Breeze TTS 2 shipped as an inference model. We wanted to know whether its synthesis stack could be adapted on one 24 GB GPU, which model components had to change, and whether full supervised fine-tuning would improve on a much smaller LoRA adapter.

We built the missing training surface, trained both routes on the consented FEMALE_01 subset of Singapore's National Speech Corpus, and released the source toolkit and selected research models.

60-second takeaway

- **Both training routes completed.** The source toolkit reconstructs supervised text and audio labels, trains LoRA or full SFT, resumes checkpoints, exports artifacts and generates matched evaluation packs.
- **LoRA was the smaller practical release.** The selected rank-8 adapter trains 12,092,448 parameters across the semantic backbone, depth decoder and three projection families.
- **Full SFT changed much more of the synthesizer.** The selected checkpoint updates 2,387,151,872 synthesis parameters while keeping the text encoder and audio codec frozen.
- **The objective comparison did not establish a winner.** Full SFT reached mean ECAPA speaker similarity 0.6973, compared with 0.6810 for LoRA. Both reached 0.0467 WER, and the paired ECAPA confidence interval crossed zero.
- **One blind listener preferred LoRA for cadence and long-form listening.** The three short-prompt comparisons were ties. Both models mispronounced the same tested local word.
- **These are non-commercial research releases.** The source code is Apache 2.0, while the model artifacts remain governed by the BreezeBlue Research and Non-Commercial License Agreement version 1.1.

Open the released work

- Source toolkit on GitHub
- Rank-8 LoRA adapter on Hugging Face
- Selected full-SFT checkpoint on Hugging Face
- Breeze TTS 2 fine-tuning collection on Hugging Face

- Upstream Breeze TTS 2 model

The two Hugging Face repositories are public but gated. A user must accept the licence conditions before downloading the files. The model cards remain visible so that the provenance, intended use, evaluation and limitations can be reviewed before access is requested.

What we had to add

The upstream release provided inference rather than a supported training interface. A useful adaptation repository needed more than a loop that made the loss decrease.

Our source toolkit adds:

1. deterministic reconstruction of text labels and 16-codebook target-audio tokens;
2. a one-example BF16 forward-and-backward feasibility gate;
3. LoRA across the semantic backbone, depth decoder, text projection, depth-input projection and output projection;
4. full SFT with a frozen text encoder and audio codec;
5. checkpoint resume, adapter export, merge, fresh-process reload and artifact identity checks;
6. validation-based checkpoint selection and bounded hyperparameter sweeps;
7. matched generation, ASR, speaker similarity, acoustic diagnostics, confidence intervals and opaque blind-listening packs; and
8. release checks that exclude training audio, optimizer state, caches and private receipts.

The repository was exercised on an NVIDIA RTX 3090 Ti. This establishes the documented paths on that 24 GB CUDA system. It does not show that every dataset, GPU or sequence length will fit.

LoRA route

The selected LoRA artifact uses rank 8 and alpha 16. It came from step 500 of a 1,000-step schedule and contains 12,092,448 trainable parameters.

We did not restrict adaptation to the main transformer backbone. Breeze predicts speech through both a semantic backbone and a depth decoder, with projections connecting text, depth input and output tokens. The feasibility work showed that all of those families participate in the supervised objective, so the released adapter covers them together.

The adapter manifest pins the upstream model revision and required base-file checksums. The custom loader refuses a mismatched revision or file identity. This matters because an adapter can load successfully against the wrong mutable base and still produce a misleading comparison.

Full-SFT route

The selected full-SFT artifact came from step 750 of a 1,000-step run. It updates 2,387,151,872 synthesis parameters while freezing the text encoder and audio codec. The run used FP32-master SGD with BF16 model weights.

The public checkpoint contains inference roles only. It excludes optimizer, scheduler, random state, trainer state, source recordings, caches and private receipts. That smaller release boundary is intentional: users receive what is needed for inference without receiving private data or unnecessary training state.

Matched results

We compared the selected LoRA and full-SFT artifacts under the same reference-free prompts and generation conditions.

Measurement	LoRA rank 8	Full SFT
Selected checkpoint	Step 500	Step 750
Mean ECAPA speaker similarity	0.6810	0.6973
Word error rate	0.0467	0.0467
Blind cadence preference	Tie on short prompts; preferred on long-form	Preferred in one cadence prompt
Blind long-form preference	Preferred	Not preferred

The ECAPA difference is descriptive. Its paired confidence interval crossed zero, so the checked objective sample did not establish that full SFT preserves speaker identity better.

The listening study contained one listener. On neutral, ordinary Singapore English and local-pronunciation prompts, the listener recorded ties: LoRA sounded smoother, while full SFT sometimes sounded closer to Singaporean rhythm. Full SFT won one cadence prompt, but LoRA was preferred for the long passage because it stayed more expressive and was less tiring.

This mixed result is useful. A larger checkpoint is not automatically the better voice, and a speaker-embedding score does not measure cadence, pronunciation or listening fatigue.

What the experiment does not prove

- One voice profile does not establish performance across Singaporean English speakers or accents.
- One blind listener does not establish a population preference.
- ECAPA similarity is a diagnostic proxy, not proof of identity fidelity.
- WER checks whether an independent recognizer recovered the words. It does not measure naturalness or pleasantness.
- Both selected artifacts mispronounced the tested word paiseh. Neither route solved local pronunciation merely by seeing Singapore English training data.
- The releases are research artifacts, not evidence of production readiness or commercial permission.

Which artifact should you try?

Start with LoRA when you want a small artifact, rapid iteration and an explicit adaptation layer over a separately pinned base. In this experiment, LoRA also received the stronger long-form listening preference.

Try full SFT when you specifically need to test whether deeper synthesis changes improve your own voice and prompts, and you can carry a much larger checkpoint. Our objective comparison did not prove that full SFT was better overall, so its extra size should not be treated as quality evidence.

For either route, use held-out prompts and listen blind. Keep pronunciation, accent, cadence, long-form monotony and fatigue separate from word accuracy and speaker similarity.

Licence and responsible use

The GitHub toolkit is source-only and licensed under Apache 2.0. It does not include model weights, adapters, checkpoints, training audio or generated speech.

The upstream model and both Instavar derivatives are governed separately by the BreezeBlue Research and Non-Commercial License Agreement version 1.1. Review the licence and notices in each model repository before use. Do not use the released models for a product, paid service, client work, advertising, revenue generation, production deployment, impersonation, deception or non-consensual voice cloning.

The FEMALE_01 source material and our consent record do not override the model licence. Dataset rights, voice consent, code licensing and model licensing are separate checks.

Reproduce and inspect

The source repository documents installation, deterministic target preparation, LoRA, full SFT, resume, merge and evaluation commands. It also records the exact upstream revision required by the released artifacts.

Before trusting a result:

1. verify the base revision and checksums;
2. keep validation data separate from training data;
3. compare base, LoRA and full SFT with identical prompts and seeds;
4. check words, speaker similarity and acoustic measurements;
5. listen to an opaque pack without model identities; and
6. choose a checkpoint only after the objective and listening evidence agree, or record the disagreement explicitly.

Related research:

- LoRA vs Full SFT for Voice Models

- Best Open-Source TTS Models for Production
- IMDA NSC Voice Cloning Benchmark