ICITKM
Proceedings of the First International Conference on Information Technology and Knowledge Management pp. 107–111
DOI: 10.15439/2018KM40
ISSN 2300-5963 ACSIS, Vol. 14

The Use of Deep Learning in Speech Enhancement

Rashmirekha Ram[1], Mihir Narayan Mohanty[2]

Electronics & Communication Engineering, ITER,
Siksha 'O' Anusandhan, Deemed to be University, Bhubaneswar, India
{ram.rashmirekha14, mihir.n.mohanty}@gmail.com

***Abstract*—Deep learning is an emerging area in current scenario. Mostly, Convolutional Neural Network (CNN) and Deep Belief Network (DBN) are used as the model in deep learning. It is termed as Deep Neural Network (DNN). The use of DNN is widely spread in many applications, exclusively for detection and classification purpose. In this paper, authors have used the same network for signal enhancement purpose. Speech is considered for the input signal with noise. The model of DNN is used with two layers. It has been compared with the ADALINE model to prove its efficacy.**

***Index Terms*—Speech Enhancement; Neural Networks; Adaptive Linear Neuron; Deep Neural Network; Signal-to-Noise Ratio; Perceptual Evaluation of Speech Quality.**

I. Introduction

SPEECH enhancement in noisy conditions is always a fascinating and challenging task for speech recognition, mobile communications, teleconferencing systems, hearing aids design etc. The objective of speech enhancement is to reduce the noise as well as to increase the SNR of the noisy speech signals in adverse environment. From the several decades, researchers have focused more attention in this area. But the results are not always satisfactory in terms of quality and intelligibility [1].

Speech signals are nonstationary in nature. Adaptive filters perform better in real time environment. Many adaptive algorithms are designed, such as Least Mean Squares (LMS), Recursive Least Squares (RLS), Normalized LMS (NLMS) and different variations in LMS. The authors compared LMS and RLS with the State Space Recursive Least Squares (SSRLS) algorithm. The improvement in SNR of the proposed algorithm is much better than the existing algorithms [2] [3]. Spectral subtraction (SS) algorithm suppresses background noise and proves better for stationary noise. S.Vihari *et.al.* proposed a noise estimation algorithm based on the Decision Directed approach. The Wiener filter and the SS algorithms are tested for nonstationary noise and outperform better [4].

In this digital world, machine learning approaches are more demanding day to day. Earlier Adaptive Linear Neuron Network (ADALINE) is designed as the single layer neural network. It is based on the principle of Multilayer Perceptron (MLP) [5]. The network consisting of the activation function and the function's output is utilized for adapting the weights. Generally Fourier Transform (FT) is used for extracting the features as the magnitude and phase and passed to the ADALINE for training. The Discrete Cosine Transform (DCT) and the Fractional DCT (FrDCT) coefficients are extracted from the noisy speech signal and ADALINE trains these features. The better enhanced signal is obtained in terms of SNR and PESQ for FrDCT ADALINE [6]. Artificial Neural Network (ANN), Convolutional Neural Network (CNN) are also designed for speech enhancement. An overview of the Neural Network is proposed in [7] [8].

Understanding of speech is difficult in noisy environment. To improve the quality and intelligibility of the speech signal, neural network based speech enhnacement is proposed in [9]. To acquire the high SNR, the time-frequency bins are decomposed and extracted. These features are fed to the network for better accuracy. Yong Xu *et.al.* proposed a regression based Deep Neural Network (DNN) for speech enhancement. A mapping function is calculated between the noisy features and the clean features. Different hidden layers are considered for SNR measurement [10] [11]. An improved LMS adaptive filter combines with the DNN for speech enhancement. The adaptive filter coefficients are estimated by the Deep belief Network (DBN) and the enhanced speech is prevailed through ILMSAF [12]. Reinforcement learning can be used for optimization of the large set of DNN training sets. The cochlear implant is designed based on the application of the DNN used for speech enhancement [13] [14]. Ram *et.al.* performed the enhancement the speech signals through the DNN with the hidden layers three [16]. Audio and Visual enhancement can also be achieved through the DNN [17].

The rest of the work is organized as follows: speech enhancement using ADALINE is explained in Section 2. Section 3 presents the speech enhancement using DNN considering two hidden layers: DNN_1 and DNN_2. Comparison results of DNN and ADALINE for speech enhancement are presented in Section 4 and Section 5 concludes the work.

II. Speech Enhancement using ADALINE

ADALINE is a simple neural network used for noise cancellation and is based on the principle of the LMS

algorithm. Fig.1 represents the block diagram of the ADALINE used for speech enhancement.

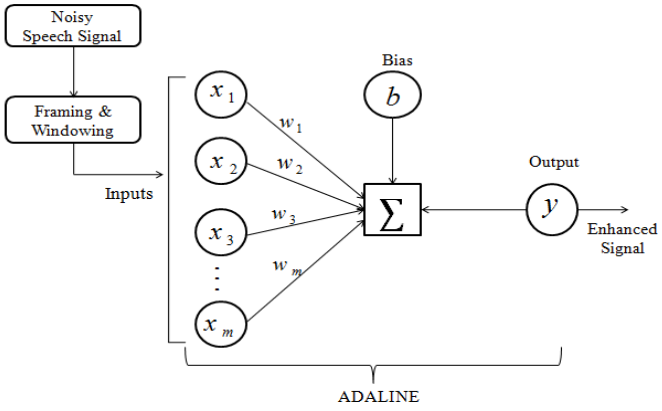

Fig.1. Speech Enhancement using ADALINE

This adaptive network is consisting of a single neuron with connected weights and a single bias. Because of nonstationary nature of speech signals, it is divided into overlapping frames before processing. To avoid spectral leakage, hamming window of length 256 is multiplied to all the overlapping frames .These overlapping windowed frames are processed in the network as inputs for enhancement.

