A Compact Whisper+LoRA Baseline for Taiwanese Hakka ASR in FSR-2025

Hung-Ting Hsieh
Independent Researcher
hsiehk0214@gmail.com

Abstract

We present a compact baseline for the Formosa Speech Recognition (FSR-2025) Taiwanese Hakka ASR challenge. Our system fine-tunes *Whisper-large-v2* (Radford et al., 2022) with LoRA (Hu et al., 2021), using consistent text normalization and balanced dev splits. On the official warm-up set, we obtain 10.94% CER for Track 1 (Hanzi) and 28.48% SER for Track 2 (Pinyin). We provide simple yet reproducible pipelines covering data preparation, training, inference, and evaluation. Code is available at github.com/Kevindic0214/FSR-Challenge-2025.

Keywords: Automatic speech recognition; Hakka; Whisper; LoRA; CER; SER

1 Introduction

Taiwanese Hakka is a low-resource language variant of significant cultural value. FSR-2025 defines two tracks: Track 1 evaluates character error rate (CER) on Hanzi, and Track 2 evaluates syllable error rate (SER) on Pinyin. We aim to provide a strong, minimal-requirement baseline using *Whisper-large-v2* fine-tuned with low-rank adaptation (LoRA), emphasizing practical engineering choices and reproducibility over model complexity. In this work, we follow the official specification of the FSR-2025 challenge (FSR2025).

2 Task and Data

We train on the HAT-Vol2 corpus (˜60 hours; Dapu and Zhao'an dialects; 16 kHz mono) and evaluate on the FSR-2025 warm-up set (˜10 hours; 4,299 utterances total). We build manifests via dedicated scripts for each track, apply Unicode NFKC normalization, remove zero-width characters, and adopt track-specific text processing: Hanzi cleaning for Track 1 and Pinyin digit-tone policy for Track 2. Dev speakers are selected in a balanced way across DF/DM/ZF/ZM groups for stable validation.

We rely on the HAT-Vol2 dataset (HAT-Vol2) and the official warm-up set (FSR2025) for training and evaluation.

Split	Size	Notes
HAT-Vol2 (train)	$\sim$60 h	Dapu/Zhao'an; 16 kHz m
Warm-up (eval)	$\sim$10 h / 4,299 utt	Official FSR-2025 set
Dev speakers	12 (balanced)	DF/DM/ZF/ZM allocatio

Table 1: Dataset overview and evaluation split.

3 Related Work

Low-resource ASR has been explored in multilingual programs such as Babel (Harper, 2014). Whisper (Radford et al., 2022) is a strong multilingual recognizer; we adapt it to Hakka via parameter-efficient fine-tuning. LoRA (Hu et al., 2021) reduces trainable parameters for seq2seq models while retaining quality, enabling practical fine-tuning on 24 GB GPUs.

4 Approach

We fine-tune *Whisper-large-v2* (Radford et al., 2022) with LoRA (Hu et al., 2021) (rank 16, α=32, dropout 0.05). Training uses gradient checkpointing, bf16 when available, and label smoothing. For Track 1 decoding, we force Chinese transcription via the decoder prompt; Track 2 uses language-appropriate decoding without language forcing. Beam search with 5 beams and temperature 0.0 is used unless specified.

Implementation details: we apply LoRA adapters to attention and MLP modules (`q_proj`, `k_proj`, `v_proj`, `out_proj`, `fc1`, `fc2`); enable TF32 for faster, stable training on recent GPUs; and use label smoothing of 0.1.

5 Experiments

We train for 3 epochs with per-device batch size 2 and gradient accumulation 16 on an RTX 4090 (24 GB). Evaluation metrics are CER (Track 1) and SER (Track 2) with sentence-level exact match for reference.

6 Results

On the warm-up set: Track 1 reaches 10.94% CER with 58.06% exact match; Track 2 reaches 28.48% SER with 12.17% exact match. These numbers are obtained with the shared pipelines and no external data beyond the provided corpora. We observe stable validation under balanced speaker splits and consistent normalization.

Track	Metric	Score
Track 1 (Hanzi)	CER / EM	10.94% / 58.06%
Track 2 (Pinyin)	SER / EM	28.48% / 12.17%

Table 2: Warm-up evaluation results. EM: exact match.

7 Reproducibility

Code and scripts are available at github.com/Kevindic0214/FSR-Challenge-2025.

We provide end-to-end scripts for data preparation, training, inference, and evaluation. Minimal examples:

Track 1:

```
python prepare_hakka_track1.py --root
    HAT-Vol2 \
  --drop_mispronounce
      --relative_audio_path
python train_whisper_lora_track1.py \
  --train_jsonl
      HAT-Vol2/manifests_track1/train.jsonl
      \
  --dev_jsonl
      HAT-Vol2/manifests_track1/dev.jsonl
python infer_track1.py --eval_root
    FSR-2025-Hakka-evaluation \
  --outfile predictions_track1.csv
      --model openai/whisper-large-v2 \
  --lora_dir runs/track1/lora_v2_r16_e3
python eval_track1_cer.py --key_dir
    FSR-2025-Hakka-evaluation-key \
  --hyp predictions_track1.csv
```

Track 2:

```
python prepare_hakka_track2.py
    --data_root HAT-Vol2 \
  --out_dir HAT-Vol2/manifests_track2
      --exclude_mispronounced
python train_whisper_lora_track2.py
python infer_track2.py --eval_root
    FSR-2025-Hakka-evaluation \
  --outfile predictions_track2.csv
      --model openai/whisper-large-v2 \
  --lora_dir
      exp_track2_whisper_large_lora
python eval_track2_ser.py --key_dir
    FSR-2025-Hakka-evaluation-key \
  --hyp predictions_track2.csv
```

8 Error Analysis

Common errors include character/phonetic substitutions and occasional short repeats; we monitor n-gram repetition to detect degeneration. Performance degrades mildly for longer utterances; bucketed analysis suggests length-aware decoding or better segmenting could help.

Examples. Sampled warm-up mismatches:
(003jh5p8hd.wav) ref: ; hyp: .
(03qw9gfad7.wav) ref: ; hyp: .
(04qied7gz8.wav) ref: . . . . . . ; hyp: . . . . . .

These illustrate homophone/near-neighbor substitutions and local phrase alterations; stronger language modeling or constrained decoding may mitigate such errors.

9 Conclusion

We provide a concise, reproducible baseline for both tracks of FSR-2025 Hakka ASR using Whisper+LoRA. Future work includes dialect-aware adaptation, LM-rescoring for Hanzi, refined Pinyin normalization, and temperature/beam tuning.

Limitations

Our results are based on the provided HAT-Vol2 training data and the official warm-up set. We do not explore external language models or data augmentation; Pinyin normalization choices (e.g., starred syllables) can affect SER.

Acknowledgments

We thank the FSR-2025 organizers and dataset providers.

References

FSR2025. 2025. Formosa speech recognition challenge 2025: Hakka asr. Challenge. Warm-up evaluation set and official task description.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Association for Computational Linguistics.

HAT-Vol2. 2024. Hat-vol2: Taiwanese hakka speech corpus. Dataset. ˜60 hours; Dapu and Zhao'an dialects; 16 kHz mono.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.