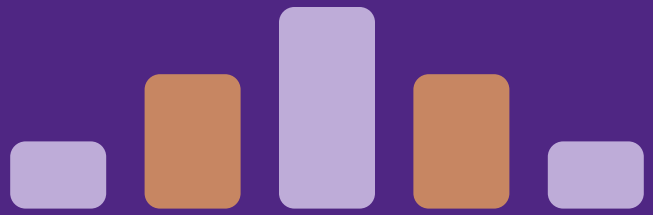
Western

CS HONOURS THESIS · WESTERN UNIVERSITY

Fine-tuning Whisper for Arabic Dysarthric Speech Recognition under Data Scarcity

Hussein Mourad

Department of Computer Science



The problem: no Arabic dysarthric speech data

- Dysarthria (motor-speech impairment) makes speech slower, distorted, and harder for standard ASR to transcribe
- Modern ASR systems (e.g. Whisper) are trained overwhelmingly on healthy, fluent speech
- No large, labelled Arabic dysarthric speech corpus currently exists
- This absence blocks the clinical fine-tuning step needed to adapt ASR for this population



OBJECTIVES

Research questions

Q1 Can a synthetic Arabic dysarthric corpus support useful fine-tuning of a modern ASR model?

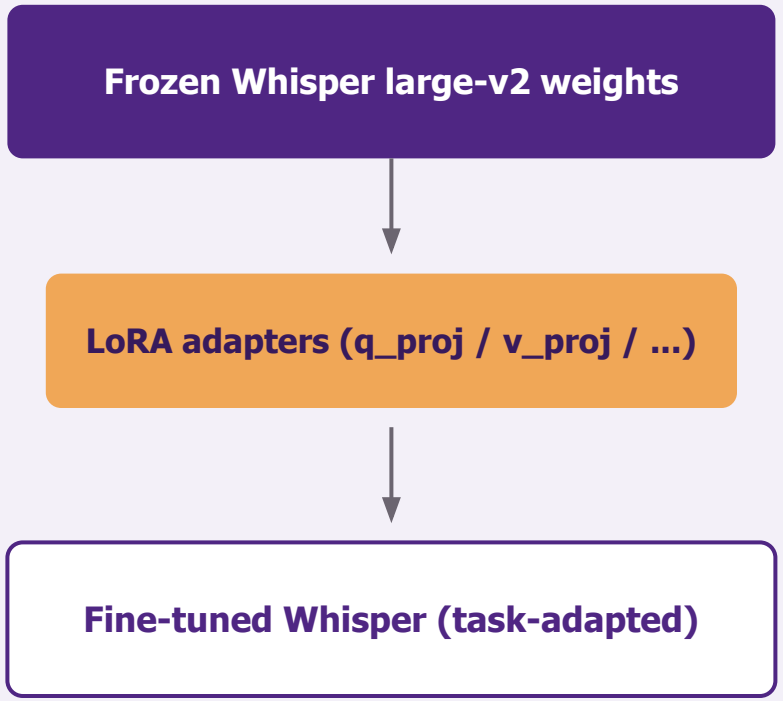
Q2 Can Whisper large-v2 be parameter-efficiently adapted on a single consumer-grade GPU?

Q3 How do LoRA capacity, learning rate, and checkpoint-selection metric affect Word Error Rate?

Objective: Identify a training recipe that minimizes WER under severe data scarcity.

Whisper + LoRA, briefly

- Whisper large-v2: ~1.5B-parameter encoder-decoder ASR model (OpenAI)
- LoRA (Low-Rank Adaptation): small trainable adapter matrices injected into attention projections; base weights stay frozen
- 8-bit quantization (bitsandbytes / LLM.int8) shrinks the base model's memory footprint
- Combined, these make large-model fine-tuning feasible on a single 16GB GPU



[FIGURE: LoRA]

Why synthetic data

- No open, large-scale Arabic dysarthric speech corpus is available to fine-tune on
- Real clinical recording and annotation is slow, costly, and ethically sensitive
- Approach: perturb healthy Egyptian Arabic speech (FLEURS ar_eg) with a deterministic DSP chain to approximate dysarthric characteristics
- Enables controlled, reproducible experimentation before any clinical data collection is attempted



PIPELINE

System pipeline overview

1

Local CPU (MacBook)

generate_data.py builds the synthetic dysarthric corpus

2

Colab GPU (T4 / L4, 16GB)

LoRA + 8-bit fine-tuning of Whisper large-v2

3

Evaluation

Word Error Rate via Hugging Face evaluate

Original FLEURS transcriptions are kept throughout, and only the audio is distorted.

