Speech Recognition using Artificial Neural Network

Arpita Gupta and Akshay Joshi

***Abstract*—This paper propose two approaches for speech recognition via supervised and unsupervised learning. Speech signals are non-stationary signals. Treating speech in computing domain falls under sequential learning task i.e. if we want to make a sense of current statement, we may need to go through the context in which it was spoken. Recurrent Neural Networks (RNN) is been used in speech recognition problems because of its powerful sequence modeling capacity. In this paper we have proposed Bi-directional Recurrent Neural Network with Long Short Term Memory model (LSTM), so that speech signal reconstruction can be done in a proper way without performance loss. For unsupervised learning, model is designed on the basis of Restricted Boltzmann Machine (RBM) which generates a reconstruction based output and helps in conversion of voice into text, each letter by letter.**

***Index Terms*—LSTM, MFCC features, Recurrent Neural Network, Restricted Boltzmann Machine.**

I. Introduction

SPEECH signal is composed of complex time dependent signals which have complex correlations with each other at a range of different timescales. Speech recognition includes conversion of spoken words into text. Speech recognition allows conversion of speech into text, making it easier to create and use the speech information. Speech signal is easier to generate but sometimes it is fast and intuitive and other times it is slow and not predictable So it is really hard to index a particular a speech. On the other hand, it is easier to store text so speech to text conversion becomes an important application.

As Artificial Neural Networks mimic the human brain, they are widely used for feature extraction in image processing, natural language processing and other similar tasks. In neural networks, the more the data is available for training the more the accuracy is obtained. So for speech recognition it becomes beneficial as a lot of datasets are available that contains a large amount of data to train the network. As amount of data increases, number of neurons are increased to maintain the network accuracy. As the number of neurons are increased, network complexity and the amount of time taken to train the network increases significantly.

Arpita Gupta, Dept. of Electrical and Electronics Engg., Birla Institute of Technology and Science, Pilani (333031), Rajasthan (Email: h2016102@pilani.bits-pilani.ac.in).

Akshay Joshi, Dept. of Electrical and Electronics Engg., Birla Institute of Technology and Science, Pilani (333031), Rajasthan (Email: akshu.jai@gmail.com).

Still they have been proven as the most efficient algorithms to train a network that does similar task as of a human.

Recurrent Neural Networks (RNNs) contain cyclic connections that makes them more powerful tool to model sequential data of speech than Feed Forward Neural networks. RNNs contain sequence modeling capacity, that is why they are preferred over other models for speech recognition or speech synthesis.

The rest of the article is organized as follows. Section II presents a brief literature survey of work performed in the field of speech recognition using Bidirectional RNN and RBMs. It addresses as how RNN is better than existing neural network architectures, how they face issues over long durations and how LSTMs eliminate those issues. Section III presents implementation of Bidirectional RNN and RBM mathematically, and the internal working of these models. Section IV and Section V describes the details of the experiment performed, categorized widely into data preprocessing, training of the network and finally testing the network. Section VI briefly discusses the results obtained. Section VII concludes the work and gives a glimpse of future scope in Section VIII.

II. Literature Survey

Previously, many people have worked to improve accuracy for speech recognition and tried to find out a robust methodology for this task. Artificial Neural Network can be chosen as an optimal solution for speech recognition, as they can be trained for any amount of data. Also when a task is computationally intensive, neural networks act as the best solution as the neurons are acting in parallel, i.e. in a single layer, the computations they perform are independent of other neurons, until there is some feedback mechanism introduced. In some research papers[1][2], people have used recurrent neural networks along with the convolutional neural networks. Mainly to analyze the audio signals 2D representation i.e. time-frequency analysis is performed. In above paper[1] speech signal is transformed using Short Time Fourier Transform (STFT) after some pre-processing. Then it's time-frequency representation is analyzed through convolutional neural network and Long Short-Term Memory (LSTM) model. Using convolutional neural network high level features of sound files such as MFCC coefficients are extracted and then using recurrent neural network, network is trained efficiently.

