Fleet-1: multilingual Whistle adapter continuation

Lucas van de weerd | Lucasgitguy-co | 7 November 2026 | Version 1.0

Repository: Lucasgitguy-co/Fleet-1

Reviewed revision: 0cc581616f7e5493e2e3999e71c134cedc879a6d

Abstract

This report documents continuation of a Korean-fine-tuned Whistle model on eight languages. Recorded short-clip evaluation shows small error-rate decreases in five languages and regressions in Spanish, Dutch and especially Korean. The results do not demonstrate improvement across all eight languages.

1. Scope and related work

This technical report reconstructs the release from pinned repository artifacts and the publisher's earlier portfolio report. It is an AI-assisted documentation synthesis, not an independent rerun or a peer-reviewed publication. Reported measurements belong to the original release records. [1, 2, 3, 4]

Conneau et al. (2022), FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech, https://arxiv.org/abs/2205.12446, describes the multilingual speech benchmark used as a data source. This report evaluates only the explicitly selected subset.

2. Model and method

The starting model is Sky-Kim/whistle-finetuned-ko at revision b8a4f3188a828a0e87e2713b2d3e83d71d893068, retaining its vocabulary, language tags and LoRA weights. FLEURS supplies 400 training and 50 validation clips per language. Official splits remain separate and repeated normalized transcripts are excluded across selected splits. The card records two epochs, batch size 1, AdamW at 0.00002, clipping 1, warmup/cosine scheduling and W4/A8 quantization-aware CPU training. Epoch 2 was selected by macro validation loss. [1, 2, 3, 4]

3. Evidence and results

Both source and Fleet-1 are evaluated with automatic language detection in the same native Cactus engine. Each language has 100 test clips filtered to at most 12 seconds and 158 text tokens. Lower error rate is better. WER and CER have different units and must not be averaged as if they were one common metric. The large Korean CER increase is a material regression. [1, 2, 3, 4]

Language	Metric	Clips	Source	Fleet-1
en	WER	100	14.87%	14.39%
de	WER	100	22.71%	22.18%
fr	WER	100	25.73%	25.18%
es	WER	100	11.83%	11.93%
it	WER	100	16.60%	16.32%
nl	WER	100	29.39%	29.53%
pl	WER	100	43.68%	43.13%
ko	CER	100	33.68%	52.35%

4. Limitations and intended use

These are filtered short-clip measurements, not full FLEURS benchmark results or statistical-significance tests. Long, noisy and conversational speech remain unevaluated. Upstream pretraining exposure is incompletely known. Use 16 kHz mono audio and the recorded Cactus runtime; whistle.cact is not a Transformers checkpoint. Retain Apache-2.0 model attribution and FLEURS CC BY 4.0 dataset attribution.

5. Reproducibility and conclusion

The source snapshot above fixes what was reviewed. The linked source files and source-manifest.json record the evidence used, including file hashes. Consult the repository's usage instructions and recorded dependencies for execution. No new model training, inference benchmark, or claimed state-of-the-art result was introduced in preparing this report. Evidence status: recorded held-out asr evaluation.

References

[1] Lucasgitguy-co. README.md at reviewed revision.

[2] Lucasgitguy-co. evaluation.json at reviewed revision.

[3] Lucasgitguy-co. training_manifest.json at reviewed revision.

[4] Lucasgitguy-co. UPSTREAM_MODEL_CARD.md at reviewed revision.

[5] Lucasgitguy-co. Local Model Adaptation Across Vision Diagrams Geometry and Game Decisions. Historical portfolio report, 5 November 2026; its proposed v2 experiment is superseded by later release records.