Optimization of Deep Neural Network for Automatic Speech Recognition

Aqbal Waris
Department of Computer Engineering
National Institute of Technology, Kurukshetra
Kurukshetra, India
aqbalwaris@gmail.com

R.K Aggarwal
Department of Computer Engineering
National Institute of Technology, Kurukshetra
Kurukshetra, India
rka15969@gmail.com

***Abstract*—Neural network has achieved improvements in various fields of image processing, natural language processing, object recognition, and acoustic signal classification in automatic speech recognition system (ASR). ASR is the task of mapping the speech signals into the corresponding text without the involvement of human. Recently, paradigm has been shifted from GMM-HMM to Deep Neural Network for speech recognition process. Performance of ASR system depends on how accurately it recognizes the acoustic signals. The recognition rate is directly related to training process of the Deep Neural Network (DNN) i.e. how accurately weights are adjusted in matrix. It short, it can be said that more fine training more accurate results. Therefore, there is a need to propose such technique that optimizes the weight matrix of neural network. In this paper, Meta-heuristic algorithm pigeon inspired optimization (PIO) technique is proposed to optimize weight matrix of DNN model. This technique optimizes the weight matrix using heuristic available. By this way, training time of DNN reduces and recognition rate of system also increases. The result of optimization of weight matrix is evaluated on TIMIT database for phoneme recognition.**

Keywords—Convolutional Neural Networks, Deep NN, Hidden Markov Model, Pigeon Inspired Optimization Speech Recognition.

I. INTRODUCTION

Speech is communication channel through which people interact with each other to share their information. Automatic Speech Recognition (ASR) is a computer-driven program which converts human audio signal recorded by the microphone into the meaningful textual format. Transcription of the speech signal into its corresponding words with high accuracy is a difficult task due to speaker accent, gender, age, and unwanted noise [1]. ASR system works in two stages. In the first stage, features are extracted from raw speech signals [2] which are representations of short window power spectrum of frequency commonly derived using Fourier transform of signal then take logs powers for each frequency. Usually, features are extracted using mel frequency cepstral coefficients (MFCCs) [3] or perceptual linear prediction (PLPs) [4]. These extracted features are related to prior knowledge of acoustic speech production. These features are modeled either using the Gaussian Mixture Models (GMMs) [5] or Artificial Neural Networks (ANNs) [6] to evaluate state emission probability. In the second stage, the most likelihood of phonemes is calculated using either a statistical model or conditional models. For decoding purpose, a Hidden Markov (HMM) [7] is used to estimate the most likely sequence of utterance or phonemes. This sequence is mapped with most likely stored words. The recent improvement in training and learning paradigm makes the system able to recognize the sequence of words with more accuracy and more robustness for large vocabulary recognition. Few years ago, GMM model was generally used but ANN has been replaced GMM model to overcome the consideration of data structure. GMM models are inefficient to classify the data which lie on or near to classification boundary of data surface line. ANN uses supervised learning mechanism for training from data and produces a discriminative function to map untrained data. Backpropagation technique like gradient descent algorithm is used to optimize the neuron's weight by calculating gradient or loss function. From few years, researchers of Microsoft, Google and IBM have been achieved great achievement in training result using Deep Neural Networks (DNNs) [8] that have many hidden layers and a large output layer. Large output layer benefits to accommodate many HMM states which offer high discriminative power for phonemes recognition. The other benefit of deep architecture is that it enables the ASR system to overcome the variation present in acoustic speech signal which is given as an input to the first input layer. Many hidden layers and many neurons per layer of DNN make them more capable to create a complex and nonlinear relationship between acoustic inputs to an output. It has the capacity to handle large vocabulary dataset for training purpose and consequently reduces overfitting. Convolutional Neural Network (CNN) [9] is also a deep learning architecture, which processes the raw data in an end to end manner, i.e. many steps are learned simultaneously. However, DNNs are not much capable to overcome translational variance present in acoustic speech signal due to vocal track length. CNN's are sufficient to handle translational invariance by using weight sharing across time and frequency. CNN's have been shown better results in fields of computer vision and image classification over DNN architecture. CNN architecture consists of mainly three stages like convolution layers, pooling layers, and activation function. Here convolution and pooling work alternative in feature learning and filter stage. Researchers have been using CNN in speech recognition which makes the system more capable to train and learn features from unprocessed acoustic speech signal as an input in place of cepstral features. CNN uses the locality in convolutional layer to extract a noiseless features from a purely voice spectrum band of speech signals. Although DNN acoustic models are still in leading role due to its ease of use and easy architecture. However, limited performance of DNN is still issue. The architecture modification is not so easy so the performance can be risen only through other methods like training. To optimize the training process meta-heuristic algorithm pigeon inspired optimization (PIO) technique is proposed, which update its weight matrix accordingly. By this method, the PER reduces to 16.4% on TIMIT dataset and a relative gain of 0.6% is achieved on convolutional deep maxout networks [10].

