TL;DR
SpeechRole (Aug 2025) confirms **E2-TTS** and **F5-TTS** as the most reliable finetuning targets when you can supply 10-60 minutes of labelled speech.
The benchmark used an older GPT-SoVITS build; the repo now ships **V2Pro** with upgraded flow-matching and acoustic encoders that close much of the gap.
Pick your model based on latency, multilingual coverage, and how much you want to lean on diffusion vs. autoregressive decoding, then lock in a data hygiene + evaluation loop before touching production voices.

---

1 Why SpeechRole matters for teams with voice data

SpeechRole is the first large-scale benchmark (Aug 2025) scoring voice cloning systems on **naturalness, role fidelity, and robustness** across curated role-play scenarios. Key takeaways for practitioners with proprietary speech libraries:

- **Finetuning still wins.** Open-weights models trained from scratch underperform when you have 30+ minutes of clean target speech. Finetuning on role-specific emotion tags lifts MOS and reduces pronunciation drift.
- **Diffusion models are maturing.** F5-TTS leverages a flow-matching decoder that the benchmark shows outperforming autoregressive baselines on long-form stability.
- **Evaluation must be multi-dimensional.** SpeechRole reports MOS (Mean Opinion Score), CER/WER, and "role accuracy" scored by LLM judges. Optimising for one metric can hide gaffes elsewhere.

ByteDance, streaming studios, and localisation vendors reading the report are now re-balancing their pipelines: diffusion for expressive reads, flow-matching for fast adaptation, and still keeping SoVITS-like architectures for low-latency chat.

---

2 Model snapshots: what changed in 2025

2.1 E2-TTS (2025 refresh)

- **Architecture:** End-to-end neural codec TTS with diffusion-based duration modelling; unified encoder handles phoneme, prosody, and speaker embeddings.
- **Benchmark showing:** SpeechRole ranks E2-TTS at or near the top on MOS and role accuracy for English and Mandarin tasks when finetuned on 20-40 minutes of aligned speech.
- **Why teams pick it:** Low inference jitter, native multilingual tokeniser, and a mature recipes repo (E2-VITS) for batching speakers.
- **Watch-outs:** Training is compute-heavy (>=4xA100 recommended) and still benefits from external alignment (MFA) for best results.

2.2 F5-TTS (Flow-matching + Diffusion)

- **Architecture:** Combines HuBERT-style content encoders with a flow-matching decoder (FastFlow) that bakes alignment into continuous time steps.
- **Benchmark showing:** SpeechRole recognises F5-TTS as the most balanced option: high MOS, low WER, strong emotion retention, especially when conditioning on style prompts.
- **Why teams pick it:** Excellent long-form stability (>3 min) and expressive cloning without heavy prompt engineering. Finetunes converge in ~30k steps with mixed precision.
- **Watch-outs:** Still emerging tooling; you need to bring your own inference server and consider quantisation for real-time use.

2.3 GPT-SoVITS V2Pro (repo release, post-benchmark)

- **Context:** SpeechRole likely tested a pre-V2Pro branch, which explains the gap. The official repo now advertises **V2Pro** with flow-matching decoders, NeMo-style SSL front-ends, and better VAD.
- **What's new:**
  - **Dual-stage encoder** (content + timbre) with continual learning hooks.
  - **Global Style Tokens** updated for surface-level emotion control.
  - **Faster autoregressive path** for chat assistants (~120 ms first token on A100).

- **Why teams still care:** SoVITS remains community-friendly, with easy dataset tooling and decent quality on 5-10 minute speaker packs.
- **Watch-outs:** Requires careful LR scheduling to avoid catastrophic forgetting; out-of-the-box benchmarks still lag F5/E2 on expressive passages until you apply the new finetune recipes.

3 Finetuning playbook when you control the data

Stage	Actions	Notes
1. Data audit	Deduplicate, remove cross-talk, normalise loudness (-23 LUFS), annotate emotion/intent.	SpeechRole shows noisy labels drag MOS by 0.2-0.4.
2. Alignment	Force-align transcripts (MFA, Aeneas) and export phoneme durations.	E2/F5 expect aligned phoneme or syllable spans for stable timing.
3. Feature bake	Precompute content features (HuBERT, Wav2Vec2) and pitch contours.	Speeds finetuning; reuse across experiments.
4. Finetune	Run mixed-precision training with speaker-specific configs.	Keep learning rates low (1e-5 to 3e-5) to preserve base timbre.
5. Evaluate	MOS panel (5-7 listeners), automatic WER, LLM-based role scoring.	Mirror SpeechRole to track parity.
6. Deploy	Export ONNX/FP16, add guardrails (profanity filters, watermarking).	Confirm latency budget vs. target platform.

4 Choosing the right model: decision matrix

Scenario	Recommended stack	Why
Long-form narration (>5 min) with emotional arcs	F5-TTS + expressive prompts	Diffusion handles breath + phrasing; stable for chapters.
Multilingual customer support	E2-TTS finetuned with phoneme sharing	Superior cross-lingual alignment; add locale-specific prosody.

Conversational agent with tight latency	GPT-SoVITS V2Pro fast path	Autoregressive decoder keeps latency low; upgrade timbre with V2Pro recipes.
Rapid A/B testing of scripts	E2-TTS distilled + VALL-E style prompting	Keeps quality while batch-generating variations.
Prototype with minimal data (less than 10 min)	GPT-SoVITS V2Pro + speaker encoder	Few-shot still decent; upgrade later to F5/E2 once data grows.

5 Practical tips from recent deployments

- **Use SpeechRole as regression testing.** Recreate its evaluation prompts internally; track MOS/WER deltas after each finetune cycle.
- **Plan for emotion transfer.** Tag your corpus with emotion labels (AROUSAL/VALENCE or Ekman) so diffusion models know what to emphasise.
- **Stabilise phoneme lengths.** For F5/E2, freeze duration predictors for the first 5k steps; unlock once losses flatten.
- **Guard voice rights.** Maintain consent logs and watermark generated audio (ultrasonic or content-ID) to flag misuse.
- **Automate dataset refresh.** Build nightly jobs to ingest new scripts, run diarisation, and push clean segments into your training bucket.

6 Open questions heading into 2026

1. **Public V2Pro benchmarks.** The community still needs independent MOS/WER runs of GPT-SoVITS V2Pro to validate repo claims against SpeechRole numbers.
2. **Cross-lingual emotion transfer.** F5/E2 excel in bilingual setups, but expressive Cantonese/Japanese remains tricky without large role-play datasets.
3. **Latency vs. quality.** Diffusion-based decoders are closing the gap, yet on-device voice agents may still require autoregressive fallbacks.
4. **Watermark standards.** Expect regulation to demand verifiable synthetic voice markers; plan for Spectrum or KIT watermark integration.
5. **Licensing clarity.** Check EULAs (especially for commercial use of Tencent/Bilibili-backed checkpoints) to avoid legal surprises.

Voice cloning is no longer about picking the flashiest demo. With SpeechRole as your north star, align model choice with data volume, latency targets, and compliance guardrails. Finetuned E2-TTS and F5-TTS set the quality bar today, while GPT-SoVITS V2Pro offers a nimble path for rapid experimentation; just keep your evaluation harness honest and iterate with consent front and centre.

References

- Coqui TTS (GitHub)
- NeMo Toolkit (GitHub)
- NeMo Toolkit (arXiv)
- SV2TTS (arXiv)