Voice Recognition

Introduction:

Voice recognition is a computer software program or hardware device with the ability to decode the human **voice**. **Voice recognition** is commonly used to operate a device, perform commands, or write without having to use a keyboard, mouse, or press any buttons. Voice Recognition is also known as software application that has ability to recognise human voice. Voice Recognition converts words and sentences into machine acceptable format. Voice recognition has gained prominence and use with the rise of AI and intelligent assistants, such as Amazon's Alexa, Apple's Siri and Microsoft's Cortana. This technology helps users by google searching, setting alarm, getting information about weather, playing their favourite music using voice command giving a hands-free experience.

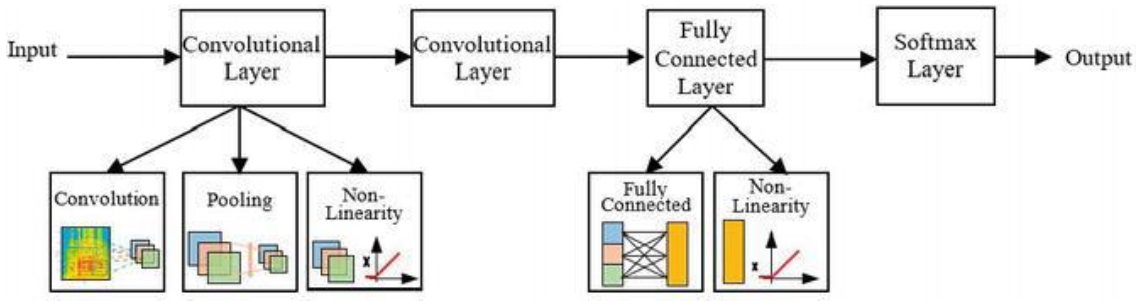
How does it work ?

- Voice recognition converts Analog signals into digital signals, known as Analog-to-Digital converter. Librosa is a in build library in python which will help to convert Analog signal into digital. It will convert your voice file(.wav file) into array.
- The first step in speech recognition is obvious, we need to feed sound waves into a computer. Sound waves are a one-dimensional. At every moment in time, they have a single value based on the height of the wave.
- To turn this sound wave into numbers, we just record of the height of the wave at equally-spaced points. This is called sampling.
- We are taking a reading thousands of times a second and recording a number representing the height of the sound wave at that point in time. Audio is sampled at 44.1khz(44,100 reading per second)
- Once your data is in machine acceptable format we have to do feature extraction which is done using 'MFCC' or 'Mel-spectrogram'. After this training on this dataset will be done. Neural network is used for training purpose. Sequential model is used in most cases for voice recognition system. In neural network their are basically three layers.

- Input **Layer**: Input variables, sometimes called the visible **layer**.
- Hidden **Layers**: **Layers** of nodes between the input and output **layers**. There may be one or more of these **layers**.
- Output **Layer**: A **layer** of nodes that produce the output variables.

- The convolutional neural network (CNN) can be regarded as a variant of the standard neural network. Instead of using fully connected hidden layers as described in the preceding section, the CNN introduces a special network structure, which consists of alternating so-called convolution and pooling layers. We decided to test how CNN works for speech data. With applications ranging from speech controls for online games to issuing commands to IoT devices, classifying speech data has a lot of charm.
- Long Short-Term Memory (LSTM) layers are a type of recurrent neural network (RNN) architecture that are useful for modeling data that has long-term sequential dependencies. This context is useful for speech recognition because of its temporal nature.

- Neural Network Structure:



Dataset is separated into 80:20 ratio. 80% for training and 20 % for testing.

After training the dataset new data is fed for prediction.

Application:

1. Voice Search: Most of us know how to use Siri and Google voice search on our phones – it is as easy and natural as asking our moms to do something for us. Although not perfect, smartphone voice assistants are getting better at recognising obscure vocabulary, managing accents and understanding many languages. As the tech gets better, so do smart speaker sales and predictions about their future proliferation, with some sources claiming 30% – 50% of all web searches are to be done screen-free within the next 2-3 years.

2. Text processing: Thanks to advanced voice recognition tech, we no longer have to type long emails, text messages or documents – we can simply speak our thoughts into our smartphones, tablets or computers' microphones – text is magically typed in by the apps. This is not only handy while driving, cooking or otherwise having your hands occupied, it's also great for multi-tasking and saving time.

Advantages:

- Increases productivity.
- Can help with menial computer tasks, such as browsing and scrolling.
- Can help people who have trouble using their hands.
- The software learns to recognize a Doctors unique speech patterns.
- The software spells every word correctly.
- Doctors can write as quickly as they speak, 100+ words per minute.
- Voice recognition software use is expanding rapidly. Both Windows and Macintosh operating systems have voice recognition built in.

Disadvantages:

- Can be hacked with pre-recorded verbal messages.
- Less accurate when there is background noise.
- The software has to be trained to recognize the user's voice. This is accomplished by reading passages provided by the program.
- Doctors have to speak distinctly in order for the software to work well. If the Doctor has non-standard speech, tends to run words together, or mumble, the training process may be long. Some punctuation must be dictated.
- The software spells every word it recognizes correctly. Typically, it recognizes 5–20% words incorrectly. It cannot recognize homonyms.
- Voice recognition uses a lot of memory. The software has specific hardware requirements.