Deep belief networks for phoneme recognition in continuous Tamil speech–an analysis

Laxmi Sree Baskaran Raguram[1,*], Vijaya Madhaya Shanmugam[2]

1. G R Damodaran College of Science,
Avanashi Road, Peelamedu, Coimbatore 641014, India

2. PSGR Krishnammal College for Women,
Avinashi Road, Peelamedu, Coimbatore 641004, India

viporala@yahoo.com

ABSTRACT. *A combination of Gaussian Mixture Model and Hidden Markov Model has been used successfully in building acoustic models for speech recognition. These models have dominated this area for nearly three decades. Re-entry of neural networks in many clustering, classification and pattern recognition problems have triggered current researchers to focus in making use of its power in the area of speech recognition. This article compares the performance of Bernoulli-Bernoulli Deep Belief Networks (BBDBN) and Gaussian-Bernoulli Deep Belief Networks (GBDBN) on phoneme recognition of spoken speech in Tamil. In addition to that the impact of feature representation in the performance of acoustic model is also studied by using three different datasets built using different feature representation for the phoneme samples extracted from the continuous Tamil speech.*

RÉSUMÉ. *Une combinaison du modèle de mélange gaussien et du modèle de Markov caché a été utilisée avec succès dans la construction de modèles acoustiques pour la reconnaissance automatique de la parole. Ces modèles jouent des roles dominents depuis près de trois décennies. La réinsertion de réseaux de neurones dans de nombreux problèmes de regroupement, de classification et de reconnaissance de formes a amené les chercheurs actuels à se concentrer sur l'utilisation de son pouvoir dans le domaine de la reconnaissance automatique de la parole. Cet article compare les performances des réseaux de croyances profondes Bernoulli-Bernoulli (BBDBN en anglais) et des réseaux de croyances profondes Gaussian-Bernoulli (GBDBN en anglais) sur la reconnaissance phonémique de la parole tamoule. En outre, l'impact de la représentation de caractéristique sur les performances du modèle acoustique est également étudié à l'aide de trois bases de données différents construits en utilisant la représentation de caractéristique différentes pour les extraits de phonèmes dans la parole tamoule continue.*

KEYWORDS: *deep belief networks, phoneme recognition, speech recognition, artificial neural networks, deep learning, tamil speech, acoustic model, continuous speech, bernoulli-bernoulli, gaussian-bernoulli.*

MOTS-CLÉS: *réseaux de croyances profondes, reconnaissance de phonèmes, reconnaissance automatique de la parole, réseaux de neurones artificiels, apprentissage en profondeur, parole tamoule, modèle acoustique, parole continue, bernoulli-bernoulli, gaussian-bernoulli.*

DOI:10.3166/TS.34.137-151 © 2017 Lavoisier

1. Introduction

Automatic Speech Recognition is one of the alternative ways in human machine interaction. Milestones in this area have shown huge improvements in recognition accuracy using various methods to build acoustic models like Hidden Markov Model (HMM), Support Vector Machine (SVM), Gaussian Mixture Models and Artificial Neural Networks (ANN). Recent researches have shown the success of ANNs in modelling complex problems including speech recognition.

Vijayaditya Peddinti *et al.,* (2015) have proposed time delay neural network architecture. The long term temporal dependencies between acoustic events are modelled effectively using Recurrent Neural Networks (RNN). But the sequential nature of RNN learning algorithm takes higher time for training feed forward networks. The model uses sub-sampling method which considerably reduces computation time during training. Experimental results are carried out on various LVCSR tasks and the effectiveness of that proposed architecture is proved by varying data ranging from 3 to 1800 hours.

Mohamed *et al.,* (2012) have proposed a deep neural networks model that performs better phone recognition than Gaussian mixture model when it is applied to TIMIT dataset. The deep neural networks include many layers of features with large number of parameters. The model consists of two phases. In the first phase, the networks are pre-trained as multiple layers with window of spectral features without the use of any discriminative information. The second phase involves the use of back-propagation technique that performs discriminative fine tuning.