The rest of the paper is organized as follows. Section II provides related work in ASR system. In section III, Pigeon inspired optimization algorithm is proposed. Which is applied in NN-HHM to train the neural network and Section IV introduce experimental setup, and the result of optimization, finaly conclusion is given in sections V.

II. RELATED WORK

Researchers began work on speech recognition in year 1950s with the help of a digital computer in which they used analog to digital converters and frequency spectrogram for feature extraction from speech sound. HMM is a statistical model, described by Leonard E. Baum [11] in the 1960s. HMM is defined as a well established approach recognition application. Its parameters contain the characteristics of self learning from its training data. It was first used in automatic speech recognition in mid-70s to enhance the efficiency of speech recognition system. Various paradigms like neural network techniques, discriminative and connectionist approaches with HMM are also proposed. On the other hand, various generative models are explored such as CDHMM that uses Baum-Welch algorithm A Gaussian mixture model was used in feature classification in 1995s by Renolds and Rose. This is a probabilistic model successfully used in ASR system based on expectation maximization algorithm. First hybrid model of NN-HMM was successfully used by Joe Tebelskis in the year 1995 [12]. In hybrid model, neural networks were used for acoustic modeling and HMM for decoding the most probable sequence of phonemes. Neural network that have a few hidden layers are used for classifying phonemes from cepstral features. Recently, deep neural networks that have many hidden layers offer better result in acoustic modeling in ASR system. Geoffrey Hinton et al. [13] successfully implemented DNN architecture for acoustic modeling in speech recognition in year 2012s. O. Abdel Hamid et. al. [14] gave an idea of hybrid NN-HMM in 2014 in which firstly, CNNs are applied for spatial features and acoustic modeling; and HMM perform the phonemes decoding. CNN can be regarded as an upgraded version of a neural network which gives 6-10 % error reduction over DNN architecture.

III. PROPOSED WORK

In this section, we optimize the weight matrix of hybrid NN-HMM model for speech recognition where training of NN is performed with the help of pigeon inspired optimization (PIO) technique. The original motive of this optimization approach is to achieve better accuracy and performance by optimizing the training processes of NN-HMM model for better accuracy and performance. A meta-heuristic optimization technique is used to train the model with the help pigeon inspired optimization [15] to minimize mean square error. Hybrid NN-HMM is baseline model just like GMM-HMM model. For each given inputs, the acoustic feature vector is estimated and given to NN that calculates posteriori probabilities means state transition probability from one state to another state and visible symbol emission probability. The NN training objective is a selection of HMM state.

A. NN-HMM

An HMM [16] is a stochastic process with related to Markov chain process that cannot observe directly but can be observed with the help of other stochastic process which is produced by Markov process. HMM is used to model a word in a vocabulary where each hidden state represents a phoneme and calculates the most probable sequence of phonemes. After the completion of training, HMM is used for decoding the sequence of words or pattern matching. HMM has three major works in speech recognition.

1) Evaluation problem: HMM model calculates the probability of a sequence of visible state v^{T} given θ model.

$$P\left(\mathrm{v}^{\bar{\mathrm{T}}}/_{\theta}\right) = \sum_{r=1}^{r_{max}} P(\mathrm{v}^{\bar{\mathrm{T}}} / \omega_r^T)\, P(\omega_r^T) \tag{1}$$

$$P(\omega_r^T) = \prod_{t=1}^{T} P\,(\omega(t)/\omega(t-1)) \tag{2}$$

$$P(v^{\bar{T}}\Big/_{\omega_r^T}) = \prod_{t=1}^{T} P\,(v(t)/\omega(t)) \tag{3}$$

where T indicates a number of visible states r_{max} is a number of the possible sequence of ω^t.

2) Decoding problem: A θ model of HMM calculate the most likely sequence or most probable sequence of the hidden state over which the machine has transition during generating a sequence of the visible state $\mathrm{v}^{\mathrm{T.}}$

3) Learning problem: HMM model is trained by supervised learning. HMM model trained the state transition probability (a_{ii}) and visible symbol emission probability b_{ik} using the backward algorithm. Backward algorithm calculates the probability that the machine will be in state (ω_i) at time instant t and will generate the remaining part of set visible symbol (v^T).

ANN [17] is trained with the help of back propagating error derivative technique where difference between actual output and expected output derivative is fed to input node. There exist more than one hidden layer unit in between inputs and outputs layer. Every hidden unit, j, uses sigmoid function to map its total inputs from previous layer, x_j, to scalar state, y_j which is sends to next upper layer unit.

$$y_j = logistic(x_j)\,\frac{1}{1+e^{-xj}};\quad x_j = b_j + \textstyle\sum_i y_i w_{i_j} \tag{4}$$

where b_j is bias of unit j, I is an index of unit in previous layer, w_{i_j} is weight between below unit i to upper unit j.

