

Task-Lens: Cross-Task Utility Based Speech Dataset Profiling for Low-Resource Indian Languages

This repository contains the supplementary data and analysis code for the paper "**Task-Lens: Cross-Task Utility Based Speech Dataset Profiling for Low-Resource Indian Languages**" ([arXiv:2602.23388](https://arxiv.org/abs/2602.23388)).

Overview

The rising demand for inclusive speech technologies amplifies the need for multilingual datasets. This project investigates the cross-task utility of existing Indian speech datasets to alleviate the data scarcity challenge. We assess the readiness of 50 Indian speech datasets spanning 26 languages for nine different downstream speech tasks (e.g., ASR, TTS, LID, Speaker Recognition).

Contents of the Supplementary Material

Dataset Profiling & Metadata (CSV Files)

- **Dataset Languages.csv**: Contains comprehensive metadata for the 50 analyzed datasets, including total duration, unique voices, sampling rates, transcription types, licensing, and annotation granularity.
- **Dataset Languages - Languages.csv**: Provides a detailed, language-wise breakdown of speech duration (in hours/minutes) available across the different datasets.
- **feature_booleans.csv**: A boolean matrix mapping specific dataset properties and metadata features (labeled **f1** to **f12**) to each surveyed dataset.
- **task_readiness_summary.csv**: A cross-task readiness matrix indicating whether a dataset contains the necessary properties and metadata to be utilized for nine specific downstream tasks (ASR-M, ASR-ML, LID, SV/SID, Deepfake Detection, SER, TTS-M, TTS-ML, GI).
- **language_task_hours.csv**: Summarizes the cumulative hours of speech data available per Indian language for each of the nine profiled downstream tasks.

Code & Analysis

- **utility.ipynb**: A Jupyter Notebook containing the Python code (utilizing **pandas** and **numpy**) used to parse the CSV metadata, evaluate dataset eligibility across tasks, and compute the cross-task statistics presented in the paper.

Usage

To run the analysis notebook locally:

1. Ensure you have Python 3 installed along with `pandas`, `numpy`, and `jupyter`.
2. Clone this repository or download all files into a single directory.
3. Open `utility.ipynb` and run the cells. The notebook will automatically read from `Dataset Languages.csv` and the other provided CSV files in the same directory.

License

Creative Commons Attribution Non Commercial 4.0 [cc-by-nc-4.0]