A novel Context-Dependent (CD) model for Large Vocabulary Speech Recognition (LVSR) has been proposed by Dahl *et al.* (2012) that implements deep belief networks and context dependent hidden markov model for phone recognition. A pre-trained Deep Neural Network Hidden Markov Model (DNN-HMM) models the distribution over tied triphone states called senones. The pre-training algorithm is mainly used to initialize deep neural network that helps in optimization and reduction of errors. Business search dataset is used to study the performance of algorithm and shows that it significantly outperform the conventional context-dependent Gaussian mixture model (GMM)-HMMs with increase in sentence accuracy.

Adeli and Jiang (2006) have proposed a novel dynamic time-delay fuzzy wavelet neural network model for nonparametric discovery of structures using the nonlinear autoregressive moving average with exogenous inputs. This model is based on the combination of four different concepts namely dynamic time delay neural network, wavelet, fuzzy logic, and the reconstructed state space concept from the chaos theory. The discrete wavelet packet method is used to remove noise in the signals. The reconstructed state space from the chaos theory is employed in this model to preserve the dynamics of time series and to construct the input vector. In order to capture the characteristics of the time series sensor data accurately and efficiently, techniques such as neural networks and fuzzy logic are employed in this model. Experimental results are carried out on a five-story steel frame to validate the performance of the proposed model.

Abiyev and Kaynak (2008) have proposed a model that combines fuzzy set theory and wavelet neural networks to solve the problem of uncertainty. The algorithm constructs fuzzy WNN based on a set of fuzzy rules where each rule consists of a wavelet function. The rules of the system are updated using Gradient Descent technique.

Hinton *et al.* (2012) deals on the poor performance of held-out test data for a large feed forward neural network. The author handles the problem of overfitting by randomly omitting half of the feature detectors on each training case. This prevents complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors. This approach helps each neuron learns to detect a feature that acts as a key feature in the process of classification given the combinatorially large variety of internal contexts in which it must operate. Random "dropout" have big improvements on many benchmark tasks and sets new records for speech and object recognition.

Lee *et al.,* (2008) being motivated in part by the hierarchical organization of the cortex, have proposed algorithms that try to learn hierarchical or deep structure from unlabeled data. While several authors have formally or informally compared their algorithms to computations performed in visual area V1 (and the cochlea), little attempt for mimicking computations at deeper levels in the cortical hierarchy has been made to evaluate these algorithms in terms of their fidelity. This paper presents an unsupervised learning model that faithfully imitates certain properties of visual area V2. Specifically, a sparse variant of the deep belief networks has been developed. Nodes in two layers of the network are learnt and identifies that the first layer results in localized, oriented, edge filters, similar to the Gabor functions known to model V1 cell receptive fields. Further, the second layer in the model encodes correlations of the first layer responses in the data. Specifically, it picks up colinear ("contour") features as well as corners and junctions. A quantitative comparison resulted witha an interesting fact that the encoding of these more complex "corner" features matched well with the results from the Ito & Komatsu's study of biological V2 responses. This suggests that the sparse variant of deep belief networks holds promise for modeling more higher-order features.

Graves *et al.,* (2013) show Recurrent neural networks (RNNs) are a powerful model for sequential data. End-to-end training methods such as Connectionist Temporal Classification make it possible to train RNNs for sequence labelling problems where the input-output alignment is unknown. The combination of these methods with the Long Short-term Memory RNN architecture has proved its strength in cursive handwriting recognition delivering state-of-the-art results. The paper investigates deep recurrent neural networks, which combine the multiple levels of representation that have proved so effective in deep networks with the flexible use of long range context that empowers RNNs. It has been found that deep Long Short-term Memory RNNs achieve a test set error of 17.7% on the TIMIT phoneme recognition benchmark when it was trained end-to-end with suitable regularisation.

Application of Deep Neural Networks in various complex problems successfully has motivated its application in Speech Recognition. The objective of this work is to analyse the performance of DBNs in phoneme recognition of continuous Tamil speech and its comparability with other existing methods based on HMM, GMM, etc. In this article, we have applied two types of DBN, namely Bernoulli - Bernoulli DBN (BBDBN) and Gaussian - Bernoulli DBN (GBDBN) on Tamil continuous speech data to identify the spoken phonemes. Here we have compared the performance of BBDBN and GBDBN on the Tamil speech data using the performance measures Root Mean Square Error (RMSE) and Phoneme Error Rate (PER). The performance of both variants of DBNs are also analysed for its accuracy by increasing the network depth.