[FIGURE: architecture]

DSP dysarthria simulation (fixed chain)

1	**Tempo stretch**	rate = 0.75 (slow speech, pitch preserved)
2	**Pitch shift**	−1.5 semitones
3	**Fluency cue**	mid-clip 100ms block repeated ×3 (stutter cue)
4	**Normalize & export**	peak normalize → 16kHz mono, 16-bit PCM WAV

[FIGURE: DSP flow]

Output

```
sample_0001.wav
...
sample_0500.wav
```

in synthetic_dysarthria_fleurs/

Deterministic — same input always yields the same output.

Dataset & split

- Source: Google FLEURS, Egyptian Arabic (ar_eg), open / ungated
- First 500 clips used, seed = 42
- Original FLEURS transcriptions kept as labels
- Only the audio is DSP-perturbed; text is untouched



50 val (10%)

500
total clips, 90/10 split

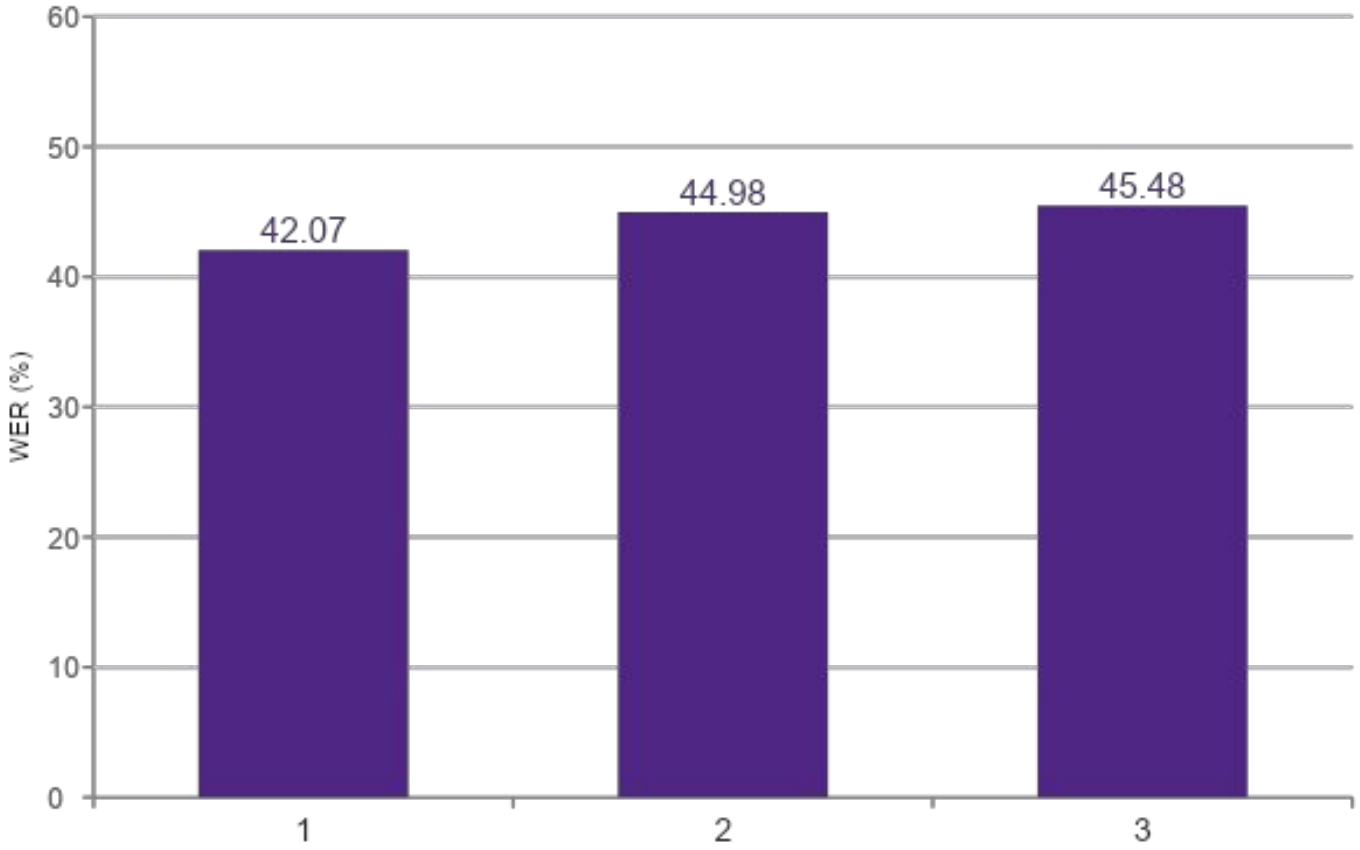
Experimental design: three controlled runs

