Performance Evaluation of Neural Networks for Speaker Recognition

Nishant Mishra,
Department of Electronics and Communication Engineering,
Birla Institute of Technology,
Mesra, Ranchi, Jharkhand 835215
mnishant2@gmail.com,

Prasun Anand,
Department of Electronics and Communication Engineering,
Birla Institute of Technology,
Mesra, Ranchi, Jharkhand 835215
prasundps@gmail.com

Mahesh Chandra
Department of Electronics and Communication Engineering,
Birla Institute of Technology,
Mesra, Ranchi, Jharkhand 835215
shrotriya@bitmesra.ac.in

Zainab Feroz,
Department of Electronics and Communication Engineering,
Birla Institute of Technology,
Mesra, Ranchi, Jharkhand 835215
zainabferoz101@gmail.com

***Abstract*—Speaker Recognition is one of the principle problems in Speech processing. The performance of speaker recognition systems can be improved by carefully choosing and calculating suitable features, which is an arduous task. Therefore, the learning based approach has been found to be simpler, more general and with the rapid growth in Artificial Intelligence, more accurate. This paper is a comparative study of the performance of different neural networks in speaker recognition. The focus of this work is to find which of these learning algorithms is more accurate, less complex, and more generic when it comes to speaker recognition. A database of *5000* utterances, *100* for each of the *50* different speakers, in both clean and noisy environment, with varying levels of noise was used. The MFCC (Mel Frequency Cepstral Coefficients) of these utterances were used as features to train and evaluate the neural networks. Accuracy of all neural networks was expectedly very high *(>90%)* for clean data, large variations coming in with introduction and change in the level of noise. RBFNN has been shown to consistently perform well under all conditions. DNN was the other consistent performer and has the potential to outperform other techniques, if trained on more data.**

Keywords—MFCC, Neural Network, SLFN, PNN, RBFNN, DNN, Speaker Recognition

I. INTRODUCTION

Automatic speaker recognition consists of identifying the speaker based on the utterance. Speaker recognition [1] identifies the various speakers based on their voice characteristics. It can be classified into two types: Text Dependent Speaker Recognition and Text Independent Speaker Recognition. In text dependent systems, a predefined utterance is used to train as well as test the system. Text Independent systems have no constraint on their speech content. The utterances used for testing are independent of those used for training. The basic approach for designing a text dependent speaker recognition system includes preparing a suitable database for training and testing the system, extracting features from the different speech samples, followed by the feature classification step.

The two most common techniques for feature extraction are LPC – Linear Predictive Coding [2] and MFCC- Mel-Frequency Cepstral Coefficients [3-4]. LPC parameters depend on speech production and are a linear combination of past values while MFCC depends on human hearing perception. Feature extraction is an important step which gives the specific information contained in the speech signal. These features depend on various parameters such as intensity, frequency, zero crossing rate, level crossing rate, etc. and also on the age of the speaker, gender, accent, speaking rate, dimensions of the vocal tract and environmental conditions.

Feature classification includes dividing the data in the category which it belongs to. The various models/algorithms that can be used for classification include Hidden Markov Model, Gaussian Mixture Model, Self Organising Maps, Neural Networks, etc.

The objective is to devise a system which gives the best results and improved performance over previous systems. Here, in this paper, MFCC features have been used and a comparative study of different neural networks –Single Hidden Layer Feed Forward Neural Network (SLFN), Probabilistic Neural Network (PNN), Radial Basis Function Neural Network (RBFNN) and Deep Neural Network (DNN) is shown for noisy as well as clean data.

II. METHODOLOGY

A. Database

The database used for training and testing the automatic speaker recognition system included Hindi Digit (0-9) samples by 50 different speakers. Each digit was spoken 10 times, thus making the number of utterances by each speaker, 100 and the total number of utterances, 5000. The database consisted of clean data and noisy data with the noise levels varying from -5dB, 0dB, 5dB, 10dB, 20dB and 30dB.

B. Mel-Frequency Cepstral Coefficients

978-1-5386-8158-9/19/$31.00 ©2019 IEEE

The entire dataset was divided into training and testing data where approximately three-fourths of the samples were used to train our system and the rest were used for testing. The samples were labelled according to the speaker. 13 Mel-Frequency Cepstral Coefficients were obtained for each utterance giving a sequence of acoustic feature vectors. These feature vectors were then used for training the neural networks.

