-dubbing — job-queue architecture, every model MIT/Apache-2.0. A face and a voice are the most privacy-sensitive inputs there are; here they never leave the machine. Public repo: github.com/MushiSenpai/mushishi-audio-stack

Measured (not projected) — 34.5 s reference clip

JOB	MEASURED	NOTE
Voice clone (Demucs + Fish Speech)	5 s	reusable profile, ~4 GB
TTS narration, cloned voice	25 s	34.5 s WAV, ~145 wpm
Transcription (WhisperX, word-aligned)	15 s	~2.3× realtime
Avatar — draft (MuseTalk)	78 s	1024², 25 fps, ~7.7 GB
Avatar — production (LatentSync 1.6)	242 s	~20 GB; cleaner blend, teeth a tie with draft
Avatar — cinematic (Hallo2)	1197 s	~35× slower than realtime; real head motion
Music — 30 s stereo clip (ACE-Step)	10 s	incl. model load
Music — 15 s song w/ vocals (YuE)	176 s	mono is a model limit, stated
Dub, same language	20 s	cross-language 3–4 min (phased VRAM)

> **The honest ceiling:** avatar quality is social-grade across all three engines. Broadcast close-ups still go to a cloud path — I say that before you pay, not after. The three lip-sync engines run as isolated, version-pinned services; the 25+ install lessons are public.

Engagement

- **Avatar video:** Draft SGD 80 · Production SGD 150 · Cinematic quoted per job.
- **Turnkey deployment** — the pipeline on your hardware (marketing ad-reads, narration, dubbing), with runbook and quality-tier guidance.

Madhan Kumar Reddy · Singapore · reddy.madhankumar.sg@gmail.com · theinvalid.me — every number above is a measured run on one RTX 5090 (32 GB); benchmarks and failure logs are public in the repo. Generated 2026-07-02.