This article is organized as follows. Architecture of Deep Belief Networks, its learning procedure and pre-training are discussed in the following section 2. Section 3 discusses about the experimental setup used in the analysis followed by Experimental results in section 4, Discussion in Section 5 and finally conclusion in section 6.

2. Deep belief networks

A DBN is an artificial neural network which comprises of many hidden layers. One of the challenges faced while modelling DBNs is formulating an appropriate training strategy for train the network. Greedy method and random method are methods generally used to initialize the parameters of the network. Solution trapped to local optima is one of the challenges faced while training a DBN. In this study we use a general discriminative training method which considers each pair of layers that is bipartite in nature as a Restricted Boltzmann Machine (RBM). This method initializes the DBN parameters randomly and further trains using back propagation technique.

A DBN is a Multi-layer Perceptron, which is considered as a stack of RBMs. An RBM is a bipartite network having two layers, the visible layer and the hidden layer. In RBMs, the connections are restricted to visible-hidden connections. The visible layer of the first RBM is fed with the feature vectors, which is passed on to the output layer of that RBM modelling the posterior probabilities of the hidden units of DBN. The output of one RBM acts as input to the succeeding RBM in the stack. Based on the distribution of vectors in observations the RBMs can be modelled either Bernoulli-Bernoulli or Gaussian-Bernoulli. In simple binary RBM/Bernoulli-Bernoulli RBM, both the visible and hidden units are binary and stochastic in nature. Gaussian-Bernoulli RBMs are usually used to model real-valued data. Thus, the DBN acts as a non-linear classifier with each of the hidden layer expressed as posterior probabilities. Each neuron in the hidden layer uses the logistic function to convert its input received from its lower layer to a scalar value which is passed on to the next layer. Learning in BBDBN and GBDBN is as follows.

2.1. Learning in bernoulli-bernoulli RBM (BBRBM)

In BBRBM, the connection weights and the biases of the neural units define the probability distribution over the joint states of the visible and hidden units through energy function:

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

where θ=(w,b,a) are model parameters and w_ij is the connection weight between the ith visible unit and the jth hidden node. V and H are the number of visible and hidden units respectively. The probability of the visible vector v for the given model parameters θ is given by:

$$p(v|\theta) = \frac{1}{Z}\sum_{h} e^{-E(v,h)} \tag{2}$$

The conditional probability distribution of hidden units, given the model parameters and visible units is represented as:

$$p(h_j = 1|v,\theta) = \sigma(a_j + \sum_{i=1}^{V} w_{ij}v_i) \tag{3}$$

where the σ is the sigmoidal function, given by $\boldsymbol{\sigma(x) = (1 + e^{-x})^{-1}}$.

The conditional probability distribution of visible units, given the model parameters and hidden units is represented as:

$$p(v_i = 1|h,\theta) = \sigma(b_j + \sum_{j=1}^{H} w_{ij}h_j) \tag{4}$$

2.2. Learning in gaussian-bernoulli RBM (GBRBM)

Similar to BBRBM, the connection weights and the biases of the neural units define the probability distribution over the joint states of the visible and hidden units through energy function which is stated as follows:

$$E(v,h|\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_i h_j - \sum_{i=1}^{V}\frac{(v_i - b_i)^2}{2} - \sum_{j=1}^{H} a_j h_j \tag{5}$$

The conditional probability distribution of visible units, given the model and hidden units are represented as a Gaussian function ($\mathcal{N}$):

$$p(v_i|h,\theta) = \mathcal{N}(b_i + \sum_{j=1}^{H} w_{ij}h_j, 1) \tag{6}$$

2.3. Pre-training the DBN

In this analysis, pertaining DBNs are done using Contractive divergence technique. Contrastive divergence is an efficient training procedure performing an approximate training for RBMs. The procedure repeatedly tries to reconstruct the visible vector from the hidden vector generated from the visible vector thus updating the weight parameters of the RBM. The weight parameters w_{ij} are updated as follows,

$$\Delta w_{ij} = \langle v_i h_j \rangle_d - \langle v_i h_j \rangle_r \tag{7}$$

In the above equation (7), the change in weight parameter Δw_{ij} is calculated, where the first term $\langle v_i h_j \rangle_d$ denotes the measured frequency for visible units with current training data and hidden units the posterior probabilities determined usin Eq(3) and the second term $\langle v_i h_j \rangle_r$ denotes the measured frequency for visible units being the reconstructed data constructed using Eq(4) when the hidden units were constructed using Eq(3), the one referred in the previous term. Figure 1 show how contrastive divergence is used in training RBM.