Run	LoRA r / α / targets	Learning rate / schedule	Epochs	Checkpoint selection
Run 1	r=16, α=32 q_proj, v_proj	1e-3, fixed	10	Manual
Run 2	r=32, α=64 q/k/v/out_proj	3e-4, cosine + 0.1 warmup	15	Best eval_loss
Run 3	r=32, α=64 (same as Run 2)	3e-4, cosine + 0.1 warmup	15 planned (interrupted)	Best WER (predict_with_generate)

Same base model, dataset, and evaluation metric throughout; only capacity, learning rate, and selection criterion vary.

RESULTS

Best WER achieved per run



Lower is better

Run 1 wins

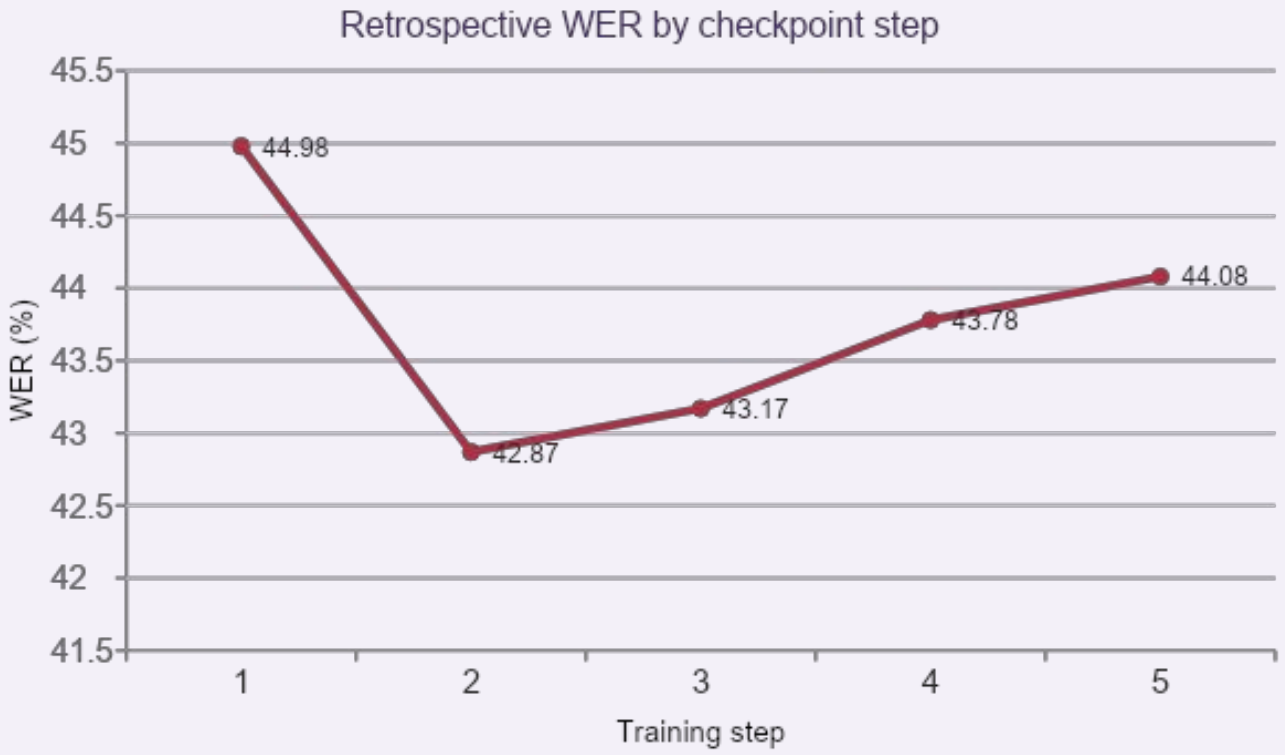
The smallest, simplest configuration, lower LoRA capacity, higher learning rate, and manually selected checkpoint achieved the lowest WER of all three runs.

