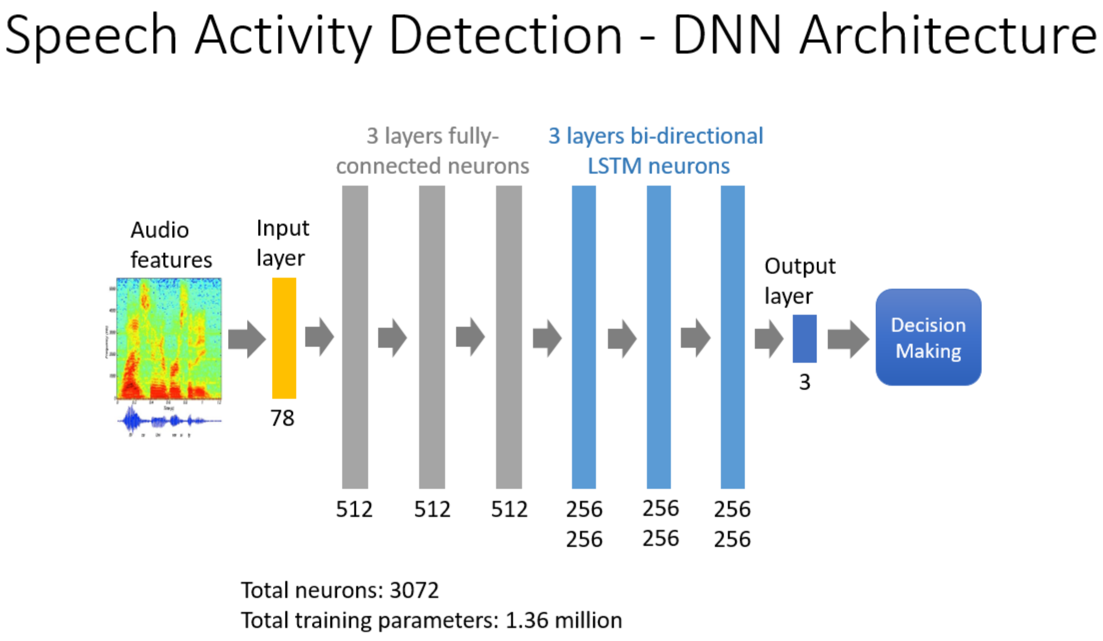
Introduction

The Speech Activity Detection Engine is an audio extraction tool that distinguishes non-speech audio inputs; e.g. music, background-noises, silence, among others.

This document explains the architecture of the model.

Deep Neural Network Architecture



Speech Activity Detection - DNN Architecture

Total neurons: 3072
Total training parameters: 1.36 million

- Consist of
  - Input Layer
    - Number of Inputs which are pre-specified by the available data
  - 2 different sets of neurons
    - 3 Layered Class Fully Connected Neurons (Fire Neural Network)
      - Classified into distinctive layers of neurons
      - Neurons between two adjacent layers are fully pairwise connected while neurons within a single layer share no connections
    - 3 Layered Bi-Directional Long Short-Term Memory (LSTM)
      - An extension of traditional LSTMs that improves sequence classification problems
      - Bi-directions LSTMs runs two LSTMs on the input sequence, one *from the past to the future* and the other *from future to past*
      - Preserves information from both past and future
      - Provides additional context to the network, understand it better which results in faster and fuller learning

- Output Layer
  - Has a Linear identity activation function which the last output layer is usually taken to represent the class scores based on the type of classifications
  - Classification scores are arbitrary real-valued numbers, e.g. in regression
- Decision Making Model
  - A Model that cleans up noise signal from the output of the neuron’s classification