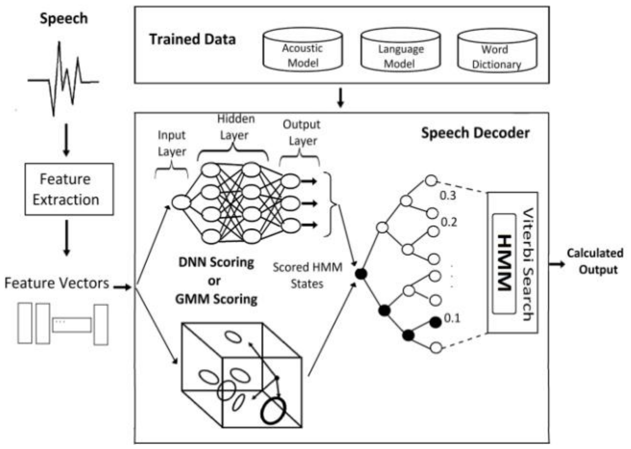

Fig 1. ANN-HMM Model

Algorithm:

```
1. Let $\partial_0$ be (max) number of calls a pigeon p will make to group and search.
2. Let n be random pigeon population.
3. consider a population $x_i$ with $iinrange : 1\ ton$
4. Select a flock of pigeon population having better searching time.
5. Let location of food be $A$
6. While food is located or $f(x_i) > \partial_0$do
    a. Evaluate vision radius of group$R_g$ and vision radius of pigeon$R_i$.
       Where
       $R = \sqrt{(lengthoflineofsight)^2 - (height)^2}$
    b. If food is not in $R_g$for a predefined time bound
           Redo the procedure after changing route
    c. If food is in $R_g$ for a predefined time bound
           Do until pigeon reaches food
           i. Evaluate the distance D for all pigeons in flock and food.
              $D_{gA} = \sqrt{(x_g - x_A)^2 + (y_g - y_A)^2}$
           ii. $min_D =$ minimum of $D_{gA}$
           iii. Return optimized path = $min_D$
           iv. $f(x_i) = f(x_i) + 1$ for pigeon with minimum distance from food
7. For optimized solution
    a. Evaluate $f(x)$ for pigeon p in flock
    b. Maximum value of $f(x)$ is returned and it would be leader in flock
```

Pigeon optimization algorithm inspired by bio-inspired optimization [18] based on swarm behavior like fireflies, ant and bee which is implemented for optimization problems. The leader of pigeon flock initiates conversation and signal to other pigeon in flock who acknowledge back by emulating the behavior of calling pigeon and manage side by side structure emerge in a flock of definite shape. The leader of pigeon of flock is chosen on the basis of the number of times calls to other pigeon in flock. A fitness function $f(x)$ attach to every pigeon that count how many times a particular pigeon called to other pigeon in given population. POA has capability of problem-solving can be used in various field of optimization like a shortest path in traveling salesman problems.

IV. RESULT AND DISCUSSIONS

A. *Experimental Setup*

Human acoustic speech recognition observations are performed on the TIMIT corpus set to evaluate performance of the proposed optimization technique for hybrid NN-HMM models. It consists of 6300 utterances from the 630 speakers. We used target class labels (61 phones * 3 state/phone). For decoding, a phone tri-gram model is used. After decoding, there 61 phone classes are mapped into 39 useful classes. MFCC feature extraction technique is used for extracting the features from raw speech signals. The sliding window size is taken 25-ms with a fixed shift of 10-ms. 13 coefficients + their first and second coefficient + energy i.e. 40 observations are supplied as input feature vector. The proposed system evaluated on MATLAB version 2017a for developing feature extractor module of ASR system. The acoustic module and decoding module have been developed using HTK 3.5 β-2 version toolkit. For neural network training, we have used PIO is to optimize the weights of NN-HMM model and objective of PIO is to minimize mean square error. The input layer on neural network includes a context window of 15 frames. The input of NN is divided into 40 bands. Each band includes one of the 40 filter-bank Mel cepstral coefficients along the 15 frames context window including their first delta and second delta derivatives. We performed an experiment on a personal computer with Intel i7-8core processor; 8GB RAM using Windows 10 operating system. We used NNSTART tool for neural network configuration on MATLAB environment. PIO will run for 1000 number of iterations with 100 sizes of the population.

