

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* (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.

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)	~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

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, $\alpha=32$, 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.

100 5 Experiments

101 We train for 3 epochs with per-device batch size
 102 2 and gradient accumulation 16 on an RTX 4090
 103 (24 GB). Evaluation metrics are CER (Track 1) and
 104 SER (Track 2) with sentence-level exact match for
 105 reference.

106 6 Results

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

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

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

122 7 Error Analysis

123 Common errors include character/phonetic substi-
 124 tutions and occasional short repeats; we monitor
 125 n-gram repetition to detect degeneration. Perfor-
 126 mance degrades mildly for longer utterances; buck-
 127 eted analysis suggests length-aware decoding or
 128 better segmenting could help.

129 **Examples.** Sampled warm-up mismatches:

130 (003jh5p8hd.wav) ref: ; hyp: .

131 (03qw9gfad7.wav) ref: ; hyp: .

132 (04qied7gz8.wav) ref:; hyp:

133 These illustrate homophone/near-neighbor sub-
 134 stitutions and local phrase alterations; stronger lan-
 135 guage modeling or constrained decoding may miti-
 136 gate such errors.

138 Limitations

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

145 8 Conclusion

146 We provide a concise, reproducible baseline for
 147 both tracks of FSR-2025 Hakka ASR using Whis-
 148 per+LoRA. Future work includes dialect-aware

149 adaptation, LM-rescoring for Hanzi, refined Pinyin
 150 normalization, and temperature/beam tuning.

153 References

154 FSR2025. 2025. Formosa speech recognition challenge
 155 2025: Hakka asr. Challenge. Warm-up evaluation
 156 set and official task description.

157 Mary Harper. 2014. Learning from 26 languages: Pro-
 158 gram management and science in the babel program.
 159 In *Proceedings of COLING 2014, the 25th Inter-
 160 national Conference on Computational Linguistics: Tech-
 161 nical Papers*, page 1, Dublin, Ireland. Association
 162 for Computational Linguistics.

163 HAT-Vol2. 2024. Hat-vol2: Taiwanese hakka speech
 164 corpus. Dataset. ~60 hours; Dapu and Zhao'an di-
 165 alects; 16 kHz mono.

166 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan
 167 Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen.
 168 2021. Lora: Low-rank adaptation of large language
 169 models. *arXiv preprint arXiv:2106.09685*.

170 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brock-
 171 man, Christine McLeavey, and Ilya Sutskever. 2022.
 172 Robust speech recognition via large-scale weak su-
 173 pervision. *arXiv preprint arXiv:2212.04356*.