Introduction to Automatic Speech Recognition

OUKIT Basma [1], MAHMOUDI Abdelhak [2], HIMMI Mohammed Majid [1]

[1] LIMIARF, Faculty of Sciences, Mohammed V University, Rabat, Morocco, {basma.ouk94 , himmi.fsr}@gmail.com
[2] LIMIARF, Ecole Normale Supérieure, Mohammed V University, Rabat, Morocco, abdelhak.mahmoudi@um5.ac.ma

Abstract

Automatic speech recognition (ASR) is the process of converting speech signals automatically into text. It can be used in diverse environment for different purposes such as in medical records, radio station, etc. Researches in ASR has been done for many years, starting with Hidden Markov Models, Dynamic Time Wrapping, Support Vector Machine and many others, but since the emergence of Deep Artificial Neural Network (Deep ANN), speech recognition research has been upgraded.

Deep Learning methods offer significantly lower speech recognition error rates compared to the traditional methods. We will review the pipeline of ASR from sampling audio and feature extraction techniques to the theory and implementation of the Recurrent Neural Networks (RNN) architecture well suited for ASR.

The aim

Showing the steps of building an automatic speech recognizer using Deep Recurrent Neural Networks

Objectives

The main objectives are reviewing and studying the existing and current speech recognizing systems to be able to master the main steps of ASR. As well as to fully comprehend the preprocessing phase of audio signals, to transform theses audios to valid input to a well structured Deep Neural Network architecture.
And also assimilating how Deep Neural Networks work and how they perform.

Problem Context

Some of the problems that we encountered are lack of Data ,the limited processing power of the computer's microprocessor.
As many studies examined the progress made in implementing voice recognition. There is still a lot to develop for Arabic language

Methods

1-Sampling

- Each audio sample contains data that provides the information necessary to accurately reproduce the original analog waveform

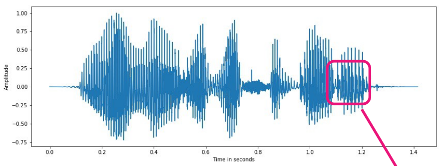


Figure1 : Speech Signal

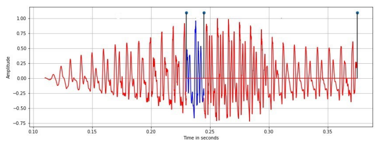


Figure 2 : Signal sample

- Audio data should be represented in frequency domain by using a variation of Fourier Transform
- We obtain then the spectogram

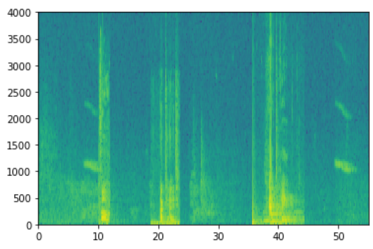
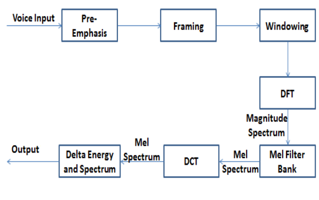


Figure 3 : Signal specter

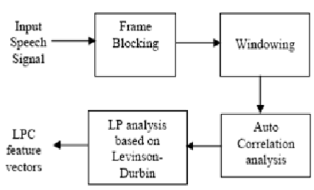
1-Feature Extraction

Feature extraction helps reducing input size

-Mel Frequency Ceptstral Coefficients(MFCC) :It is one of he most dominant method to extract cepstral features. It is based on the known variations of the human ear's critical bandwidths with frequencies which are below a 1000 Hz. First we Frame the signal into short frames.For each frame we create the it magnitude spectrum. Then we apply the mel filterbank to the power spectra,Take the logarithm of all filterbank energies.Take the DCT of the log filterbank energies. Then we obtain MFCC vectors.



-Linear predictive coding(LPC) :The basic idea behind LPC analysis is that a speech sample can be approximated as linear combination of past speech samples. It provides auto-regression based speech features. The speech signal is approximated as a inear combination of its p previous samples.The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech signal and the estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients.



