Low-Resource Constraints for Speech Recognition using HDNN

S. Kousar Bhanu[#1]
[1] M.tech Scholar, Department of CSE, JNTUA College of Engineering, Anantapuramu, Andhra Pradesh, India

Abstract

In Speech Recognition acoustic model is a document to represent the communication between audio signals that make up speech. In many approaches Neural Network develops an attractive acoustic modelling like Speaker Adaptation whereby adapted to acoustic features. The study determines that in Speech Recognition the Highway Deep Neural Network (HDNN's) contains the two gate units that secured over all the hidden layers to supervise the way of the highway networks. These gate units are shared beyond all the hidden layers to reduce the size of model parameters, all the model parameters are updated in sequence training to improve the results. In this paper, HDNN is used for implementation of Gate functions using Stacked Autoencoder (SAE), a layer-wise approach to train Deep Neural Network based on Speech Repository analysis. These Encoders decides a Machine Learning Model to locate a Low level Dimensional portrayal of model parameters has taken from Direct Voice Input (DVI). DVI intended to voice command-and-control to read the parameters from the speech utterances for each speaker.

Keywords — *Stacked Auto-Encoders, Automatic Speech Recognition, Stacked Auto-Encoders, Speech Recognition, HDNN, Acoustic Models*

I. INTRODUCTION

Speech Recognition accuracy has been improved approximately by Deep Neural Network (DNN) phonic model over the past several years. DNN has leaped forward naturally in Speech Recognition. Hinton et.al [1] Gaussian Mixture Model (GMM) based structure reports word error rate limiting between 10-32% ranges of tasks in comparison of these models. DNN and HMM (Hidden Markov Model) are the two Neural Networks used as a selective parameter for the GMM-based system [2]. Similar to these models, present neural network make for larger and deeper. Multilayer neural models like Recurrent Neural Networks (RNNs) alongside Convolutional Neural Network (CNN) are adaptable and viable structures have been adjusted to a scope in Natural Language, Image and Speech Processing.

A few utilizations of speech automation measure the clean speech from noisy environments, like Automatic Speech Recognition (ASR) system. Direct voice input (DVI) and Large Vocabulary Continuous Speech Recognition (LVSCR) is two extensive type of ASR. DVI intended to voice command-and-control and LVSCR frameworks are utilized for voiced-based document creation.

The Acoustic model is challenged to move on the resource-constrained platforms that have a large tendency when compare to the conventional Gaussian model. In past work, HDNN is designed for tune-up Small-Footprint Acoustic Model [3]. Usually, HDNN contains transform and carry gates, which regulate the information in the whole network. These gate units are shared beyond all the hidden layers to reduce the size of model parameters, all the model parameters are updated in sequence training to improve the results. To overcome this problem, the proposed system has an unsupervised model-based adaptation of HDNN Acoustic models with Stacked Auto-Encoders (SAE).The ASR using SAE which is stochastic machine learning process to get the results in a probabilistic manner. A Deep learning based speech recognition method that comprises a Stacked Auto-Encoders model which is used to learn generic linguistic features, it uses auto-encoder as building blocks to create a Deep network and layer-wise greedy fashion for training.

However, Small-footprint models are good in a few viewpoints, for example, less computational cost and less memory to pertinent for low-asset dialects which contains humble number of model parameter where the data is small for training. The system also focused on computational footprint particularly system parameters like CPU, Memory, Disk, and Network. Memory is the main estimating factor of the computational footprint. The Existing system performance is also based on these computational parameters. The experiment was performed on effective individual headset based on Direct Voice Input (DVI). Unsupervised Speaker adaptation Model gate functions are fine-tuned using Stacked Auto Encoder. Overall, the Small-print Highway Deep Neural Network acoustic model achieved better results based on computational parameters.

II. HIGHWAY DEEP NEURAL NETWORKS

A. Deep Neural Network

Deep learning is an area of best in class frameworks in different directions, especially computer vision and Automatic Speech Recognition (ASR). Although Convolutional Neural Networks

(CNNs) and Recurrent Neural Networks can receive high recognition accuracy with slight model parameters comparison in DNNs [6], [7], they are summing high cost for appliances on resource-constrained platforms.

A network with N hidden layers for multi-layer network is given as

$$h_1 = \boldsymbol{\sigma}(x, \theta_1) \quad (1)$$

$$h_n = \boldsymbol{\sigma}(h_{n-1}, \theta_n), \text{ for } n=2\ldots N \quad (2)$$

$$y = g(h_N, \theta_c) \quad (3)$$

$\boldsymbol{\sigma}(x, \theta_1)$, where x is an input parameter, $\boldsymbol{\sigma}$ denotes non-linear activation function, e.g., sigmoid; along with input h_{n-1} with the parameter θ_n. $g(h_N,\theta_c)$ is the output behavior within the output layer that is calculated by θ_c.

