Capstone Project Progress During Semester Break

## **1. Introduction**

During the semester break, I focused on advancing my capstone project, **Somali Accent Database (SomaliAccentDB)**, which aims to collect, structure, and analyze voice recordings from Somali speakers. The project's goal is to create a **well-organized dataset** that can be used for linguistic studies and AI-based accent analysis. In addition, I explored **existing Kaggle datasets** to learn from their data structuring techniques and applied these insights to improve my project.

## **2. Work Completed**

### **A. Data Collection**

* Contacted **150+ participants** from different Somali regions.
* Guided participants to record voice samples in a **controlled format**.
* Ensured diversity in **gender, age, and regional dialects**.
* Organized files systematically using **unique identifiers**.

### **B. Metadata Organization**

* Created a **CSV metadata file** to store:
  + **Speaker information** (age, gender, region).
  + **Sentence ID** and text content.
  + **Audio file path** for easy reference.
* Ensured **consistent formatting** in data storage.
* Maintained a structured **file-naming convention** to streamline access.

### **C. Audio Processing**

* Used **Audacity** to process recordings:
  + Removed noise and normalized audio levels.
  + Segmented long recordings into **individual words**.
* Established a **folder hierarchy**:
  + unprocessed \_Audio/ for unprocessed recordings.
  + Processed\_Audio/ for cleaned and segmented files.
  + Metadata/ for CSV and documentation.

### **D. Learning from Kaggle Datasets**

I analyzed two Kaggle datasets related to **speech and accents**:

1. **Global Speech Accent Recognition Dataset**
   1. Structure: Organized metadata with speaker information and accent labels.
   2. Takeaway: Inspired my approach to structuring metadata, including accent-based categorization.
2. **Speaker Accent Recognition Dataset**
   1. Structure: Featured structured metadata with audio file references and speaker demographic details.
   2. Takeaway: Helped refine the CSV file format, ensuring all essential speaker attributes were included.

**Implementation:**

* Applied **structured CSV metadata** similar to Kaggle datasets.
* Created a **clear folder structure** to avoid data fragmentation.
* Used **speaker IDs** and **sentence mappings** to improve retrieval efficiency.

### **E. GitHub Repository Setup**

* Created **GitHub Repository**: [SomaliAccentDB](https://github.com/Jeylani-2526/SomaliAccentDB).
* Uploaded **metadata files** and **processed audio samples**.
* Wrote a **detailed README.md** covering:
  + Project description
  + Dataset structure
  + Tools used (Excel and Audacity)
  + Future expansion plans

## **3. Challenges and Solutions**

|  |  |
| --- | --- |
| **Challenge** | **Solution** |
| Managing large amounts of voice recordings | Implemented a structured folder system & linked metadata files to audio paths |
| Ensuring metadata consistency | Followed Kaggle dataset structures for formatting |
| Need for scalable storage & accessibility | Organized audio files with unique IDs & mapped them in CSV |

## **4. Future Work**

* **Expand the dataset** by collecting more voices from participants and reaching a total of 200.
* **Finalize the dataset structure**, ensuring consistency and accessibility.
* **Analyze collected data** to identify patterns in Somali accents.

## **5. Conclusion**

During the semester break, I made significant progress in **data collection, metadata structuring, and repository setup**. Learning from **existing Kaggle datasets** helped me refine the **organization and accessibility** of my dataset. Moving forward, I will focus on **completing data collection, finalizing the dataset structure, and exploring patterns within the data** to enhance its value for researchers.