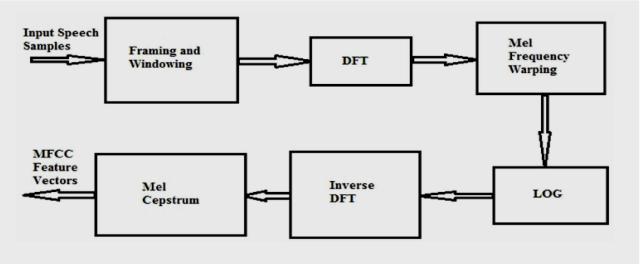

Fig.1. MFCC Feature Extraction

III. Neural Networks

Neural network is a computing system composed of many parallel operating processing elements which has the capability of acquiring, storing and utilizing experiential knowledge. These computing systems draw their inspiration from biological neural networks that constitute animal brain. The idea is to make the computer learn and perform a task, through experiencing data, without explicitly hard coding to perform that task. The architecture and type of the network depends on the kind of problem statement in hand. The following neural networks have been used for Speaker Recognition and a comparative study of their performances is provided.

A. SLFN

In this architecture there are three layers of neurons. Each neuron applies a nonlinear activation function to its input to generate its output except the neuron in the first layer which just acts as interface between the network and its environment. The output of the last layer is compared to that of required result and the error between them is obtained. This error is then reduced by using backpropagation [5-6] iteratively on the training data. The trained model is used to classify unseen data to its proper class.

B. PNN

A PNN [7] learns by approximating the probability distribution function of training dataset. The closeness of input data is compared with all the training neurons and the data is classified into the category with maximum closeness.

C. DNN

The Deep Neural Network [8] architecture used in this study is a 4-layer perceptron [9]. By definition any neural network architecture that has more than one hidden layer is a DNN. It has a similar working principle to that of a 3-layer perceptron and uses the same method for reducing error. The advantage of an additional layer is that it is better at non-linear separation and has better noise tolerance. DNNs, though, are very computationally intensive and data hungry.

D. RBFNN

Radial Basis Function Neural Network [10-11] is based on the principle of Cover's theorem [12] which states that a complex pattern classification problem when casted into a higher dimensional space nonlinearly, is more likely to be separable than in a low-dimensional space. In the most basic form it is a three-layer network with the following function:

First layer acts as an interface between the network and the environment i.e. it accepts the input data. Second layer is the only hidden layer and is used to map the input space to higher dimensional space through a nonlinear transformation. Gaussian functions, multi-quadrics, inverse-multi quadrics etc. can be used for the same. Third Layer gives the output of the network which is then compared to required output to obtain error. This error is then reduced to requisite level iteratively, using Least Mean Square algorithm [13] to train the network.

IV. Results and Discussion

MFCC features of each utterance were found out using the MIR Toolbox [14] in MATLAB [15]. The features were vectors of length 13 each. These vectors were used to train all the four neural networks for speaker recognition for 10, 20, 40 and 50 speakers respectively, on both clean and noisy data. The accuracy obtained for all four neural networks has been tabulated below. Assuming SNR of clean data to be 40dB (P_{speech} = 10000 P_{noise}), a comparison of performances of all four neural networks for different number of speakers is also shown as plots of Accuracy vs SNR (in dB). As can be inferred, the overall performance of RBFNN was the most consistent for all noise levels, while DNN learnt more complex features and, provided more data, should trump RBFNN in terms of accuracy. PNN on the other hand, was the fastest to train and compute, although it requires more memory and shows the largest fluctuations in accuracy with varying levels of noise.