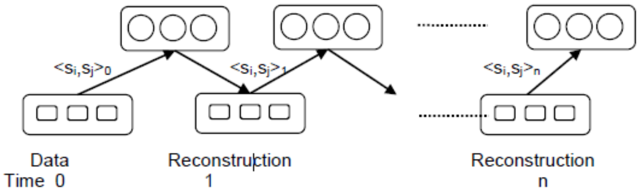

Figure 1. Training RBM using contrastive divergence

The outline involved in building the BBDBN/GBDBN is listed below in the algorithm:

Algorithm 1. Steps to build BBDBN/GBDBN based acoustic model

1. Segment the continuous Tamil speech data into phonetic segments using Graphcut based segmentation algorithm.
2. Build the monophone training dataset and test dataset.
3. Decide the DBN architectural parameters number of layers in DBN and number of neurons in each layer and design the BBDBN/GBDBN.
4. Initialize the weight and bias parameters of BBDBN/GBDBN with random values in the range (0,1).
5. Pre-train the DBN using contrastive divergence.
6. Train DBN with back propagation technique with monophone train dataset.
7. Test the acoustic model build with monophone test dataset.

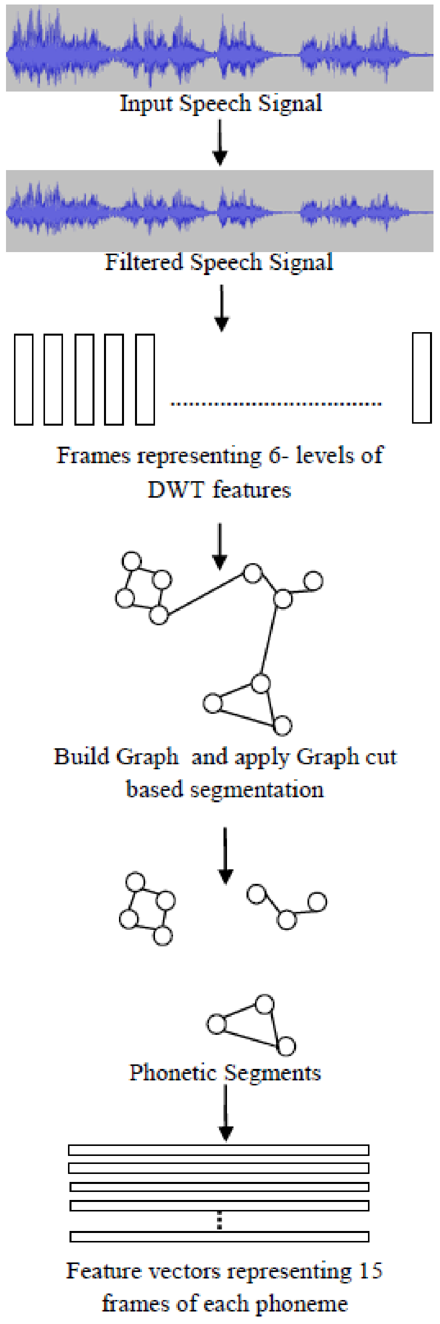
Input Speech Signal
Filtered Speech Signal
Frames representing 6- levels of DWT features
Build Graph and apply Graph cut based segmentation
Phonetic Segments
Feature vectors representing 15 frames of each phoneme

