Language Recognition from Audio for Pan Indian Languages
Project Proposal : CSE343 Machine Learning, Monsoon 2024, IIIT-Delhi

Ritika Thakur 2022408 | Saksham Singh 2022434 | Sarthak Gupta 2022451 | Sidhartha Garg 2022499

Abstract

This project aims to create an advanced system for recognizing Pan Indian languages by leveraging classical machine learning techniques. Starting with an extensive dataset of audio recordings, we focus on extracting rich acoustic features that capture the unique phonetic signatures of each language. By applying and refining multiple models, our approach demonstrates significant potential in accurately classifying a wide range of Indian languages.

1. Motivation

India's rich linguistic diversity, with 23 official languages and countless dialects, presents unique challenges for language recognition systems. Effective language identification from audio is essential for improving communication technologies and making digital platforms more accessible in multilingual environments.

We want to develop a language recognition system for Pan Indian languages using classical machine learning techniques on a comprehensive dataset of audio recordings. By analyzing acoustic features like pitch and spectral properties, we aim to build models capable of accurately classifying languages from audio samples.

2. Related Work

- **Spoken Language Identification using Gaussian Mixture Model-Universal Background Model in Indian Context** by Sreedhar Potla, Vishnu Vardhan B.: The study identifies one of twenty-three Indian languages using Mel-Frequency Cepstral Coefficient (MFCC) features, showing improved accuracy with a GMM-UBM model compared to a standard GMM.
- **Language Identification based on Auditory and Vocal Characteristics** by Hua Ying-jie, Duo Lin : An auditory-based feature extraction algorithm using Gammatone filters and ERB model-derived features outperforms MFCCs in language identification.
- **Language Identification Using Gaussian Mixture Models** by Calvin Nkadimeng : Language Identification for African languages using GMMs is feasible but requires addressing classification challenges and exploring phonetically transcribed methods for more rigorous research

3. Timeline

We plan to use a dataset, created by Chaitanya Bharadwaj H B on Kaggle, containing approximately 257K five-second audio samples evenly distributed across 10 Indian languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, and Urdu. The audio samples were sourced from regional videos on YouTube.

We plan to divide the project timeline in 2 phases (The given timeline is subject to change based on what we learn in the course) :

- **Phase 1:** To be able to implement the pre-processing on data to extract features relevant to the model, and visualise the data. Implement a GMM/ SVM model to judge accuracy on the data, and visualise the results.
- **Phase 2:** To implement and explore other models like random forests and Neural Networks to increase accuracy.

4. Contribution and Division of Work

We will try to equally divide all the work in each phase and the same will be reflected throughout our reports.

5. Final Outcome

In a world with diverse cultures and rapid technological advancements to bridge these gaps, this project can serve as the foundation for translation and transcription applications in multi-lingual settings.

The project can also provide insights into various linguistic features, such as tonality and frequency, for the specified languages. This data can be critical for the field of linguistics.