Speech Recognition

Ashutosh Upadhye, SST Siddhardha, Rahul Dhawan

March 2018

This report includes the implementation procedures of our Project as part of TPDS course at the Indian Institute of Technology, Palakkad. This report includes the introduction section, literature survey, implementations followed by results and conclusions.

1 Introduction

Speech recognition is one of the most recently developing field of research at both industrial and scientific levels. Until recently, the idea of holding a conversation with a computer seemed pure science fiction. If you asked a computer to “open the pod bay doors”—well, that was only in movies. But things are changing, and quickly. A growing number of people now talk to their mobile smart phones, asking them to send e-mail and text messages, search for directions, or find information on the Web. The first step towards the world of speech recognition problem is digit recognition.

2 Problem Statement

For a given audio file in which the speaker will speak a sequence of numbers, recognize the number being said.

3 Literature Survey

Books are available to read and learn about speech recognition ,these enabled us to see what happens beyond the code.

”Automatic Speech Recognition: A Deep Learning Approach” (Publisher: Springer) written by D. Yu and L. Deng published near the end of 2014, with highly mathematically-oriented technical detail on how deep learning methods are derived and implemented in modern speech recognition systems based on DNNs and related deep learning methods.This gave us an insight into the conversion algorithm used by Google.

Here are some IEEE and other articles we referred :

- Waibel, Hanazawa, Hinton, Shikano, Lang. (1989) ”Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech and Signal Processing.”
- Reynolds, Douglas; Rose, Richard (January 1995). ”Robust text-independent speaker identification using Gaussian mixture speaker models” (PDF). IEEE Transactions on Speech and Audio Processing (IEEE) 3 (1): 72–83. doi:10.1109/89.365379. ISSN 1063- 6676. OCLC 26108901. Retrieved 21 February 2014.

4 Model and Algorithm

4.1 Overview of the model

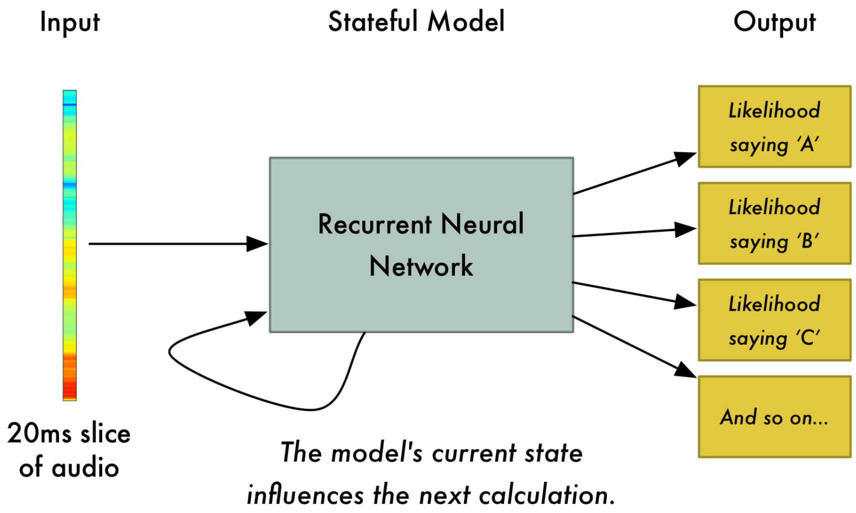

Figure 1: Overview

4.2 Step 1: Acoustic Features for Speech Recognition

STEP 1 is a pre-processing step that converts raw audio to one of two feature representations that are commonly used for ASR.

4.2.1 Spectograms:

A spectrogram is a visual representation of the spectrum of frequencies of sound or other signal as they vary with time. Spectrograms are sometimes called spectral waterfalls, voiceprints, or voicegrams.

Spectrograms may created from a time-domain signal in one of two ways: approximated as a filterbank that results from a series of band-pass filters, or calculated from the time signal using the Fourier transform. These two methods actually form two different time–frequency representations, but are equivalent under some conditions.

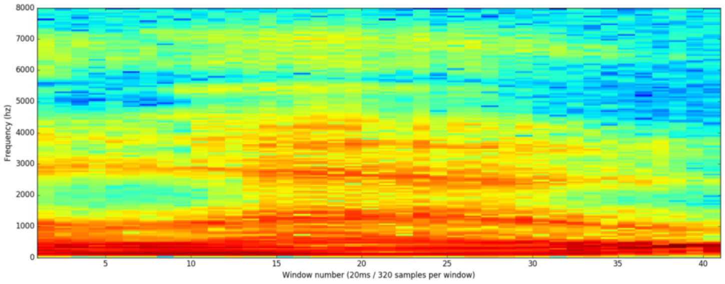

Figure 2: Normalized Spectogram

4.2.2 Mel-Frequency Cepstral Coefficients (MFCCs):

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone.

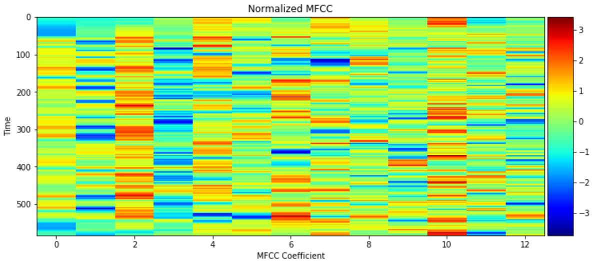

Figure 3: Normalized MFCC

4.3 Step 2: Deep Neural Networks for Acoustic Modeling

STEP 2 is an acoustic model which accepts audio features as input and returns a probability distribution over all potential transcriptions.

4.3.1 Recurrent Neural Networks:

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed graph along a sequence. This allows it to exhibit dynamic temporal behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

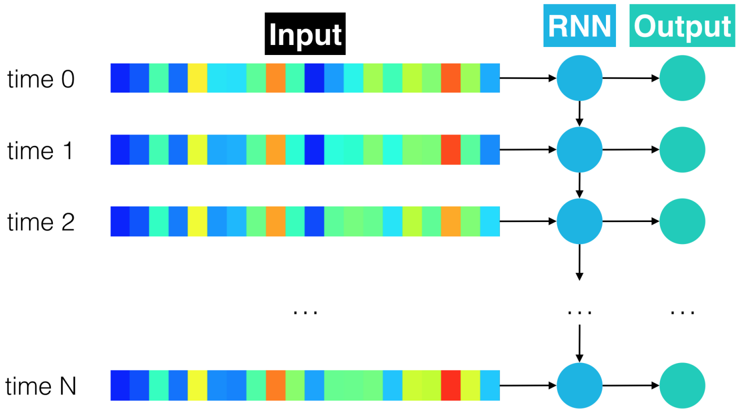

Figure 4: Recurrent Neural Network

4.4 Step 3: Obtaining Predictions

STEP 3 in the pipeline takes the output from the acoustic model and returns a predicted digit. This is rather trivial.

5 Data Collection and Implementation

For this project we used the audio files with spoken numbers as input and trimmed the files to maintain the consistency. These trimmed files are used to generate the final dataset.

6 Experiments

To be added.