Figure 1. Steps in building the Dataset

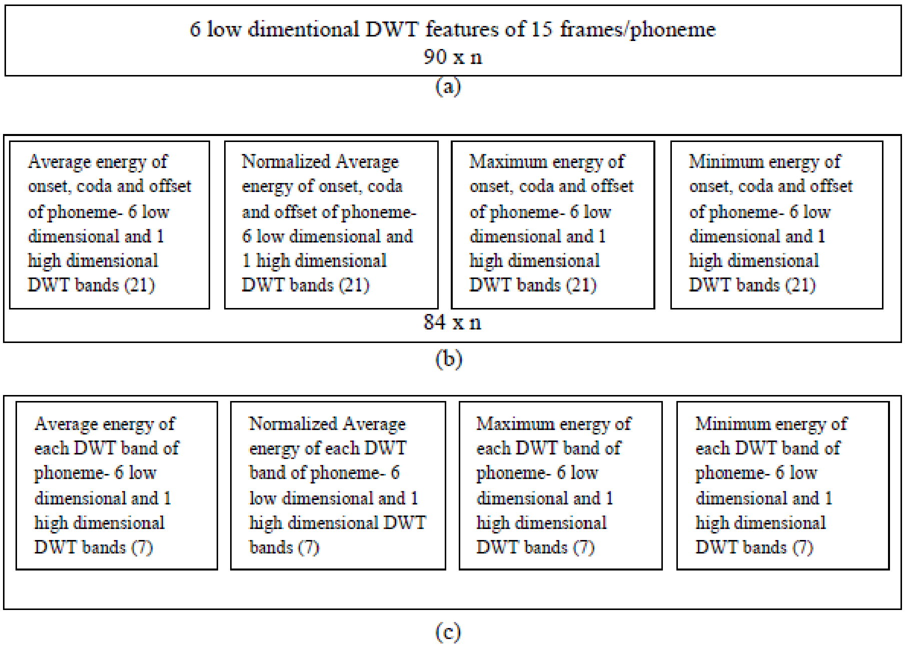
6 low dimentional DWT features of 15 frames/phoneme
90 x n
(a)
Average energy of onset, coda and offset of phoneme- 6 low dimensional and 1 high dimensional DWT bands (21)
Normalized Average energy of onset, coda and offset of phoneme- 6 low dimensional and 1 high dimensional DWT bands (21)
Maximum energy of onset, coda and offset of phoneme- 6 low dimensional and 1 high dimensional DWT bands (21)
Minimum energy of onset, coda and offset of phoneme- 6 low dimensional and 1 high dimensional DWT bands (21)
84 x n
(b)
Average energy of each DWT band of phoneme- 6 low dimensional and 1 high dimensional DWT bands (7)
Normalized Average energy of each DWT band of phoneme- 6 low dimensional and 1 high dimensional DWT bands (7)
Maximum energy of each DWT band of phoneme- 6 low dimensional and 1 high dimensional DWT bands (7)
Minimum energy of each DWT band of phoneme- 6 low dimensional and 1 high dimensional DWT bands (7)
(c)

Figure 3. Features selected for datasets (a) DWTFS, (b) DWTES and (c) DWTNES

3. Experimental setup

3.1. Speech corpus

The corpus Kazhangiyam in (Laxmi Sree and Suguna, 2016) has been extended with additional speech data and used in this work. The corpus consists of 9 hours of speech data spoken by 40 speakers including both male and female in the age group of 18 to 45. The speech was recorded in a controlled environment. The sampling frequency was set to 16 kHz. Phonemes were extracted from the wav file using the Graphcut based segmentation algorithm discussed in (Laxmi Sree and Vijaya, 2016) to build the datasets. This Graphcut based segmentation algorithm represents the features of each speech frame as a node in the graph and the similarity between these nodes as edge weights. It then performs repeated bipartition of graphs to produce the required segmentation. The corpus includes three datasets namely Discrete Wavelet Transform Feature Set (DWTFS), Discrete Wavelet Transform Energy Set (DWTES), Discrete Wavelet Transform Normalized Energy Set (DWTNES). The datasets are built with varied representation of DWT features extracted for the speech signal. DWTFS dataset is formed with six low dimensional DWT features of phonemes segmented from continuous Tamil speech with a total of 90 features. DWTES dataset is formed by splitting the DWT features of the phoneme into three parts onset, coda and offset parts whose total energy, normalized energy, maximum energy and minimum energy in each band is considered with a total of 84 features to represent each phoneme. DWTNES dataset is formed with average DWT features, normalized DWT features, maximum and minimum energies of each band as the features of the phoneme (28 features). As a whole 6 hours of speech data has been used in this work. It comprises around 1,87,452 samples out of which 70% forms the training dataset and 30% forms the testing dataset. The dataset was split to ensure both the training and testing datasets to cover all the phoneme classes considered.

