

Isolated Word Speech Recognition System Using Deep Neural Networks

Dhavale Dhanashri and S.B. Dhonde

Abstract Speech recognition is the process of converting speech signals into words. For acoustic modeling HMM-GMM is used for many years. For GMM, it requires assumptions near the data distribution for calculating probabilities. For removing this limitation, GMM is replaced by DNN in acoustic model. Deep neural networks are the feed forward neural networks having more than one or multiple layers of hidden units. In this work, we have presented the isolated word speech recognition system using acoustic model of HMM and DNN. We are using Deep Belief Network pre-training algorithm for initializing deep neural networks. DBN is a multilayer generative probabilistic model with large number of stochastic binary units. The features used are the mel-frequency cepstrum coefficients (MFCC). Experimental results are calculated on TI digits database. Proposed system has achieved 86.06 % accuracy on TI digits database. System accuracy can be further increased by increasing the number of hidden units.

Keywords Isolated word speech recognition • Deep neural networks • Deep belief networks • Acoustic model

1 Introduction

According to definition biometric is something, you are. For every individual there are some physical characteristics that are unique. It includes mainly fingerprint, iris, speech. A combination of these characteristics can be used to build a model which will be used to identify individuals. Speech signal carries important information. It gives information about the speech and speaker identity. Speech identification systems extract information from the speech signal independent of the speaker.

Dhavale Dhanashri (✉) · S.B. Dhonde
Department of Electronics, AISSMS-IOIT, Pune, Maharashtra, India
e-mail: dhanashrimd@gmail.com

S.B. Dhonde
e-mail: dhondesomnath@gmail.com

Since 1950 speech recognition has been in the research field for developing techniques for isolated digits or for continuous speech. Speech recognition can be considered as a process which is converting speech signals into spoken words [1]. In this paper, we are explaining about the speech recognition system for isolated words identification.

Much research has been done in the speech recognition field for isolated digits as well as for continuous speech. Many algorithms are invented like Dynamic Time Wrapping, Hidden Markov Model, and Gaussian Mixture Model [2]. Neural networks are used as a classifier for speech recognition system. There are many types of neural networks like feed forward neural networks and feedback neural networks and recurrent neural networks [3]. Many researchers came with the new learning algorithms for the neural networks. Later it came to know that more the deeper network, performance of speech recognition system is higher.

Deep neural networks are used in speech recognition for classification. Deep neural networks are the multilayer, feedforward neural networks. In this paper our approach introduces a system containing acoustic model of HMM and DNN used for identifying digits from zero to nine.

2 Literature Review

Much research has been done in the field of speech recognition technology related to the application of neural networks. Neural networks are similar to human brain. They consist of a processing element called neuron. Combination of artificial neural network and hidden markov model is used for speech recognition technology. Previously ANN is used to estimate the posterior probabilities of a continuous density HMMs' state given the acoustic features and then backpropagation algorithm is used to train that network [4]. But by using backpropagation algorithm it is found that it is difficult to train network having more than two layers.

Under this section, we are addressing the survey about research made in neural networks. In [5] they have proposed a spectral masking system to noise robust speech recognition with the help of deep neural networks [5]. A spectral masking system has been proposed by them in which power spectral domains are calculated using DNN. For further increasing performance, Lin Adaptation is applied to mask estimator and acoustic model of DNN. They have evaluated the system accuracy on Aurora2 and Aurora4 database. The system yielded word error rates of 4.6 and 11.8 %. Polur et al. in [6] implemented a system using artificial neural networks. They have calculated system accuracy for two words 'yes' and 'no.' They have used artificial neural networks in conjunction with cepstral analysis in isolated word recognition system.

In [7], they have introduced two generalized maxout units called as p-norm and soft-maxout. They evaluated their results on large vocabulary and calculated that p

norm generalization works better. In [8] they have introduced the use of neural networks for speaker identification. As for neural networks it is difficult to train on large number of units. In this paper, they have proposed neural network array combining the binary partitioned approach with decision trees. This approach reduces the computational cost and classification error rate. Mohamed et al. in [9] proposed an acoustic model of HMM and DNN for speech recognition system. They have evaluated the system performance on TIMIT database. A context-dependent model for phone recognition has been presented by Dahl [4]. This model consists of hybrid architecture of HMM and DNN. They have applied this model to large vocabulary tasks. A new algorithm for learning deep neural networks is given in this paper. It is the DBN pre-training algorithm which helps in reducing recognition error. They performed experiments on business search dataset. They have compared their results with conventional HMM-GMM system. They got improvement in accuracy up to 5.8 and 9.2 %. In [10] they have proposed an acoustic model of HMM and DNN for speech recognition system. They have implemented the system using auto encoders for pre-training algorithm. Their results shows that system gives better performance for auto encoders.