~42% WER is high relative to clean speech but broadly consistent with dysarthric ASR being substantially harder in the literature (e.g., Vachhani et al., 2018; Mansour, 2026).

Western

DEEP DIVE

Run 2: eval_loss vs. WER diverge



Step	eval_loss	WER (%)
80 (min eval_loss)	~0.585	**44.98**
410 (best WER)	—	**42.87**
430	—	43.17
420	—	43.78
435	~0.79	44.08

Key finding: the checkpoint with the lowest eval_loss (step 80) was not the checkpoint with the lowest WER (step 410). Train loss fell to ~0.011 while eval_loss rose again by step 435. A classic overfitting signature that eval_loss alone does not translate cleanly into generation-time WER.

Western

[CHART: Run 2 divergence]

Why more capacity failed

- ~450 training samples is too small to support a doubled LoRA rank (r=32) without overfitting
- Higher capacity + lower learning rate reached near-zero training loss faster, not better generalization
- Teacher-forced eval_loss and free-running generation quality (WER) diverge under this overfitting
- Run 3 confirms WER-based selection is methodologically sound; it still did not beat Run 1, pointing to a data-scale limit rather than a flawed selection metric

FINAL MODEL

Selected model

Run 1 · checkpoint-87 · epoch 3

42.07% Word Error Rate

(For reference: same run, epoch 10 — the final scheduled epoch — measured 42.97% WER, slightly worse than epoch 3.)

- Simpler LoRA configuration (r=16) outperformed both higher-capacity configurations
- Manual, early checkpoint selection avoided the later-epoch overfitting seen in Run 2

Limitations

- Synthetic pathology simulation, not real dysarthric Arabic speech
- Tiny dataset (450 train / 50 validation clips)
- No real clinical test set used for evaluation
- Fixed, unvalidated DSP parameters (not checked against real dysarthric recordings)
- Run 3 interrupted at step 227 of 290 (Colab disconnect); trajectory was already worsening

Contributions & conclusions

- A controlled comparison of LoRA capacity, learning rate, and checkpoint-selection metric for Arabic dysarthric ASR under severe data scarcity
- Empirical evidence that eval_loss is an unreliable proxy for WER in this low-resource fine-tuning setting
- A reproducible synthetic-data generation pipeline and fine-tuning artefacts for future work

Thesis takeaway

"Under severe Arabic dysarthric-data scarcity, a compact LoRA setup selected early by WER (42.07%) beat higher-capacity runs, and validation loss alone was a misleading checkpoint criterion."

Future work

Validate DSP parameters

against real dysarthric speech samples, once available

Collect real clinical data

partner for genuine Arabic dysarthric recordings

Scale beyond ~500 clips

test whether higher-capacity LoRA helps with more data

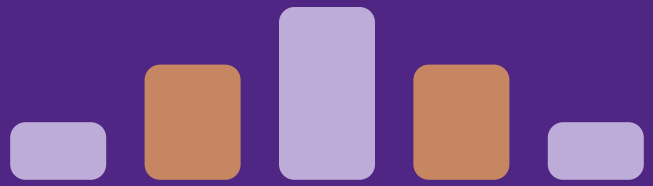
Broaden PEFT exploration

validate additional configurations once data scale allows

Artefacts

Artefact	Purpose
`generate_data.py`	Local corpus builder (DSP chain)
`synthetic_dysarthria_artefacts.ipynb`	Verify / regenerate the corpus
`Thesis_CoLab_Artefact.ipynb`	Colab training & evaluation
`synthetic_dysarthria_fleurs.zip`	500 generated WAV files

Run order: generate_data.py → verify notebook → Colab fine-tune (README included)



Western

Thank you

Questions & discussion

Hussein Mourad · CS Honours Thesis · Western University