4. Experimental results

Both DBNs used here are pre-trained using Contrastive Divergence technique. The pre-training trains one RBM at a time in the stack and proceeds to the next one in the stack. Each RBM was pre-trained with 1000 epochs. Once the pre-training is complete, the whole DBN is trained using back propagation learning technique. The training is conducted for 1000 epochs with a batch size of 100 data points, step size 0.1, initial momentum 0.5, final momentum 0.9 and weight cost 0.0002.

Table 1. RMSE values of 4 layer BBDBN and GBDBN while training and testing

Datasets	DWTFS	DWTES	DWTNES
BBDBN-Training	0.015124	0.048565	0.015577
GBDBN-Training	0.05973	0.059793	0.063981
BBDBN-Testing	0.014913	0.048145	0.015571
GBDBN-Testing	0.059303	0.059718	0.063829

Experiments have been conducted on BBDBN and GBDBN with all the three datasets. The results of the experiment are as follows. A comparison on the RMSE values while training BBDBN and GBDBN with three datasets DWTFS, DWTES and DWTNES is shown in the Table 1. The table shows the results of four layer DBNs. The RMSE values while training/testing BBDBN seems to be much lower than training/testing GBDBN for all the three data representations. The RMSE values while training BBDBN for various datasets DWTFS, DWTES and DWTNES are 0.015124, 0.048565 and 0.015577 respectively. The RMSE values while training GBDBN for various datasets DWTFS, DWTES and DWTNES are 0.05973, 0.059793 and 0.063981 respectively. It is clear that DWTFS dataset provides lower RMSE value when compared to the other two datasets DWTES and DWTNES. Figure 4 show the Phone Error Rate (PER) of the 4-layer BBDBN and GBDBN during both training and testing with different datasets under consideration. It is clear from the figure that the performance of BBDBN is better for all the three datasets when compared to GBDBN in terms of PER. In addition, it is observed that the data representation in DWTFS performs better than DWTES and DWTNES with respect to PER.

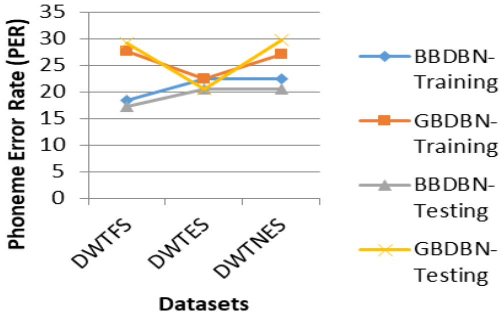

Figure 4. Phoneme error rate of BBDBN and GBDBN on train and test data

Table 2. Performance comparison through RMSE values of BBDBN and GBDBN networks by increasing the number of layers (for DWTFS dataset)

Model/No. of Layers	4	5	6
BBDBN-Training	0.015124	0.015266	0.015232
GBDBN-Training	0.05973	0.061828	0.063849
BBDBN-Testing	0.014913	0.01529	0.015281
GBDBN-Testing	0.059303	0.062027	0.064215

Table 3. Phone error rate (PER) during training and testing 4-layer, 5-layer and 6-layer BBDBNs and GBDBNs (for DWTFS dataset)

Model/No. of Layers	4	5	6
BBDBN-Training	18.42	18.78	14.62
GBDBN-Training	27.56	22.34	22.34
BBDBN-Testing	17.28	19.94	14.97
GBDBN-Testing	29.24	20.53	20.53

Analysis has been proceeded by increasing the number of layers in the DBN and the respective change in the RMSE values of both BBDBN and GBDBN networks have been recorded (refer Table 2). It is observed that still the RMSE values of 4, 5 and 6-layers BBDBN seem to be much smaller than that of their GBDBN counterparts for DWTFS dataset. DWTFS dataset has been used for this analysis. Figure 5 plots the PER for the train and test dataset of DWTFS on BBDBN and GBDBN. It is seen that the Phone Error Rate is better in BBDBN with various depth of the network when compare to their equivalents in GBDBN. The observation also shows that increasing the number of layers in the network reduces the error rate by increasing the performance the network classification except very few situations.