In [11] they have proposed the use of deep neural networks for extracting features from mel scale filter banks and have used DBN in combination with HMM. They have evaluated their results on Tagalog database. In [3] they have make use of speech enhancement to remove noise from speech signal before extracting it. Dynamic Programming algorithm is used to calculate the similarity between templates Use of deep recurrent neural networks is presented in [12]. They have introduced end to end learning method for recurrent neural networks. They have evaluated their results on TIMIT dataset with 462 speakers. They have achieved test error of 17.7 % on TIMIT phoneme recognition benchmark.

There are many approaches of using neural networks for speech recognition. Many authors have implemented the systems using neural networks. Some of them used combination of HMM and DNN for large vocabulary speech and others have used artificial neural networks for small vocabulary. In this work instead of using that approach, we are using deeper neural network to replace traditional neural network. Also we are trying to improve the earlier hybrid approach using DBN pre-training algorithm for training DNN.

3 System Overview

In this paper, we are going to introduce a system, i.e., isolated word speech recognition system using acoustic model of HMM and DNN. Related to learning algorithms of neural networks there are many sentences made in past. Many algorithms have introduced in the literature. In this approach we are using deep belief network pre-training algorithm for the deep neural networks [4]. Many

researchers proved that neural networks with pre-training achieve better response than networks without pre-training. We are using the feed forward, multilayer neural network for producing posterior probabilities of HMM states as output by using feature vectors as input [9].

It has been concluded that when using one or two hidden layers, back propagation algorithm is useful. Results are not satisfactory if the numbers of hidden layers are increased for back propagation algorithm.

In this paper, we have replaced GMM by HMM for narrating HMM states to feature vector. Here we are using multilayer, feed forward neural networks that use feature vectors as input and produces posterior probabilities of HMM states as output [9]. Previously for training neural networks backpropagation algorithm have been used. In this paper we are using pre-training algorithm to train multilayer neural networks to achieve better performance.

3.1 *Preprocessing*

The input speech signals are in the form of audible waveform, they are converted to electrical pulses. After that speech signal is preprocessed. Recognition rate can be distorted by noise and difference in amplitude of the speech signal [13]. These problems can be overcome by doing pre-processing part which consists of normalization, pre-emphasizing, windowing [14]. In order to adjust the volume of audio files to a standard level normalization is used.

In order to preprocess a signal, it should be pre-emphasized before. At the time when sound is produced, some part of speech signal is suppressed. So this suppressed part which is a high frequency part is compensated by doing pre-emphasis of speech signal.

A signal which is converted into frames is taken. It is an unstable signal. It remains stationary for very short time. In windowing method, we are applying hamming window to speech signal. It reduces the signal discontinuities mainly at the end and start of signal. Hamming window equation is given by

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \quad (1)$$

3.2 *Feature Extraction*

Mel-Frequency Cepstral Coefficients (MFCC) have been the most widely used speech features for speech recognition [13]. Feature extraction converts higher dimensional input signals into lower dimensional vectors. In the process of feature

extraction first the speech signal is converted to frequency domain by doing FFT on signal. It is applied to speech signal in order to obtain magnitude frequency response of each frame. The spectrums obtained are mapped onto the mel scale using triangular overlapping window. For mel-frequency scale it is having linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [11]. Here the frequency axis in Hz is converted into mel scale using formula

$$\text{mel frequency} = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

Finally log is taken for the output and DCT, discrete cosine transform is applied to the log mel powers. The magnitudes of the resulting spectrum are the required MFCC's. The equation for DCT is given by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2kn\pi/N}, k=0, 1, \dots, N-1 \quad (3)$$

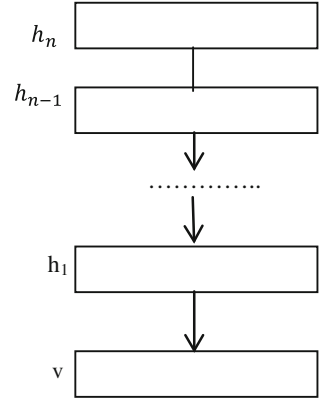
4 Acoustic Modeling

When applying deep neural networks as an acoustic model to speech recognition system in combination with HMM, it is used to produce posterior probabilities of HMM states as output [9]. Here we are using DBN pre-training algorithm for building deep neural networks. It is better to learn a generative model when there is large number of unlabeled data in combination with the training data [9]. Deep belief network is a multilayer generative model. Here the deep neural networks are trained layer by layer.

