INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 3)
Available online at: www.ijariit.com

Speech recognition system using Deep Neural Network

Dhanush N. D.
dhanushnd.ec16@rvce.edu.in
RV College of Engineering, Bangalore, Karnataka

Jagruth S.
jagruths.ec16@rvce.edu.in
RV College of Engineering, Bangalore, Karnataka

Rohini Hallikar
rohinish@rvce.edu.in
RV College of Engineering, Bangalore, Karnataka

ABSTRACT

Speech recognition is the property of a system to identify the words spoken by the user in a scripted language and convert the data to readable and writable format. The work carried out was able to convert the speech to text. Using the speech as instructions to perform many web-based services, system bound tasks. Deep learning with deep neural networks coded in python is implemented in this paper which makes the system more reliable, robust against noise and with accuracy of 70%.

Keywords— *Speech-Recognition, Deep neural network, Python, deep learning, noise-robust speech recognition.*

1. INTRODUCTION

Speech recognition is an important field which is emerging as a inter disciplinary as well as a sub field of computational linguistics which create technology and methods that empowers the acknowledgment and understanding of native language into a computer understandable language. This is most generally known as Automatic Speech Recognition (ASR). Neural network has a long history in speech recognition, and it is ordinarily in mix with Hidden Markov Models (HMMs). This have gotten interest recently with the sensational upgrades in acoustic modeling yielded by the entirely outstanding significant deep feed-forward frameworks.

The significant disadvantage of utilizing HMM is that if the expressions are equi-probable then the word would be mix of both the articulations which wasn't precise and is problematic. This prompted the utilization of deep neural networks. This system is implemented using Recurrent Neural Networks (RNN). RNN provide high accuracy and are highly reliable. This makes the speech recognition system highly efficient and can be relied upon to get better results.

2. METHODOLOGY AND IMPLEMENTATION

In the prepared dataset, it incorporates the input alongside the expected output for the comparing input. This gives the model ground truth information that is normally prepared by people else in a semi-mechanized way. During the training stage, present this data to train the model by comparing the output of the training data set with the actual data set for a given input. The test dataset contains data that is used to apply to the model. During the test period of the network, the data without expected output is utilized to estimate how well the model had been prepared and evaluated model properties. Figure 1 represents a flowchart which shows how a neural system is trained utilizing the dataset.

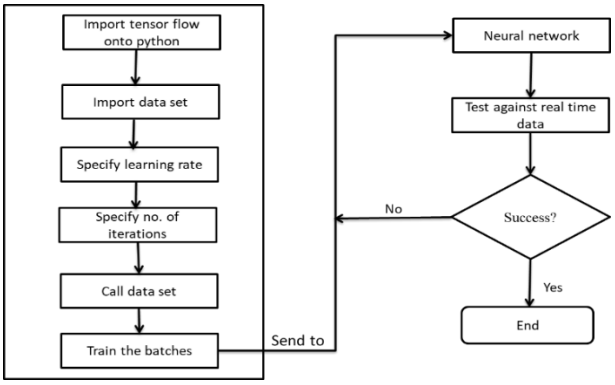

Fig. 1: Flowchart to train the neural network

1.1 Learning rate of a network

Consider a sigmoid function f(x) where the minimal of the derivative of that function is at point x =0 (x > 0), where the derivative is more than zero then user needs to go in negative (- ve) direction and when x < 0, derivative under zero one needs to consider in positive (+ve) direction. One constantly should make a step toward the direction which is inverse to the derivative. This is done by using gradient. Gradient is vector that points towards some direction in the space. It points towards the steepest increment of function. Since user will be in need to minimize the function, user must take a step in the other direction of the gradient. Variable with respect to which the graph is going to take the derivative or weights change since these are the qualities needed to change to improve the system. On the off chance that the gradient of the loss function regarding the weights and make little steps in the other direction of gradient, at that point the loss will bit by bit decline until it unites to some neighborhood negligible value.

1.2 Implementation

The speech assistant is developed by developing a code in python. If the speech is not recognized by the system, then it

© 2020, www.IJARIIT.com All Rights Reserved

will be looped back asking the user to speak again. If the speech is recognized, then it is converted to text. If the speech is recognized and if it is a command, then its run based on command and that command is displayed as text as well. Once the command is executed then the output is displayed.

Operating Systems module has been utilized to perform system bound tasks and the module gives an advantageous technique for utilizing operating system dependent functionality. If the user needs to peruse or compose a document see open() or need to manipulate paths? at that point check the operating systems path module and need to peruse all the lines in all the records on the command line? at that point check the file input module. For making brief records and reports, utilize the tempfile module and for critical level archives and index dealing see the shutil module.

The availability of these capacities is as depicted. The structure of all intrinsic operating system dependent modules of this programming language is with a definitive target is if an equivalent functionality is accessible it uses a comparable interface and subsequently decreases the overhead. Extensions bound to a specific operating system are similarly open from the operating system module yet utilizing these is obviously risk to probability. The internet browser module is used as it gives a significant level interface to permit displaying Web based records for us and this displays the contents URL using the default web browser and it is also designed to handle errors.

Implementation of speech recognition system is done using TensorFlow as a module on Python. A flowchart to set up and train a Recurrent Neural Network is drawn using Python. The assistant is developed using various modules that are pre-defined in python library.

1.3 Experimental evaluation

An alternative to utilizing a fixed learning rate, a flexible learning rate has been achieved by changing the parameters which indicate cyclic capacity (f) for example, triangular window. Each cycle has a fixed learning rate regarding the number of iterations. This method lets the learning rate cyclically change between predefined boundary. As the quantity of emphases builds, error rate diminishes. Be that as it may, an excessive amount of emphasis hinders the systems performance. From the Fig.2, number of emphases is chosen by considering the tradeoffs where the X-axis is the number of iterations and Y-axis is the Error rate in training set.