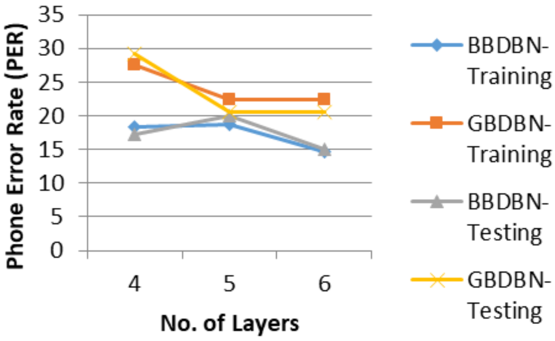

Figure 5. Phone error rate (PER) during training and testing 4-layer, 5-layer and 6-layer BBDBNs and GBDBNs (for DWTFS dataset)

Table 4 lists down the RMSE values of training and testing BBDBN and GBDBN with DWTES dataset by varying the number of layers in DBNs. The performance in terms of PER of the networks BBDBNs and GBDBNs are compared in Table 5 and Figure 6. It can be noticed that the PER of BBDBN on DWTES train and test dataset also shows that it models better when compared to GBDBN. But the PER of the networks show a decline in the performance by increasing the number of layers of network with DWTES dataset.

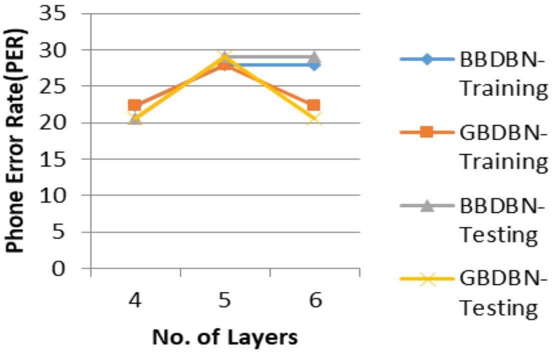

Figure 6. Comparison of phone error rate (PER) during training and testing 4-layer, 5-layer and 6-layer BBDBNs and GBDBNs (for DWTES dataset)

Table 4. Comparison of RMSE values while training and testing 4-layer, 5-layer and 6-layer BBDBNs and GBDBNs (Dataset: DWTES)

Model/No. of Layers	4	5	6
BBDBN-Training	0.048565	0.055574	0.059939
GBDBN-Training	0.059793	0.059684	0.065982
BBDBN-Testing	0.048145	0.05577	0.060274
GBDBN-Testing	0.059718	0.060007	0.066194

5. Discussion

The results presented in the previous section shows that the representation of features to build the dataset plays an important role in the performance of the acoustic model that is built using BBDBN or GBDBN. This study shows that the dataset built using direct DWT features (DWTFS) when compared to band-wise energy features dataset (DWTES) and dataset built with normalized energy of onset, coda and trailing portions of phonemes (DWTNES). For a 4-layer BBDBN using DWTFS dataset have achieved a better performance with 17.28% PER whereas the other two datasets have achieved 20.53% PER for BBDBN with test data. But in case of 4-layer GBDBN the best performance is achieved by using DWTES with 20.53% PER whereas using the other two dataset have yield PER around 29% (Table 3).

Table 5. Phone error rate (PER) during training and testing 4-layer, 5-layer and 6-layer BBDBNs and GBDBNs (for DWTES dataset)

Model/No. of Layers	4	5	6
BBDBN-Training	22.34	27.88	27.88
GBDBN-Training	22.34	27.88	22.34
BBDBN-Testing	20.53	29.12	29.12
GBDBN-Testing	20.53	29.12	20.53

Analysis on the performance of the networks by varying the number of layers in the DBNs show that there is an increase in the performance of both types of DBNs with the increase in the number of layers with the best PER of 14.62% for DWTFS train dataset on BBDBN and 14.97% for DWTFS test dataset on BBDBN. The results of training and testing DBNs with DWTES have turned up better for 6-layer GBDBN and 4-layer BBDBN with 20.53% PER.

The best result of the current analysis is compared with the results of other existing models in Table 6 to prove the stgreangth of DBN. The other models compared here uses TIMIT core test set, which uses MFCC for feature representation. In our work, KAZHANGIYAM corpus (Laxmi Sree and Suguna, 2016) which uses DWT features for input representations is used. Using BBDBN for Phoneme recognition of Tamil speech provides around 5% improvement over the previously reported PER.

