

Semester 1 Project Proposal: Data Augmentation Strategies for Low-Resource Sudanese Dialect Automatic Speech Recognition (ASR)

Course: AI Master's Program - Semester Project

Student Name: Ahmed Elnoor Adam Ahmed

Student ID: 261003037

1. Project Overview

This project focuses on building a functional Automatic Speech Recognition (ASR) system specifically tuned for the Sudanese dialect. Current AI models for Arabic are largely trained on Modern Standard Arabic (MSA), which makes them ineffective for the everyday speech of Sudanese people. My project proposes a technical solution to the "Low-Resource" data problem by combining synthetic voice generation with self-learning algorithms.

2. Statement of the Problem

Sudanese Arabic is a linguistically rich dialect with deep historical roots, yet it is one of the most underserved in the digital world. The primary issue is a circular dependency: we cannot build good ASR without labeled data, but labeled data is hard to create because we lack the basic ASR tools to help transcribe it.

Existing public datasets for Sudan offer less than 5 hours of usable audio. To build a model that actually works for a person in Khartoum or Port Sudan, we need at least 50–100 hours of specialized data. My research asks: Can we "manufacture" this missing data using technology instead of waiting for a manual dataset to be built?

3. Objectives and Scope

The goal of this project is to take a pre-trained model (Whisper) and adapt it for Sudan using a three-pronged approach:

1. **Phonetic Mapping:** Mapping the unique phonetic differences in Sudanese speech (e.g., the specific 'Qaf' and 'Jeem' pronunciations).
2. **Synthetic Augmentation:** Generating high-quality synthetic audio from written Sudanese text to "pre-train" the model.
3. **Self-Training Pipeline:** Developing a system that can take raw YouTube or radio audio and generate its own training labels (Pseudo-labeling).

4. Proposed Technical Methodology

4.1. Data Strategy

- The Seed: A small, 5-hour dataset of real Sudanese speech (labeled).
- The Expansion: Using Text-to-Speech (TTS) engines to generate 40 hours of synthetic Sudanese audio.
- The Unlabeled Pool: 100 hours of raw, untranscribed audio from Sudanese social media and broadcast archives.

4.2. Implementation Phases

1. Base Calibration: Evaluating the "Zero-Shot" performance of OpenAI's Whisper model on Sudanese speech to identify the most common error patterns.
2. Bootstrap Training: Fine-tuning the model on the synthetic (TTS) data. This helps the model learn the "vibe" and vocabulary of the dialect.
3. Iterative Pseudo-Labeling: Using the bootstrapped model to transcribe the 100 hours of raw audio. We will keep only the transcriptions with high confidence scores and use them to retrain the model.
4. Final Optimization: Fine-tuning with LoRA (Low-Rank Adaptation) to ensure the model remains lightweight and efficient.

5. Potential Impact

A working Sudanese ASR system is more than just a tech demo; it is a gateway to digital inclusion. It could enable localized customer service bots, automated subtitle systems for local media, and voice-assisted technologies for the visually impaired in Sudan.

6. Project Timeline (Estimated)

- Week 1-2: Literature review and final data collection.
 - Week 3-4: Baseline testing and TTS data generation.
 - Week 5-7: Iterative self-training and pseudo-labeling.
 - Week 8: Final evaluation, WER analysis, and report writing.
-

7. Ethical Considerations

All audio data used for training will be sourced from public domains or used with permission. We will ensure that the model does not propagate biases present in the raw data and that any pseudo-labeled data is vetted for toxic content.

8. Selected References

1. Ahmed, A. (2026). Enhancing ASR for Sudanese Dialects. Research Notes.
2. OpenAI. (2023). Robust Speech Recognition via Large-Scale Weak Supervision.
3. Ali, A. (2022). Dialectal Arabic Speech Recognition: Challenges and Opportunities.
4. Zhang, Y. (2024). Low-Rank Adaptation for Large-Scale ASR Models.
5. Hassan, M. (2024). Synthetic Data for Low-Resource Languages.