5 Deep Belief Networks

These are the multilayer generative probabilistic models with large number of stochastic binary units. In this the bottom layer is visible layer v , and others are the hidden layers, h . Figure 1 shows the graphical model of deep belief network. It shows the layer of hidden and visible layers. It shows, there is a undirected connection between top two layers and RBM is formed by this. And the bottom layers are in top-down directed manner [4]. By training deep neural networks for input feature vectors, probability distributions for HMM states can be obtained. Using some preexisting model, correct state can be obtained and weights of neural networks are adjusted in order to increase the probability for correct HMM state [11].

Fig. 1 Graphical model of deep belief network



6 Restricted Boltzmann Machines

Restricted Boltzmann machines (RBMs), these are the undirected graphical models consisting visible layer and hidden layer. These are not actually deep architectures [4]. All units are connected in a well manner. Also there is no connection between the units of same layer. RBM's are the building blocks of deep belief networks.

RBM's assign an energy value to each configuration so that dependencies between variables can be obtained with more probable configurations having a lower energy [9, 4]. Energy term associated with visible, v and hidden, h units is given by equation,

$$E(v, h) = - \sum_{i=1}^v \sum_{j=1}^H w_{ij} v_i h_j - \sum_{i=1}^v b_i v_i - \sum_{j=1}^H a_j h_j \quad (4)$$

where w_{ij} is the weight assigned between the connections of hidden and visible units, b_i and a_j are their bias terms. The probability of a given configuration is given by eq,

$$P(v, h) = \frac{e^{-E(v, h)}}{Z} \quad (5)$$

where z is the normalization factor, $Z = \sum_{v, h} e^{-E(v, h)}$

As RBM is an energy-based model, training of RBM is like obtaining probability distribution. In RBM it is done by decreasing the energy of most probable regions of the state space while boosting the energy of less probable regions. Contrastive divergence is used for learning RBM [4].

It uses the following rules for updating weights and biases,

$$\Delta W_{ij}^0 \propto \langle v_i h_j \rangle_0 - \langle v_i h_j \rangle_n \quad (6)$$

$$\Delta b_i^0 \propto \langle v_i \rangle_0 - \langle v_i \rangle_n \quad (7)$$

$$\Delta b_i^1 \propto \langle h_i^1 \rangle_0 - \langle h_i^1 \rangle_n \quad (8)$$

where $\langle \cdot \rangle_0$ is the average when visible units are imposed to the input values and hidden units are taken out, and $\langle \cdot \rangle_n$ is the average after assigning the input data to visible units.

After pre-training a stack of RBMs, the weights and biases of the hidden units can be utilized to initialize the hidden layers of a deep belief network.

Once we have trained an RBM, we can use it to represent it as data. For each input vector, v we compute a hidden unit activation probabilities, h which we can be used to train next RBM. Hence RBM weights are used to extract features from pervious layer. After completing the training of RBM, we got weights of the hidden layers which will be used to build deep neural network. After that back propagation algorithm is used for fine tuning all weights in the network.

7 Experimental Results

7.1 Dataset Description

In this paper, we have evaluated all the results on TI Digits database. It is speaker independent connected digit database. The data were collected in a quiet environment and digitized at 20 kHz. The utterances collected from the speakers are digit sequences. Eleven digits were used: “zero,” “one,” “two,” ..., “nine,” and “oh.” The data is divided into two sets, training and testing. Each contains 55 men and 57 women while 56 men and 57 women samples.

The training data is used to train the network. Each file of speaker contains two utterances of each word. The speech samples from training data are read using wavread command. Features are extracted using MFCC. We have used voicebox speech processing toolbox for our experiment (Fig. 2).

Fig. 2 Relationship between hidden units and word error rate

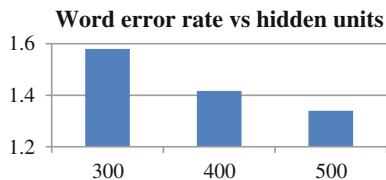
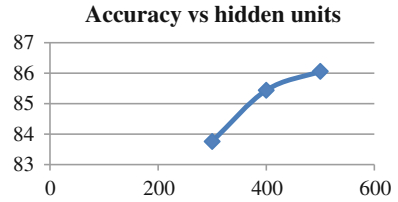


Fig. 3 Relationship between hidden units and accuracy



7.2 Results

We tested the results on dataset of 56 men and 57 women samples. In this we have calculated the system accuracy for different number of hidden layers. We have calculated it for 300, 400, 500 hidden units. Our experiment for 300 hidden units achieved 83.76 % accuracy. By increasing hidden units to 400, accuracy is increased up to 85.44 %. We have got 86.06 % accuracy for 500 hidden units. These results are shown in Fig. 3. Also the word error rate for each word is calculated. Figure 2 shows the graph of word error rate versus hidden units.