TABLE I. ACCURACY OF SPEAKER RECOGNITION USING SLFN

Number of Speakers	**Clean Data**	**Noisy Data**					
		-5dB	0dB	5dB	10dB	20dB	30dB
50	92.18	76.81	84.86	88.92	91.096	92.352	92.24
40	94.53	76.95	86.54	90.57	92.66	94.15	94.81
20	96.56	82.85	89.35	92.62	94.31	95.24	95.73
10	98.8	90.68	95.64	94.76	97.64	98.44	99.08

TABLE II. ACCURACY OF SPEAKER RECOGNITION USING PNN

Number of Speakers	**Clean Data**	**Noisy Data**					
		-5dB	0dB	5dB	10dB	20dB	30dB

50	96.096	31.47	51.72	71.44	85.99	95.23	95.76
40	97.2	34.07	54.02	73.82	86.85	95.73	97.13
20	97.82	44.32	62.54	78.68	90.84	96.61	97.72
10	99.2	59.48	73.28	75.32	96.36	99.08	99.24

TABLE III. ACCURACY OF SPEAKER RECOGNITION USING DNN

Number of Speakers	**Clean Data**	**Noisy Data**					
		-5dB	0dB	5dB	10dB	20dB	30dB
50	86.07	61.26	72.87	80.87	85.6	87.00	87.90
40	89.37	60.34	66.87	82.13	83.87	87.79	89.03
20	96.13	78.90	87.85	91.33	94.28	95.34	96.01
10	98.6	90.12	95.60	95.56	98.16	98.6	98.72

TABLE IV. ACCURACY OF SPEAKER RECOGNITION USING RBFNN

Number of Speakers	**Clean Data**	**Noisy Data**					
		-5dB	0dB	5dB	10dB	20dB	30dB
50	96.22	85.76	89.62	92.71	94.44	95.84	96.29
40	97.19	87.44	90.72	93.18	95.15	96.96	97.23
20	97.47	89.36	92.11	94.39	95.96	97.52	97.99
10	98.8	91.46	94.4	94.72	97.64	98.64	98.74