Table 6. Performance of BBDBN and GBDBN compared with results of previously reported models

Method	PER
Conditional Random Field (Morris and Fosler-Lussier, 2006)	34.80%
Large-Margin GMM (Sha and Saul, 2006)	33%
CD-HMM (Hifny and Renals, 2009)	27.30%
Heterogenous Classifiers (Halberstadt, 1999)	24.40%
Monophone Deep Belief Networks (Li and Yu, 2007)	20.70%
BBDBN (this work)	14.97%
GBDBN (this work)	20.53%

6. Conclusion

The analysis reported in this paper studies the performance of Phoneme recognition for Tamil continuous speech by building BBDBN and GBDBN acoustic models. The study has been done using the self developed speech corpus Kazhangiyam (Laxmi Sree and Suguna, 2016). The acoustic models have shown varied performance with different input representation of features that has been presented through the datasets DWTFS, DWTES and DWTNES. The input representations in DWTFS reports better results, where as the ones of DWTES lacks performance sometime and DWTNES performs poor when compared to the other two datasets.

The analysis of BBDBN and GDDBN was also done by varying the depth of the DBN (number of layers), which gives a common inference of achieving better performance by increasing the number of layers. It is observed that pre-training the DBNs consumes a lot of CPU time. Increasing the number of layers of a DBN also increases the training time. Working on the time constraint of the training part of the DBN can provide an additional advantage while building the acoustic model. The comparison of BBDBN used in this analysis with other existing works show an improvement of 5% in the performance of model.

References

Abiyev R. H., Kaynak O. (2008). Fuzzy wavelet neural networks for identification and control of dynamic plants—a novel structure and a comparative study. *IEEE transactions on Industrial Electronics,* Vol. 55, No. 8, pp. 3133-3140. http://dx.doi.org/10.1109/TIE.2008.924018

Adeli H., Jiang X. M. (2006). Dynamic fuzzy wavelet neural network model for structural system identification. *Journal of Structural Engineering,* Vol. 132, No. 1, pp. 102-111. http://dx.doi.org/10.1061/(ASCE)0733-9445(2006)132:1(102)

Dahl G. E., Yu D., Deng L., Acero A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 20, No. 1, pp. 30-42. http://dx.doi.org/10.1109/TASL.2011.2134090

Graves A., Mohamed A., Hinton G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* http://dx.doi.org/10.1109/ICASSP.2013.6638947

Halberstadt A. K. (1999). Heterogeneous acoustic measurements and multiple classifiers for speech recognition. *Massachusetts Institute of Technology.*

Hifny Y., Renals S. (2009). Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 2, pp. 354-365. http://dx.doi.org/10.1109/TASL.2008.2010286

Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *Neural and Evolutionary Computing,* pp. 1-18. https://arxiv.org/pdf/1207.0580.pdf

Laxmi Sree B. R., Suguna M. (2016). AAYUDHA: A tool for automatic segmentation and labelling of continuous tamil speech. *International Journal of Computer Applications,* Vol. 143, No. 1, pp. 31-35. http://dx.doi.org/10.5120/ijca2016910002

Laxmi Sree B. R., Vijaya M. S. (2016). Graph cut based segmentation method for tamil continuous speech. *Digital Connectivity–Social Impact. Springer Singapore*, pp. 257-267. http://dx.doi.org/10.1007/978-981-10-3274-5_21

Lee H., Ekanadham C., Ng A. Y. (2008). Sparse deep belief net model for visual area V2. *Proceeding NIPS'07 Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 873-880.

Li D., Yu D. (2007). Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing -ICASSP '07,* Vol. 4, pp. 445-448. http://dx.doi.org/10.1109/ICASSP.2007.366945

Mohamed A., Dahl G. E., Hinton G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 14-22. http://dx.doi.org/10.1109/TASL.2011.2109382

Morris J., Fosler-Lussier E. (2006). Combining phonetic attributes using conditional random fields. *Ninth International Conference on Spoken Language Processing*. http://www.isca-speech.org/archive/interspeech_2006/i06_1287.html

Peddinti V., Povey D., Khudanpur S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3214-3218.

Sha F., Saul L. K. (2006). Large margin Gaussian mixture modeling for phonetic classification and recognition. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. https://doi.org/10.1109/ICASSP.2006.1660008