**Survey on speech to text Deep learning Architecture**

What is an audio data analysis?

An audio data is a data having formats ‘.wav’, where the amplitude of the audio changes with respect to time. Since sounds are represented in audio signal with parameters frequency, bandwidth and decibel, using amplitude and time we can express an audio signal. Audio data analysis is about analyzing and understanding audio signals captured by digital devices, with numerous applications in the enterprise, healthcare, productivity, and smart cities.

What is is speech recognition?

Speech recognition is the process of converting human sound signals into words or instructions. It is based on speech and an important research direction of speech signal processing and a branch of pattern recognition. It aims to learn how to take the input audio and predict the text content of the words and sentences that were uttered.

Inputs of speech recognition system

\*\*Audio:\*\* The sound excerpts are digital audio files in .wav format.

\*\*Transcription:\*\*A text file corresponding to the given audio file

In this article, I will focus on how the Automatic Speech Recognition (ASR) of Speech-to-Text using deep learning works.

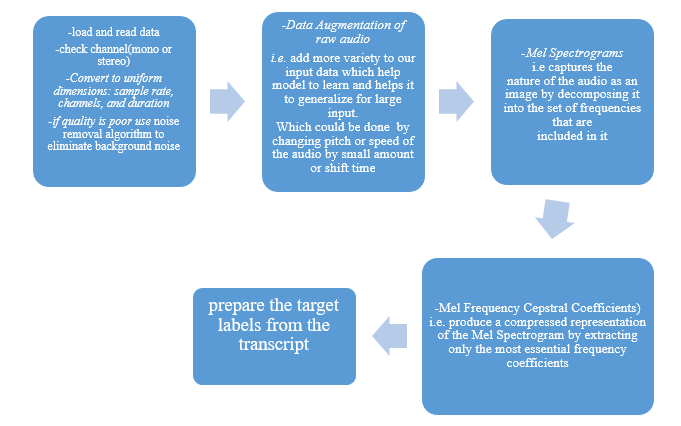
###Speech to text###

For speech to text recognition the training data contains:

* Input features (*X*): audio files of spoken words
* Target features (Y): the text transcripts

In our datasets we have audio .wav files for input features and Amharic texts for target.

*Data pre-processing*



**Deep learning Architectures**

Deep learning architecture consists of deep/neural networks. The neural networks are based on three layers that advance data. The first is an input layer (raw data), second hidden layers (process and combined to the input data), and last is an output layer (which produces aftermath: result, estimation, forecast, etc.). The most basic architecture deep learning in ASR (Automatic speech recognition) Systems are:

1. Convolutional Neural Networks (CNN)

Takes an input image, assign importance to various aspects/objects in the image, and be able to differentiate one from the others. Mostly used for image processing, image recognition, video analysis. It contain an input and an output layer and multiple hidden layers.

1. Recurrent Neural Networks (RNN)

Commonly used in speech recognition. It uses internal memory to process variable-length sequences of inputs so that each processed information is captured, stored, and utilized to calculate the final outcome. Also it may have connection with feedback which allows to maintain the memory of the past and solve problems in case for present.

There are two types of RNN:

* 1. Bidirectional RNN: which works in two ways. That is the output layer can get information from past and future states simultaneously.
  2. Deep RNN: Multiple layers are present and help us to extract more hierarchical information.

1. Long Short-Term Memory (LSTM)

It help us to process sequence of audio/video datasets and has feedback connection. It is also based on memory cell where its values as input are put as function of short or long lime so that it helps to memorize and remember.

1. Deep Neural Networks (DNN)
2. DNN-HMM Hybrid Systems for ASR