The conventional neural networks do not give much accuracy because all neurons are connected to each other in a local network and depend on the present state only. So some researchers used deep convolutional neural network[3] and they paid more attention for choosing a configuration for size of filters, pooling and input feature maps. [3] used very deep convolutional network as in-depth analysis of network

978-1-5386-3521-6/18/$31.00 ©2018 IEEE

explores its key characteristics as model scale, fast convergence speed and robustness towards noise. In case of speech recognition, it's beneficial to use past inputs also, as all speech signals are time dependent. So some researchers have suggested to use recurrent neural networks as an alternative model [4]. RNN contains cyclic models i.e. layers are connected to each other so it is easier to design sequential model. [4] has focused on end-to-end training of neural network, in which recurrent neural network is trained to map directly from acoustic to phonetic sequences of the given signal. This approach removes the need for a predefined alignment of signals to create target for training process. But conventional RNNs also have problem of vanishing gradient and exploding gradient [5]. To address all these problems, [6] proposed an architecture which uses LSTM (long short term memory model) along with recurrent neural network. When it comes to sequence labeling, LSTM model perform better task as it keeps track of past and future inputs. Recently, neural network with LSTM have been introduced on phoneme recognition task [7], robust speech recognition task [8], and large vocabulary speech recognition task [9], [10]. Also, speech is a dynamic process because at every timestamp, value of input changes. So some researchers have proposed Hidden Markov Model (HMM), but were not able to achieve as much accuracy level as is achieved by recurrent neural network [11].

III. METHODOLOGY

After observing research work done in this field, we have proposed two architectures, one consisting of supervised learning and another consisting of unsupervised learning.

Supervised learning is a process in which along with the input data, target output is also known. So the predicted output is mapped with the target output and is compared to the target output. As per the obtained error rate, weights are updated accordingly. For speech recognition we have used "Bi-directional Recurrent Neural Network along with the LSTM model" as it keeps track of past and future inputs that helps in determining the sequence labeling.

In case of unsupervised learning only input data is given. Network is trained by reconstructing the output according the input data pattern. Unsupervised learning is used basically for feature extraction and pattern matching. For speech recognition using unsupervised learning, we have used Restricted Boltzmann Machine to train the network, as with the help of RBM we can create samples that can be used for pattern recognition.

A. Bidirectional Recurrent Neural Network

It includes duplicating the first recurrent layer in neural network that contains two layers side by side, then provide input sequence to the first layer and reversed copy of input sequence to the second layer. The use of providing the input sequence in bi-directional way is justified as context of whole utterance is used to interpret what is being spoken, rather than using a linear interpretation.

In this paper we have used cross entropy function to update the weights. Gradient descent algorithm has the problem that many a times it converges at local minima so does not give optimal results. By the use of cross entropy function, this problem can be eliminated.

1. Cross entropy cost function

Cross entropy between two events is defined as the probability measure by averaging number of bits drawn from the set. It has faster convergence rate due to less derivative terms. Also, it gives good performance measurement as compared to back propagation algorithm.

2. Long short term memory model

LSTM model allows the network to keep the data in a memory cell. A memory cell has four gates according to which it is decided whether an input is to be remembered or not. It has gates to control whether a memory cell is to be overwritten with input, is to be forgotten or fed to the output gates. Cells may have the values between 0 and 1 which is decided by sigmoid activation function. As LSTM model has long term dependencies over layers, it also helps in solving vanishing gradient problem.

Fig.1 shows the structure of a basic memory cell. The original architecture contains two main gates: input gate and output gate. Later, forget gate is also added that prevents neural network from processing continuous input signals. It scales internal state of the cell by adopting and resetting cell's memory.

Mathematical equations of LSTM cell:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$
$$a_t = \tau(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$c_t = f_t c_{t-1} + i_t a_t$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$
$$h_t = o_t \theta(c_t)$$

Where, σ is taken as logistic sigmoid function, and i, a, f, o and c are respectively the input gate, cell input activation, forget gate, output gate, and cell state vectors, all of which are of same size as the hidden vector h. W_{ci}, W_{cf}, W_{co} are diagonal weight matrices respectively. τ and θ are the input of cell and cell output non-linear activation functions.

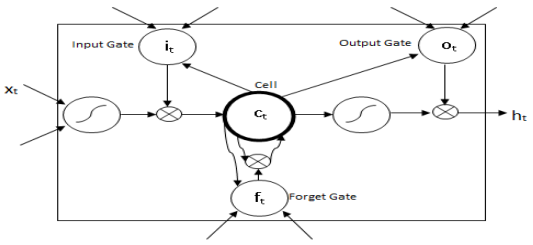

Fig. 1. Memory cell [12].

