Building a pipeline for processing and understanding audio resources on Sandalwood cultivation

Problem description

Sandalwood has played an important role in Indian cultural, religious, and therapeutic practices for a long time. It is widely cultivated in the state of Karnataka and is one of the expensive crops. It is important to capture and conserve any indigenous knowledge about sandalwood and also facilitate conservation efforts aimed at improving the population of the tree. This task aims at creating a workflow that utilizes Automatic Speech Recognition (ASR) for a speech corpus, and makes the corpus available to support querying for relevant information based on questions provided by a layman.

Dataset

The dataset is a corpus of audio files with content related to sandalwood cultivation, mostly in Kannada language. The dataset was compiled by scraping YouTube. The content of these audio files are in colloquial-style language which can differ significantly from the formal-style language. The audio files can also contain minor noise as they were recorded in public places.

The dataset can be downloaded from here

Task 1: Speech Recognition
Develop a ASR model for colloquial (Kannada) language using any foundation model. The final model can be used for the next task. The team can use the provided Sandalwood dataset for activities such as fine-tuning. Additionally, teams can also use any other publicly available dataset focused on Sandalwood cultivation including audio content in dialects.

Task 2: Speech based Question-Answering
Develop a pipeline to do the following inference,

1. A question will be asked by the user in the form of speech.
2. The provided audios will be searched and the segment where an appropriate answer for the input question exists will be returned to the user. Teams can use the developed ASR model to convert both the input question and the audios to text format to facilitate searching.