3-Reccurent Neural Network-LSTM

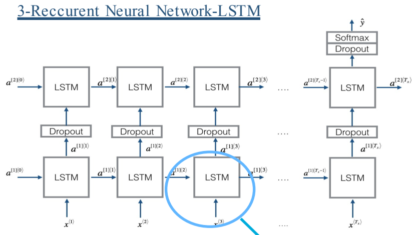


Figure 4 : Unidirectional multilayer LSTM

As a test of LSTM on speech recognition ,we put a shimmer sound whenever the word "Activation" is detected. Since RNNs are not good at catching long term dependencies because of gradient descent vanishing problems, Long Short Term Memory(LSTM) were introduced as a solution. After sampling and collecting inputs from feature extraction, we feed these inputs to LSTM model.

The intuition behind LSTM, is that it contains tree gates, the Figure 6 illustrates a single LSTM memory cell. For the version of LSTM used in this paper, it is implemented by the following functions. The first step in our LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer".The next step is to decide what new information we're going to store in the cell state .Finally, we need to decide what we're going to output. This output will be based on our cell state, but will be a filtered version.

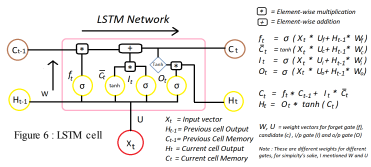


Figure 6 : LSTM cell

Printing the model:

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 5511, 101)	0
conv1d_1 (Conv1D)	(None, 1375, 196)	297136
batch_normalization_1 (Batch	(None, 1375, 196)	784
activation_1 (Activation)	(None, 1375, 196)	0
dropout_1 (Dropout)	(None, 1375, 196)	0
gru_1 (GRU)	(None, 1375, 128)	124800
dropout_2 (Dropout)	(None, 1375, 128)	0
batch_normalization_2 (Batch	(None, 1375, 128)	512
gru_2 (GRU)	(None, 1375, 128)	98688
dropout_3 (Dropout)	(None, 1375, 128)	0
batch_normalization_3 (Batch	(None, 1375, 128)	512
dropout_4 (Dropout)	(None, 1375, 128)	0
time_distributed_1 (TimeDist	(None, 1375, 1)	129

```
Total params: 522,561
Trainable params: 521,657
Non-trainable params: 904
```

Fitting the model:

```
Epoch 1/1
26/26 [==============================] - 10s 395ms/step - loss: 0.0893 - acc: 0.9716
```

Testing the model:

```
25/25 [==============================] - 2s 68ms/step
Dev set accuracy =  0.9291636347770691
```

Conclusion

In this paper we reviewed the basics of automatic speech recognition, with all the steps that are crucial for building it. First step is sampling then feature extraction and then feed our inputs into an LSTM.
We can say that building a speech recognizer is not simple work for mostly lack of Data, that leads to low performance and accuracy of the Deep Neural Architecture.

As a future work, we intend to implement more RNN models as well as combining these models for a better audio processing results. Gather as much data as possible .

References

A.J. Jerri. (1977) The Shannon sampling theorem—Its various extensions and applications: A tutorial review. Proceedings of the IEEE, pp. 1565–1596.

Kishori R. Ghule, R. R. Deshmukh. (2015) Feature Extraction Techniques for Speech Recognition: A Review. International Journal of Scientific Engineering Research, Vol. 6.

Pratik K. Kurzekar ., Ratnadeep R. Deshmukh ., Vishal B. Waghmare ., & Pukhraj P. Shrishrimal H (2014) A Comparative Study of Feature Extraction Techniques for Speech Recognition System. IJIRSET, Vol. 3.

Alex Graves., Abdel-rahman Mohamed ., & Geoffrey Hinton (2014)SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS. ICASSP.

DEEP LEARNING IndabaX Morocco April 29-30, 2019 Hassan II University of Casablanca Faculty of Sciences and Techniques (FST) of Mohammedia