B. Restricted Boltzmann Machine

Restricted Boltzmann machine is an model that is useful for reduction of dimensionality, classification of images and feature learning. Restricted Boltzmann machine is the most simplified version of the Boltzmann machine which is based on energy probabilities. Restricted Boltzmann Machines are bidirectional networks of stochastic learning which consists of visible and hidden units.

Only visible neurons and hidden neurons are connected with each other. The advantage of this kind of neural network is hidden units are not correlated with each other and they are independently connected to visible layer.

Restricted Boltzmann machine consists of following layers:

a. First layer: First layer of visible units that accepts the input data
b. Second Layer: Second layer consist of hidden units (the latent factors we try to learn)
c. A bias unit for visible units and hidden units, a way to adjust weights properly.

In the first reconstruction phase, the activations of first hidden layer act as input in the backward pass. After that these activations are multiplied by the weights given. The sum of these products obtained is added to the visible input layer bias, and finally output of these operations obtained as a reconstruction. Fig. 2 demonstrates how RBM layers are arranged/interconnected.

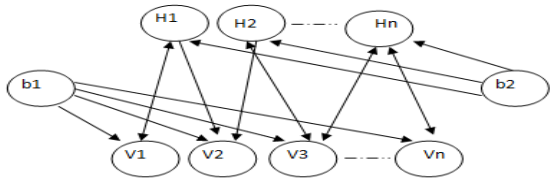

Fig. 2. RBM layer [13].

IV. Procedure

A. Bi-directional Recurrent Neural Network

1. Data Preprocessing

For getting input sound data, we have taken "LibriSpeech" dataset which contains raw sound files in .flac format. First of all conversion of .flac format to .wav format is to be done. So that using wav files feature extraction can be done. These features are feed into recurrent neural network.

For conversion of speech into text, we need to convert audio data into a form that can be fed into neural network. For this, we have calculated Mel Frequency Cepstral coefficients (MFCC) features. MFCC features are sequence of acoustic feature vectors, each vector representing information in a small time window of signal.

MFCC features are obtained as a result of inverse Fourier transform of logarithm of estimated spectrum of signal. MFCC feature extraction includes analog to digital conversion, pre emphasis, windowing of the signal, taking discrete Fourier transform, Mel filter and log, calculation of Cepstrum and finally calculation of energy using cepstral coefficients.

Text encoding is done by assigning numbers to alphabetical letters from 1 to 26. White space and apostrophe are assigned number 27 and 28 respectively for encoding.

2. Training of the network

MFCC features and text encoded data is sent as input to neural network to create forward and backward LSTM cell. After that output of these cells are fed into bi-directional recurrent neural network.

To calculate the cost function, cross entropy function is applied. Also, as number of inputs are very high, so dropout is used. Dropout method selects some neurons at a time on the basis of probability and sends it to the network. Dropout reduces the probability of overfitting of data and increases efficiency of the network. Weights are updated according to the calculated cost function.

3. Testing of the network

After training, some sound files, distinct from the training set were taken to test the network and percentage accuracy is measured.

B. Restricted Boltzmann learning Architecture:

1. Data preprocessing

Data preprocessing follows the same methodology as followed for bi-directional recurrent neural network except that here there was no text encoding.

2. Training of the network

RBM is based on the probability distribution so first of all, using visible units and sigmoid activation function, probability of hidden layer is calculated. Using these probabilities, v_prob and h_prob_1 is calculated which are known as visible neuron's probability and next hidden state's probability respectively. As there is no target output is available, so the error is calculated using reconstruction strategy. Weights are updated using Gibbs sampling rate, which is defined according to the error calculated by predicted output and given input.

3. Testing of network

To test the data, some test input is fed to the network and final accuracy is calculated.

V. Experiment

For proper training of the neural network, a large amount of sound data is required. Thus, we've selected 678 sound files for demonstration purpose, 543 (around 80%) of which are used for training the network and 135 (around 20%) are used for testing the network. The .flac files are first converted into .wav files through a python script. Then, we extract MFCC features from sound data. Since sound files are of different durations, we perform zero padding to maintain uniformity in length of MFCC features.

1. Desktop specifications:

A mainframe PC with following specifications:

a. Intel Xeon quad core processor.
b. Gigabit Ethernet connection.
c. 10GB Swap memory.

