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

Anonymous ROCLING submission

Abstract

We present a compact baseline for the Formosa Speech Recognition (FSR-2025) Taiwanese Hakka ASR challenge. Our system fine-tunes *Whisper-large-v2* (Track 1) and *Whisper-large-v3-turbo* (Track 2) (Radford et al., 2022) with LoRA (Hu et al., 2021), under a consistent normalization policy and balanced speaker-based 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, reproducible pipelines covering data preparation, training, inference, and evaluation, without using external data or language models.

Keywords: Automatic speech recognition; Hakka; Whisper; LoRA; low-resource; FSR-2025; 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* (Track 1) and *Whisper-large-v3-turbo* (Track 2) 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).

Contributions. (i) A reproducible baseline for both tracks with unified normalization that matches evaluation; (ii) a simple LoRA recipe runnable on a single 24 GB GPU with balanced speaker-based dev split; (iii) competitive warm-up results with lightweight error and length-bucket analyses.

Table 1: Dataset overview and evaluation split.

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

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.

Normalization policy. Track 1 (Hanzi): apply NFKC, remove zero-width characters, map mixed punctuation to Chinese forms, and strip spaces to align with evaluation. Track 2 (Pinyin): apply NFKC, remove zero-width characters, map ü/ǘ/… and "u:/U:" to "v", keep only [a–z0–9] and single spaces, and by default drop starred syllables (e.g., "*ki53" or "ki53*"); an optional fix merges split-tone forms (e.g., "ki 53" → "ki53").

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* (Track 1) and *Whisper-large-v3-turbo* (Track 2) (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.

We keep Whisper's default suppression behavior (do not forcibly clear `suppress_tokens`), disable the generation cache during training, and enable early stopping on the dev metric (patience 2). bf16 is automatically used when supported; otherwise fp16 on GPU.

Implementation. We implement training and inference with HuggingFace Transformers and PEFT on PyTorch. Base models: `openai/whisper-large-v2` (Track 1) and `openai/whisper-large-v3-turbo` (Track 2). Audio I/O uses `torchaudio` for Track 1 and `soundfile+librosa` for Track 2. Manifests are JSONL with fields {`utt_id`, `audio`, `text/hanzi/pinyin`, `group`}; relative audio paths are resolved via a root flag. During training we enable gradient checkpointing (non-reentrant when available), set `use_cache`=False, and turn on TF32. Decoding uses `num_beams`=5, `temperature`=0.0, `no_repeat_ngram_size`=3, `length_penalty`=1.0, and `max_new_tokens`=256; we force a Chinese decoder prompt only for Track 1. We log CER/SER, exact match, group/length-bucket scores, 3-gram repetition rate, throughput, and peak memory; training uses AdamW with a linear schedule and 500 warmup steps, and we save only the LoRA adapter and processor for lightweight deployment.

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. We use seed 1337, learning rate 1×10^{-4} for *large-v2* (Track 1) and 3×10^{-4} for *large-v3-turbo* (Track 2), with 500 warmup steps, label smoothing 0.1, gradient checkpointing, TF32, and early stopping (patience 2). The HuggingFace Trainer default optimizer (AdamW) is used.

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 (*large-v2* for Track 1 and *large-v3-turbo* for Track 2) and no external data beyond the provided corpora. We observe stable validation under balanced speaker splits and consistent normalization. Longer utterances show mildly higher error rates (notably in the 12.4–20 s bucket), and we observe small variations across DF/DM/ZF/ZM groups under the balanced-split protocol.

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

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

Final-test results. On the official final-test, our system achieves 18.78% CER for

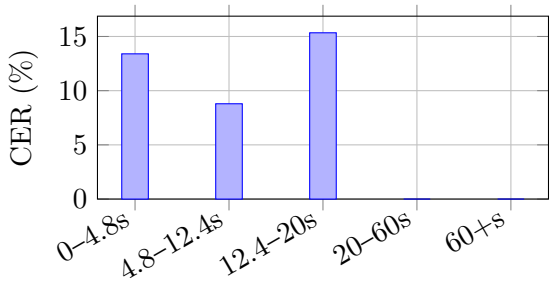

Figure 1: Track 1 warm-up CER by utterance duration.