Gradient descent is an algorithm to train the network by updating the weights by cross-entropy training which reduces to over fitting [3], where the back-propagation algorithm is a common model that used to train networks for unsupervised speaker adaptation; it measures the gradient for each output node and hidden node.

B. Highway Networks

Highway Networks are inspirited by Long-Short-Term Memory recurrent (LSTM) networks and gate units to control the flow of information. HDNN [5] was designed a high deep networks with gate functions to train by building the hidden layers in the whole network

$$h_1=\boldsymbol{\sigma}(h_{l-1}, \theta_1) \circ T(h_{l-1}, W_T) + h_{l-1} \circ C(h_{l-1}, W_C) \quad (4)$$

Where, T is the *transform* gate that measures the output of all hidden layers; C is the *carry* gate, which enhance the absorption before passing to the next layer; h_l denotes the hidden activation of l^{th} layer and o shows element-wise multiplication. The output of T(.) and C(.) gates that are limited to be within [0, 1], the sigmoid function for both gates that are characterized by W_T, W_C respectively. Sequentially previous work [3], by connecting the parameters in the gate function the model parameters are saved over the entire hidden layer.

C. Stacked Auto Encoder

Auto Encoder (AE) [8] it is an unsupervised learning model proposed by Rumelhart in 1986, indicate the learning process by reforming the input. The model carries two parts: encoding and decoding, the same size for input and output layers which makes to learn the compressed data by reducing the number of hidden layers.

A Stacked Auto Encoder (SAE) is a neural system comprising of numerous layers where the yield of each layer is associate with contributions of the following of layers, by objective function the error is fitted between input and reform input data and parameters are updated to activate the hidden layer. Figure 1 shows the SAE structure.

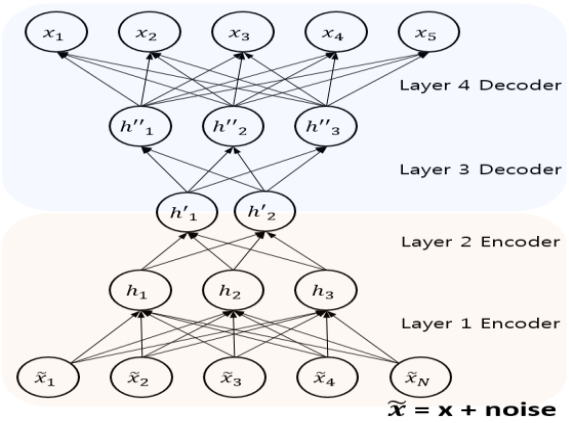

Fig.1. Structure of SAE

III. TRAINING

A. Cross-Entropy Training

In machine learning Cross-Entropy train standard neural networks for unsupervised speaker adaptation, the function is called also as cross-entropy loss function (CE) [3].

$$L(CE)(\theta) = -\Sigma j y\hat{}jt \log y jt \quad (5)$$

Where, yjt is the output for index model (3) of a neural network at the time t.

B. Stacked Auto-Encoder pre-training

Stacked Auto-Encoder neural networks will be constructed through supervised adaptation technique and stacked through the concept of sparse Auto-Encoder. In training process Batch encoder is used to process the model parameters for the same colour in fine-tuning. The nodes in a hidden layer will decrease with rising level of layers.

C. Neural Network Fine-Tuning

In Neural Networks through crest and through analysis the features have learned by the layer-wise fashion. The SAE-Fine tuning Acoustic model collected and stored only crest information data at each time slot. The model performs fine-tuning to update parameters by Stochastic Gradient Descent (SGC) to minimize the learning process of features and cost of the auto-encoders [8].

IV. EXPERIMENTAL RESULTS

A. System setup

The experiment was performed on individual headset that evaluates the effectiveness of the adaptation method. Clean word-level speech data sets with 50 utterances were used for training. The DNN

and HDNN frameworks were prepared utilizing the same training data set. The systems were prepared utilizing the Cross-entropy without pre-training, the weights of each hidden layer with a range of [-0.5, 0.5] were initialized randomly. The examined experiments on a system with Intel® Core™ i5-3320M CPU@ 2.60 GHz, 64-bit OS, 4 GB RAM, and 64-bit OS characteristics.

Table I shows the DNN and HDNN models training results for various size based on Computational Footprint. Figure 2 shows the performance evolution of models.

id	System	Utterances	Memory footprint	Duration
1	DNN	20 50	0.5M	23.2 54.5
2	HDNN	20 50	0.35M	13.9 38.2
3	HDNN +SAE	20 50	0.30M	11.8 29.9

Table I: Training Results For DNN And HDNN Models Based On Computational Footprint.

