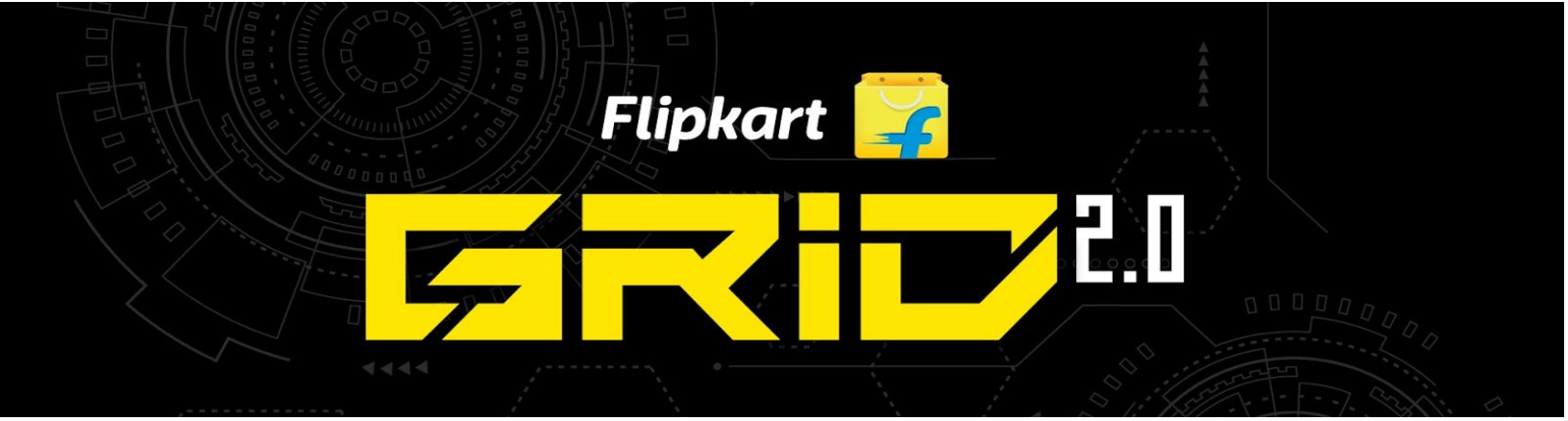


India is a country of multiple languages and dialects.The Next 200 million consumers from Bharat have different dialects and a huge potential to solve. While working on recognizing these accurately, ASR models have to be robust for handling various sources of audio.

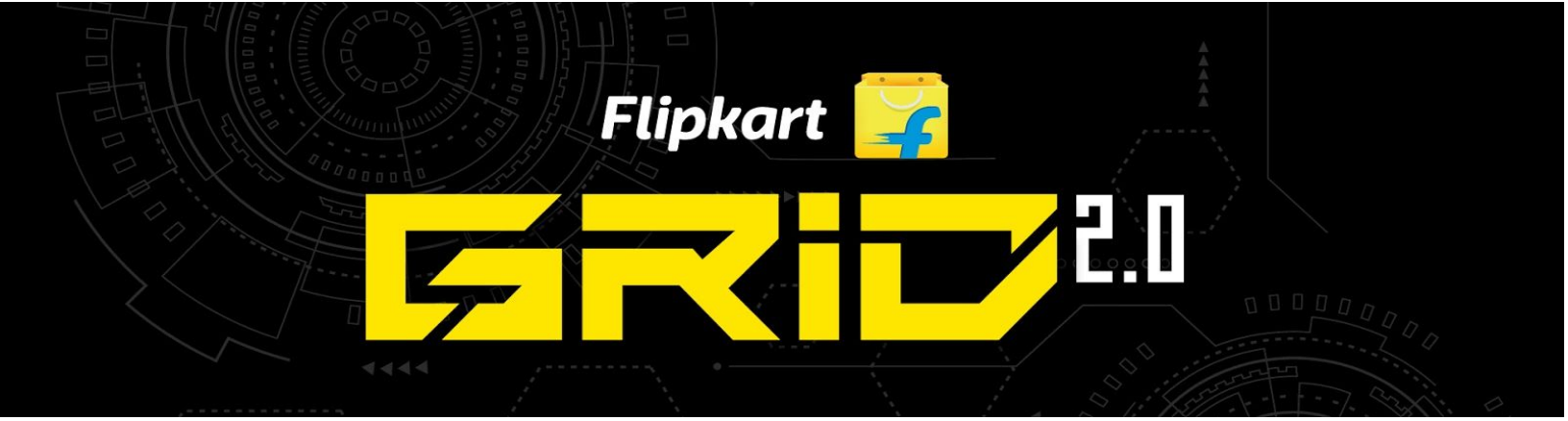
Brief

Imagine a user shopping on an Ecommerce App using a voice-based conversation assistant which is a feature available within the Application. Now while the user is sitting on a sofa in the living room of their house and browsing for fashion clothes on the Application by speaking into his/her phone, the TV in the room is blasting a movie song. For another user who is also shopping on an App using a voice assistant, his friend is also in the same room talking on his phone with someone else. It is a very common occurrence in Indian houses and neighborhoods to hear the voices of multiple people in the same area.

While it is easy and natural for humans to identify dynamic sounds and differentiate between the voice from the friend you're speaking with on the phone and a secondary sound coming out of the TV speaker in the room, how do we make machines do the same job for us?

Problem Statement

As a speech recognition system, you should be able to identify and separate out speech utterances from the user who is using voice features for shopping from the other speeches which may be present in the background of the user using the voice assistant. In order to serve the shopping intent of the user correctly, It is important for the speech recognition system to only focus on the speech of the user who is shopping and not mix it with other speech utterances also captured due to the background.



Deliverables/Expectations for Phase I (Idea Submission)

An algorithm/approach with block diagrams and detailed explanation to accomplish the below:

Given an audio which may or may not have background audible speech -

1. Identify if there are more than 1 speakers in the audio

2. Identify and separate the primary speaker audio data based on a classifier that is either learnt from data and/or assisted by some hand-engineered features.

Please provide definitions and working of the core components of the system. You can give references if any part of work is inspired by some previous work.

The solution should work for both a returning user as well as a new user. Considerations of different settings, edge cases will be given extra points.

Background speech may be from TV, music or humans. You can assume the impact of environmental noise like traffic, wind or other non-human household noise to be minimal. All other assumptions should be clearly stated.

The solution should run in real time. So a brief discussion on computational complexity would be expected.

-- End of Document--