2. Tools used:

Table I below enlist the tools used for implementing Bi-RNN and RBM models.

TABLE I
TOOLS USED FOR EXPERIMENTAL SETUP

S. No.	Software	Purpose
1.	Ubuntu 16.04 LTS	Operating system
2.	Google Tensorflow 1.4.0	Machine learning API
3.	Python 2.7.14	Implementing Tensorflow API
4.	LibriSpeech (100 hrs voice) Speech database	Sound recordings in .flac format with their text equivalents

VI. OBSERVATIONS AND RESULTS

We've measured the accuracy for 25 epochs. Table II gives the list of parameters used for training both the networks.

TABLE II
PARAMETERS CHOSEN TO TRAIN BOTH MODELS

Parameter	Value
Learning Rate	0.0001
Number of hidden layers	250
Input size	80000 X 1
Training data size	543
Epoch	25

Fig. 3 and Fig. 4 gives the epochs vs percentage accuracy curve for Bi-RNN and RBM.

1. Supervised learning

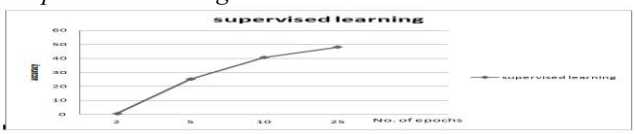

Fig. 3. Number of epochs vs accuracy in Bi-RNN model

2. Unsupervised Learning

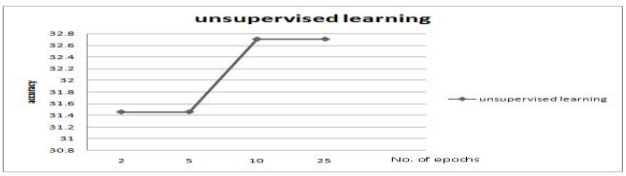

Fig. 4. Number of epochs vs accuracy in RBM model

VII. CONCLUSION

a. For supervised learning, as can be seen from above graph, there is a significant increase at initial level. Then, after epoch 8, the curve is almost flattened, indicating that rate of growth of accuracy is now decreasing.

b. For RBM, only slight increase in percentage accuracy was seen. It is because the primary disadvantage of RBM is that they are little bit tricky to train, as Contrastive Divergence algorithm requires sampling from a Markov chain model, and requires a bit of care to get things right.

VIII. FUTURE WORK

The current study focused on studying the impact on accuracy with increasing epochs. The parameters such as number of hidden layers, learning rate were kept fixed. Our future study will be to study the variations in accuracy rate by varying number of hidden layers, learning rate and other parameters so that system can be designed more robust and utilized well for real time applications.

REFERENCES

[1] Speech emotion recognition using convolutional and Recurrent Neural Networks,Wootaek Lim; Daeyoung Jang; Taejin Lee 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

[2] Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks Ying Zhang, Mohammad Pezeshki, Philemon Brakel, Saizheng Zhang, C ´ esar Laurent ´ Yoshua Bengio1 , Aaron Courville2

[3] Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition Yanmin Qian, Member, IEEE, Mengxiao Bi, Student Member, IEEE, Tian Tan, Student Member, IEEE, and Kai Yu, Senior Member, IEEE

[4] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in ICASSP, 2013, pp. 6645–6649.

[5] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Networks, vol. 5, pp. 157–166, 1994.

[6] S. Hochreiter and J. Schimidhuber, "Long short-term memory," Neural Computation, vol. 9, pp. 1735–1780, 1997.

[7] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in ICASSP, 2013, pp. 6645–6649.

[8] J. Geiger, X. Zhang, F. Weninger, and et al., "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling," in Interspeech, 2014, pp. 631–635.

[9] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in ASRU, 2013, pp. 273– 278.

[10] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Interspeech, 2014, pp. 338–342.

[11] H.A. Bourlard and N. Morgan, Connnectionist Speech Recognition: A hybrid Approach, Kluwer Academic Publishers, 1994.

[12] Generating Sequences With Recurrent Neural Networks Alex Graves Department of Computer Science University of Toronto graves@cs.toronto.edu,arXiv:1308.0850v5 [cs.NE] 5 Jun 2014.

[13] Time Series Prediction using Restricted Boltzmann Machines and Backpropagation Rafael Hraskoa , Andre G. C. Pacheco, Renato A. Krohlinga