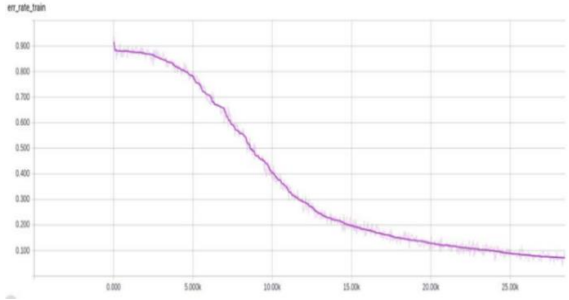

Fig. 2: Error vs iteration

2. LITERATURE SURVEY

An interesting teacher-student learning framework is proposed, which joins independent articulation improvement, e.g., IMCRA, and deep learning strategies at the preparation stage. The prepared understudy model can be exactly utilized for articulation upgrade as the preprocessor of ASR frameworks in the acknowledgment stage [1]

Discrete speech signals are unpredictable intermixing of various informative factors, and this information blending makes decoding any of the individual factors extremely troublesome. A characteristic thought is to factorize every speech outline into independent variables; however, it ends up being even more difficult than decoding each individual factor consequently, the author suggests Deep Neural Network (DNN) over HMM [2].

The speech enhancement techniques to improve the performance of voice control MRI. This is huge since errors in the presence of unwanted noise are more common and tend to make applications, for instance spoken dialogue systems, too lumbering to even consider using. The information signal is adulterated with MRI noise with changing signal-to-noise ratio. Proposing the use of recognition systems is to remove noise from the system [3]. Most speech emotion recognition systems utilize high-dimensional speech highlights, showing human feeling interpretation, to propel feeling recognition execution and demonstrates a move in the speech highlights from phenomes to feelings [4].

This paper [5] Markovian blend of language and prosodic models for better speech talk cognizance and acknowledgment is accomplished. For every condition notably, concealed Markov portrayals of the essential structure, related perceptions rise normally and awards existing speech recognizers to be gotten together with independent prepared prosodic classifiers. Fundamentally, communicating the downfalls of a typical HMM model and the arrangement of another kind of HMM model which works in two modes: to recoup shrouded structure (for instance, sentence limits), or to survey speech acknowledgment theories, in this way incorporating prosody into the recognition process [6].

Springer handbook of speech preparing, Springer Science and Business Media from standard client items for instance cell phones and MP3 players progressively current endeavors for instance, human-machine interfaces and responsive robots, speech technologies advancements are by and by everywhere. Many envision that it is simply a question of time before more uses of the investigation of speech become inevitable in the everyday life [8].

3. CONCLUSION

In this study, a RNN algorithm is used to help the performance and speed of the speech recognition system. An RNN is trained to get the desired accuracy by giving it more computational power and data to detect the word correctly. The speech recognition system can perform different web services such as opening a website, performing google search using speech as command, composing of speech to text converted data and sending an e-mail. The implementation of speech recognition will help widely in application such as mobile phone, home automation, self-driven cars and vending machine. Which will help improving the quality of living of everyone using this technology. The student teacher model was not able to be implemented due to the constraints in processing capacity. Further an Iterative Mask Estimation (IME)- based multi-channel speech redesign improving neural network can be gotten together with ordinary multi-channel speech upgrade algorithm. In [4] using Liquid State Machine (LSM) and Spiking Neural Network (SNN) to identify the emotions of the speaker. The framework used in the paper is not very generic, which cannot support many speech processing techniques. Future study would reveal a more general framework along with more optimization using the student-teacher model.

© 2020, www.IJARIIT.com All Rights Reserved

4. REFERENCES

[1] Yan-Hui Tu, Jun Du, Chin-Hui Lee, "Speech enhancement based on teacher-student deep learning using improved speech presentation probability for noise-robust speech recognition", 2019 IEEE/ACM transaction on audio, speech and language processing.

[2] D Wang, Y Chen, Y Shi, Z Tang, T F Zheng, "Deep factorization for speech signal", 2018 IEEE /ICASSP, Calgary, pp:5094-5098

[3] T Athanaselis, S. Bakamidis, G. Giannopoulos, I. Dologou, E. Fotinea, "Robust speech recognition in the presence of noise using medical data", 2008 IEEE International workshop on image systems and techniques, Crete, pp:349-352

[4] R Lotfidereshgi, P Gournay, "Biologically inspired speech emotion recognition", 2017 IEEE/ICASSP, New Orleans, LA, pp:5135-5139

[5] A Stolcke, E Shriberg, "Markovian combination of language and prosodic models for better speech understanding and recognition", IEEE Workshop on automatic speech recognition and understanding, 2001, ASRU, Madonna di Campiglio, pp:177-185

[6] Leslie N, Smith, "Cyclical learning rate for training neural networks", 2017, arXiv:1506.01186v6

[7] Dehak, Pierre Du-mouchel, Prierre Ouellet, "Front end factor analysis for speaker verification", 2011 IEEE Transaction on audio, speech and language processing, vol.19, no.4, pp.788-798

[8] Dong Wang, Thomas Fang Zheng, "Transfer learning for the speech and language processing", 2015 IEEE/APSIPA, pp.1225-1237

[9] Dong YU, "Neural network based multi factor aware joint training for robust speech recogintion", 2016 IEEE/ACM transaction on audio, speech and language processing, vol.24, no.12, pp.2231-2240

© 2020, www.IJARIIT.com All Rights Reserved