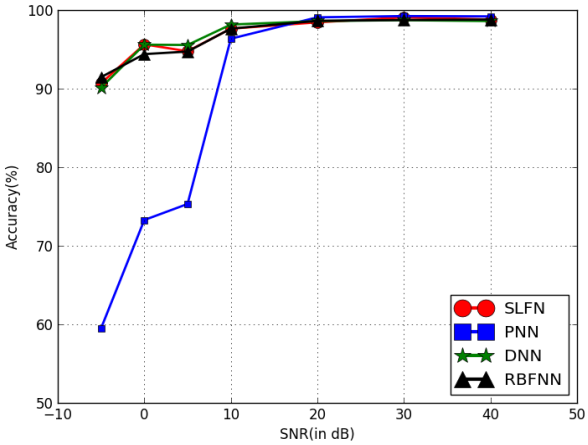

Fig.2. Accuracy Vs SNR (in dB) for 10 speakers

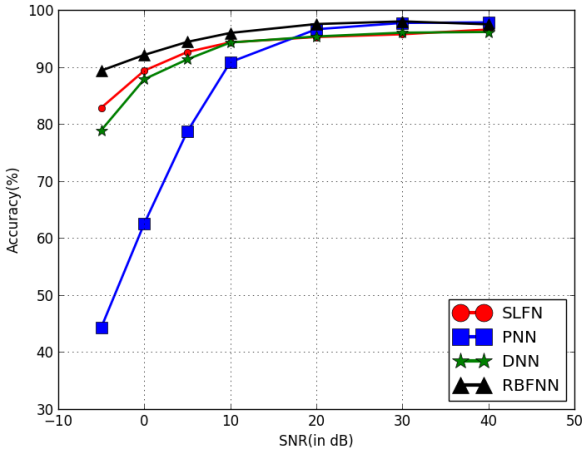

Fig.3.Accuracy Vs SNR (in dB) for 20 speakers

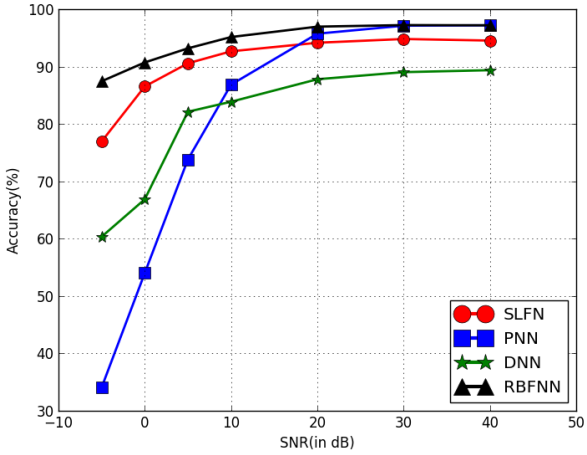

Fig.4. Plot of Accuracy Vs SNR (in dB) for 40 speakers

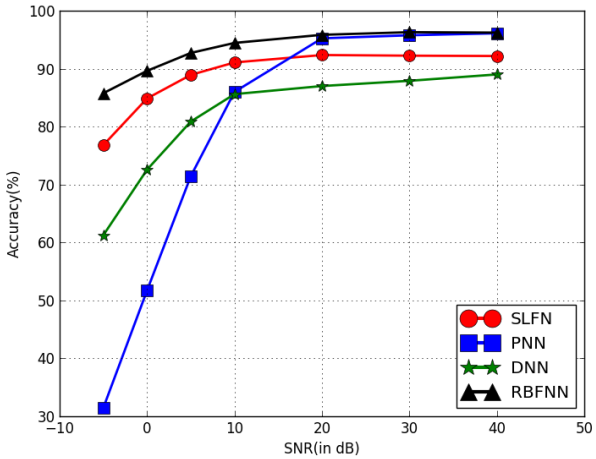

Fig.5. Accuracy Vs SNR (in dB) for 50 speakers

V. CONCLUSION AND FUTURE SCOPE

In this work, the performance of different neural networks has been compared for speaker recognition. The performance depends on number of factors such as amount of data per speaker, features extracted, amount of noise present, etc. Deep learning was observed to require much more data per speaker for better performance as the number of speakers and hence the

variance in data increases. All networks are adversely affected by presence of noise, especially PNN, which basically measures the proximity of test data with training data, shows large deviations.

It can be established that the quantity and quality of data is of utmost importance for any learning based approach. Apart from it, for future improvement in results, choosing networks by horses for courses approach or the ensemble methods are the way forward.

REFERENCES

[1] Beigi, Homayoon. Fundamentals of speaker recognition. Springer Science & Business Media, 2011.

[2] O'Shaughnessy, Douglas. "Linear predictive coding." IEEE potentials 7.1 (1988): 29-32.

[3] Tiwari, Vibha. "MFCC and its applications in speaker recognition." International journal on emerging technologies 1.1 (2010): 19-22., 92, pp.68-73

[4] Martinez, J., Perez, H., Escamilla, E., & Suzuki, M. M. (2012, February). Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques. In *Electrical Communications and Computers (CONIELECOMP), 2012 22nd International Conference on*(pp. 248-251). IEEE.

[5] Lecun, Y. (1988). A theoretical framework for back-propagation. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School, CMU, Pittsburg, PA* (pp. 21-28). Morgan Kaufmann.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing, volume 1. MIT Press, Cambridge, MA, 1986.

[7] Specht, Donald F. "Probabilistic neural networks." *Neural networks* 3.1 (1990): 109-118.

[8] Richardson, Fred, Douglas Reynolds, and Najim Dehak. "Deep neural network approaches to speaker and language recognition." *IEEE Signal Processing Letters* 22.10 (2015): 1671-1675.

[9] Gardner, Matt W., and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." *Atmospheric environment* 32.14 (1998): 2627-2636.

[10] Strumiłło, Paweł, and Władysław Kamiński. "Radial basis function neural networks: theory and applications." *Neural Networks and Soft Computing*. Physica, Heidelberg, 2003. 107-119.

[11] Park, Jooyoung, and Irwin W. Sandberg. "Universal approximation using radial-basis-function networks." *Neural computation* 3.2 (1991): 246-257.

[12] Cover, Thomas M. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition." *IEEE transactions on electronic computers* 3 (1965): 326-334.

[13] Liu, Weifeng, Puskal P. Pokharel, and Jose C. Principe. "The kernel least-mean-square algorithm." *IEEE Transactions on Signal Processing* 56.2 (2008): 543-554.

[14] Lartillot, Olivier, and Petri Toiviainen. "A Matlab toolbox for musical feature extraction from audio." *International Conference on Digital Audio Effects*. 2007.

[15] MATLAB 2015a, The MathWorks, Natick, 2015.