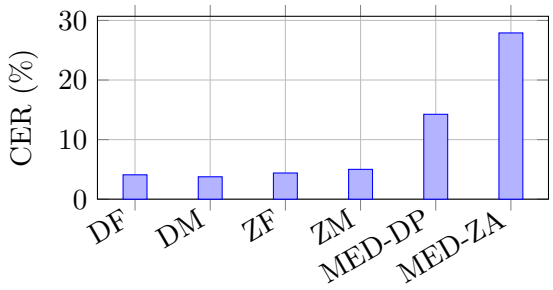

Figure 2: Track 1 warm-up CER by group (recorded DF/DM/ZF/ZM and media subsets).

Track 1 and 33.38% SER for Track 2. For analysis, we also report a tone-removed Pinyin WER of 21.30%. Figures 5 and 6 show the organizer-provided charts. In our social group, our CER ranking is **2/3** (Track 1), and our tone-removed Pinyin WER ranking is **2/2** among teams with available data (Track 2).

Length buckets and groups. We further report Track 1 warm-up CER by utterance length and by speaker groups. Length buckets follow the official seconds-based bins (0–4.8/4.8–12.4/12.4–20 s). Figures 1 and 2 summarize the results.

Track 2 buckets and groups. For Track 2 (Pinyin), we report SER by syllable length buckets and by groups in Figures 3 and 4.

Table 3: Official final-test scores (values only; charts and rankings omitted for review).

Track	Metric	Final-test
1 (Hanzi)	CER	18.78%
2 (Pinyin)	SER	33.38%
2 (Pinyin)	WER (tone-removed)	21.30%

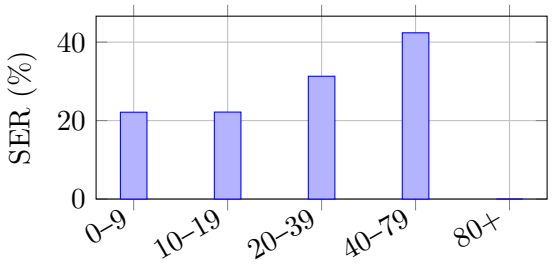

Figure 3: Track 2 warm-up SER by syllable length (tokens).

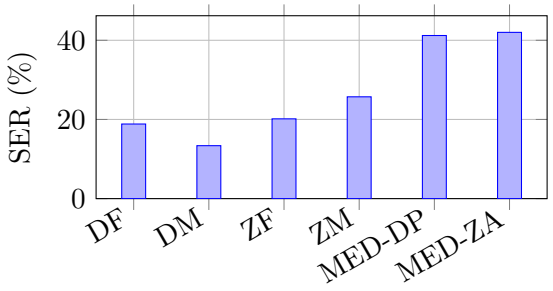

Figure 4: Track 2 warm-up SER by group (recorded DF/DM/ZF/ZM and media subsets).

7 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: 大家攏無仰子**啲**隨拹你**人救**出去；hyp: 大家攏無仰子**項**隨拹你**研究**出去。
(03qw9gfad7.wav) ref: 食**著幾**隻草蜢**乜好**啊；hyp: 食**到佢**隻草蜢**毋會好**啊。

These illustrate homophone/near-neighbor substitutions and local phrase alterations; stronger language modeling or constrained decoding may mitigate such errors. For Pinyin (Track 2), common patterns include tone-digit confusions and occasional effects from star-syllable handling; our normalization reduces such artifacts.

Limitations

Our results are based on the provided HAT-Vol2 training data and the official warm-up

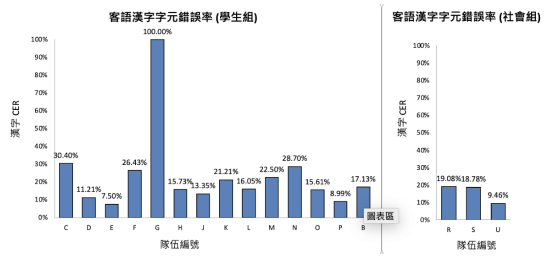

Figure 5: Official final-test chart: CER (Track 1).

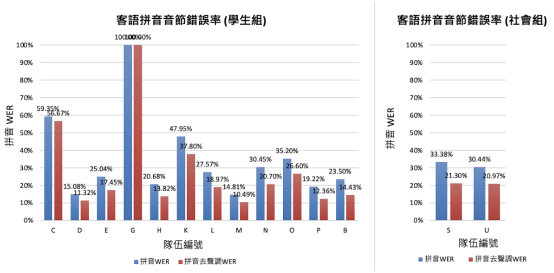

Figure 6: Official final-test chart: tone-removed WER/SER (Track 2).

set. We do not explore external language models or data augmentation; Pinyin normalization choices (e.g., starred syllables) can affect SER.

8 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.

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*.