As shown in Fig. 3, accuracy increases with increase in number of hidden units. For RBM, we have used 100 epochs. In the experiment, we have calculated results for using deep neural networks only for classification. We have got results for different hidden units such as for 300 units accuracy achieved is 66.33 %.

8 Conclusion

In this paper, we have proposed an isolated word speech recognition system using acoustic model of HMM and DNN. Having large number of hidden units, deep neural networks are used as a classifier in speech recognition system. In case of GMM they are statistically incapable at modeling high-dimensional data that has any kind of componential structure. As a replacement for GMM, DNN is used in combination with HMM in acoustic modeling. In this paper, we gave a brief view about the deep neural networks and the newly invented pre-training algorithm using deep belief networks.

In [15] they have implemented the system using DTW technique. They have used Least Mean Square algorithm for removing the noise. They have got 93 % recognition rate using enhancement and 82 % without enhancement. We have evaluated our results on TI Digits dataset. The system has achieved 86.06 % accuracy on this dataset. Training of DNN requires much time. For achieving better recognition we must parallelize the training. Research is going on for discovering new learning algorithms for deep neural networks.

References

1. M.A. Anusuya, S.K. Katti.: Speech Recognition by Machine: A Review. In: International Journal of Computer Science and Information Security, Vol. 6, No. 3, pp. 181–205, (2009).
2. Khalid Saeed, Mohammad Kheir Nammous.: A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image, In: IEEE Transactions on Industrial Electronics, Vol. 54, No. 2, pp 887–897, (2007).
3. Veera Ala-Keturi.: Speech Recognition Based on Artificial Neural Networks. In: Helsinki Institute of Technology, (2004).
4. George E. Dahl, Dong Yu, Li Deng, and Alex Acero.: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. In: IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1, pp, 30–42, (2012).
5. Bo Li and Khe Chai Sim.: A Spectral Masking Approach to Noise-Robust Speech Recognition Using Deep Neural Networks. In: IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 22, No. 8, pp. 996–1305, (2014).
6. Prasad D Polur, Ruobing Zhou, Jun Yang, Fedra Adnani, Rosalyn S.: Hobsod Isolated Speech Recognition Using Artificial Neural Network. In: 2001 Proceedings of the 23rd Annual EMBS International Conference Istanbul, pp 1731–1734, (2001).
7. Xiaohui Zhang, Jan Trmal, Daniel Povey, Sanjeev Khudanpur.: Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks. In: IEEE International Conference On Acoustic, Speech and Signal, pp 214–219, (2014).
8. Xicai Yue, Datian Ye, Chongxun Zheng, Xiaoyu Wu.: Neural networks for improved text independent speaker identification. In: IEEE Engineering In Medicine And Biology, pp 53–58, (2002).
9. Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton.: Acoustic Modeling Using Deep Belief Networks. In: IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1, pp. 14–22, (2012).
10. Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury.: Deep neural networks for acoustic modeling in speech recognition. In: IEEE Signal Processing Magazine, pp 82–97, (2012).
11. Jonas Gehring, Wonkyum Lee, Kevin Kilgour, Ian Lane, Yaije Miao, Alex Waibel.: Modular Combination of Deep Neural Networks for Acoustic Modeling. In: INTERSPEECH (2013).
12. Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton.: Speech Recognition With Deep Recurrent Neural Networks. In: ICASSP, pp 6445–6449, (2013).
13. Song Yang, Meng Joo Er, and Yang Gao.: A High Performance Neural-Networks-Based Speech4 Recognition System. In: IEEE, pp 1527–1531, (2000).
14. Wouter Gevaert, Georgi Tsenov, Valeri Mladenov.: Neural Networks used for Speech Recognition. In: Journal Of Automatic Control, University Of Belgrade, Vol. 20, pp. 1–7, (2010).
15. Siva Prasad Nandyala, T. Kishore Kumar.: Real Time Isolated Word Recognition using Adaptive Algorithm. In: International Conference on Industrial and Intelligent Information, pp 163–168, (2012).