B. *Result*

The phone recognition observations are shown in table 1, which clearly indicates a significant and persistent reduction of the phone error rate as compared to Gradient Descent training. PIO able is to perform well and achieve phone error rate reduction of 16.4%. Result clearly indicates that training using PIO is better compared to back-propagation (BP) training method for NN-HMM model. It reduces error rate up to 10% on the same TIMIT core test database. As the number of iterations in PIO increases the corresponding the phone error rate also decreases. Figure 2 represents comparison over iteration of PIO and error rate of NN-HMM model, which indicates that over large number of iterations error rate reduces vastly.

Table 1. Result comparison

Technique	**Phoneme Error Rate**	
	Training Set	***Testing Set***
Gradient Descent	16.5	17.1
Back Propagation	16.8	17.3
PIO	16.4	16

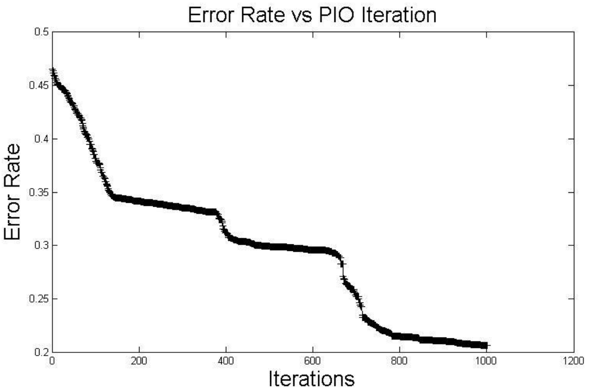

Fig. 2. Iteration vs. Error Rate comparison

V. CONCLUSION

The primary aim of designing speech recognition systems is that it's should be high. There are various methods like improved feature extraction technique, better training algorithms, hybrid acoustic model etc. which have been used for it. Training by weight optimization technique is also growing area in the field of deep neural networks. Therefore, an meta heuristic optimization tecnique i.e. PIO is proposed which optimizes weight matrix of NN-HMM model for ASR system. We performed the experiment on TIMIT dataset using HTK tool kit. It is a first attempt when PIO is used for the training of speech recognition system. It reduces phone error rate by up to 0.6% with respect to back propagation method.

REFERENCES

[1] Pasricha, Vishal, and Rajesh Aggarwal. "Hybrid architecture for robust speech recognition system." In *Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on*, pp. 1-7. IEEE, 2016.

[2] Passricha Vishal and Rajesh Kumar Aggarwal. . "Feature Extraction technique for Hindi speech recognition system,"International Journal of computing and application 13, no. 2 (2018) : 221-229

[3] Ibrahim, Noor Jamaliah, Zaidi Razak, Mohd Yakub, Zulkifli Mohd Yusoff, Mohd Yamani Idna Idris, and Emran Mohd Tamil. "Quranic verse recitation feature extraction using Mel-frequency cepstral coefficients (MFCC)." In Proc. the 4th IEEE Int. Colloquium on Signal Processing and its Application (CSPA). 2008

[4] Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." the Journal of the Acoustical Society of America 87, no. 4 (1990): 1738-1752.

[5] Burget, Lukas, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek et al. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models." In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp. 4334-4337. IEEE, 2010.

[6] Lippmann, Richard. "An introduction to computing with neural nets." IEEE Assp magazine 4, no. 2 (1987): 4-22.

[7] Huang, Xuedong D., Yasuo Ariki, and Mervyn A. Jack. "Hidden Markov models for speech recognition." (1990): 60-80.

[8] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine 29, no. 6 (2012): 82-97.

[9] Palaz, Dimitri, Mathew Magimai Doss, and Ronan Collobert. "Convolutional neural networks-based continuous speech recognition using raw speech signal." In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 4295-4299. IEEE, 2015.

[10] Tóth, László. "Phone recognition with hierarchical convolutional deep maxout networks." *EURASIP Journal on Audio, Speech, and Music Processing* 2015, no. 1 (2015): 25.

[11] Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." The annals of mathematical statistics 41, no. 1 (1970): 164-171.

[12] Tebelskis, Joe. "Speech recognition using neural networks." PhD diss., Carnegie Mellon University, 1995.

[13] Dahl, George E., Tara N. Sainath, and Geoffrey E. Hinton. "Improving deep neural networks for LVCSR using rectified linear units and dropout." In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8609-8613. IEEE, 2013.

[14] Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition." In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4277-4280. IEEE, 2012.

[15] Goel, Shruti. "Pigeon optimization algorithm: A novel approach for solving optimization problems." In *Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on*, pp. 1-5. IEEE, 2014.

[16] Rabiner, Lawrence R., and Biing-Hwang Juang. *Fundamentals of speech recognition*. Vol. 14. Englewood Cliffs: PTR Prentice Hall, 1993.

[17] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645-6649. IEEE, 2013.

[18] Duan, Haibin, and Peixin Qiao. "Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning." *International Journal of Intelligent Computing and Cybernetics* 7, no. 1 (2014): 24-37.