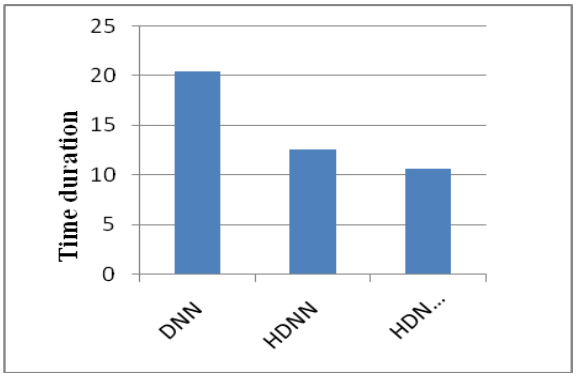

Fig.2. Performance Evolution of models.

B. Results

Stacked autoencoder by utilizing entryway capacities has control the conduct of a multi-layer neural system highlight removes with a predetermined number of model parameters. The experiment evaluates on unsupervised speaker adaptation, in which speaker-independent model used for decoding. The evolution set contains 50 utterances of the speaker, the performance observed with 10 adaptation iterations.

Figure 3 represents the convergence curves of unsupervised adaptation with the different number of iterations.

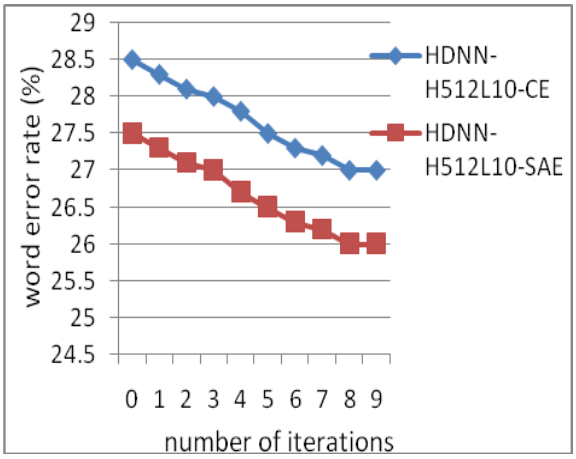

Fig. 3 convergence curve of unsupervised adaptation

VII. CONCLUSIONS AND FUTURE WORK

Deep learning has been strongly activated in speech processing, particularly in acoustic modelling for ASR. HDNN is structured and depth-gated Feed forward neural network for training acoustic model. In this paper, cross-entropy training and adaptation acoustic modelling used stacked autoencoder for fine-tuning the adaptation data. The gate functions which controls the behaviour of the entire system by using small amount of parameters. The behaviour of gate units are improved by using stacked autoencoder, a layer-wise approach to train the deep neural network based on the speaker-independent model. In training process Batch encoder is used to process the model parameter for fine-tuning. Furthermore, the investigation of computational footprint i.e., system parameters like CPU, Memory, Disk, Network, based on these computational parameters the behaviour of HDNN acoustic model has improved.

In future, stacked auto encoders used to train the deep neural network without using layer-wise scheme and also this model will evaluate in high configuration system

REFERENCES

[1] Swietojanski, Pawel, Jinyu Li, and Steve Renals. "Learning hidden unit contributions for unsupervised acoustic model adaptation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.8 (2016): 1450-1463.

[2] Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." IEEE Transactions on audio, speech, and language processing 20.1 (2012): 30-42.

[3] Lu, Liang, and Steve Renals. "Small-footprint highway deep neural networks for speech recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing25.7 (2017): 1502-1511.

[4] Lu, Liang. "Sequence training and adaptation of highway deep neural networks." Spoken Language Technology Workshop (SLT), 2016 IEEE. IEEE, 2016.

[5] Srivastava, Rupesh K., Klaus Greff, and Jürgen Schmidhuber. "Training very deep networks." Advances in neural information processing systems. 2015.
[6] Sak, Haşim, Andrew Senior, and Françoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." Fifteenth annual conference of the international speech communication association. 2014.
[7] Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." IEEE/ACM Transactions on audio, speech, and language processing 22.10 (2014): 1533-1545.
[8] Zhou, Ju, Li Ju, and Xiaolong Zhang. "A hybrid learning model based on auto-encoders." Industrial Electronics and Applications (ICIEA), 2017 12th IEEE Conference on. IEEE, 2017.
[9] Shinoda, Koichi. "Speaker adaptation techniques for automatic speech recognition." Proc. APSIPA ASC 2011 Xi'an(2011).
[10] Cao, Zihong, et al. "Auto-encoder using the bi-firing activation function." Machine Learning and Cybernetics (ICMLC), 2014 International Conference on. Vol. 1. IEEE, 2014.
[11] Lee, Jae-Neung, and Keun-Chang Kwak. "A performance comparison of auto-encoder and its variants for classification." Signals and Systems (ICSigSys), 2017 International Conference on. IEEE, 2017.
[12] Lin, Szu-Yin, et al. "A Dynamic Data-Driven Fine-Tuning Approach for Stacked Auto-Encoder Neural Network." e-Business Engineering (ICEBE), 2017 IEEE 14th International Conference on. IEEE, 2017