To obtain the output of each instant of the speech signal, the each set of weights and biases are calculated. The input layers $x_1, x_2, ..., x_m$ are connected to the output y by interconnecting the weights $w_1, w_2, ..., w_m$ and bias b. The following steps are followed for speech enhancement using ADALINE.

- Set the weights $\{w(m)\}$ at 0.25 and biases $\{b(m)\}$ at 0.825 experimentally.
- Se the learning rate parameter (l) as 0.5.
- Consider the clean signal as the target signal $\{t(m)\}$.
- Set the noisy signal as the input signal $\{x(m)\}$.
- For each time index m, the output signal $\{y(m)\}$ and the error $\{a(m)\}$ can be calculated as

$$y(m) = w(m) * x(m) + b(m)$$
$$a(m) = t(m) - y(m)$$

- The weights and biases of the network are adapted as

$$w(m)_{new} = w(m) + l\{a(m) * x(m)\}$$
$$b(m)_{new} = b(m) + l * \{a(m)\}$$

All parameters are set experimentally for proper adaptation of the network. The weights and biases are adjusted to attain the desired signal. The enhanced signal is achieved as the error signal by the ADALINE.

III. Speech Enhancement using Deep Neural Networks

DNN is based on the supervised learning and determines the mapping from the noisy features to clean features. The structure of the DNN is presented in Fig.2 and is divided into 2 phases: training phase and testing phase. The hidden layers employ as the activation function.

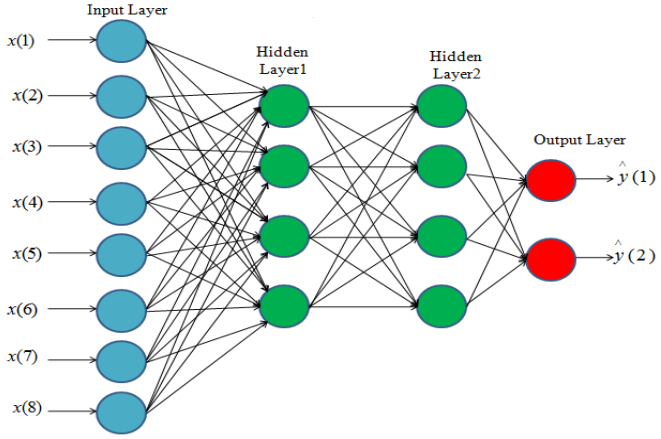

Fig.2. Structure of Deep Neural Network

In this work, two hidden layers are considered and sigmoid function is considered as the activation function for the output. To learn the DNN of noisy log spectra, the multiple restricted Boltzmann machines (RBMs) are arranged [15]. The NOIZEOUS database is taken from the softcopy of Loizou. Babble Noise, Train Noise, Airport Noise and Restaurant Noise of SNR 0dB, 5dB, 10dB and 15dB are considered for training and Drilling Noise, Street Noise are considered for testing. Total 100 speech samples of noisy as well as clean features are acquired in the training phase. Two hidden layers are considered with 512 hidden units each and 8 output units. Total 512*2=1024 hidden units are trained for noisy speech features.

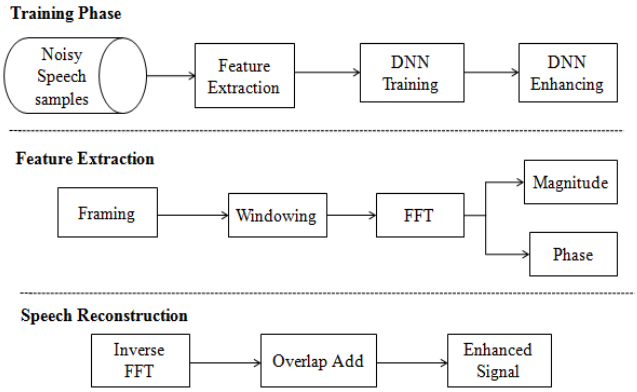

Fig.3. Speech Enhancement using Deep Neural Network

For speech enhancement, the noisy sentence is divided into overlapping frames. Hamming window of length 512 is multiplied to the framed signal to avoid signal distortion. The proposed DNN based speech enhancement method is represented in Fig.3. To extract the magnitude and phase spectra, Fourier Transform is employed. Only the magnitude

spectra are considered for training the noisy features in DNN and the phase spectra are ignored. After training, the Ideal mask is estimated for testing. The mask is enforced to the FT feature vectors of the noisy speech signal. The output of the DNN is interpreted as the predicted mask for the input. All the feature vectors are added and the overlapping frames are concatenated to reconstruct the speech signal. Subsequently, all frames are synthesized into a time domain signal by Overlap add. The following steps are followed for speech enhancement using DNN.

- Process all speech signals as overlapped frames.
- The frame length is 512 with an overlap of 40%.
- Calculate FFT of each and every overlapped windowed frame.
- Four different types of noise signals: Babble, Train, Airport and Restaurant are taken with SNRs 0dB, 5dB, 10dB and 15 dB and the clean signals are considered for training.
- To train the DNN, 156 dimensions of magnitude spectrum of the noisy signals are employed.
- For testing, two noise signals: Drilling and Street are considered to estimate the mismatch condition.
- Inverse FFT and Overlap Add method are implemented to reconstruct the speech signal.

Network pretraining, regularization are employed to make the system better.

IV. Experimental Results

In this work, 10 sentences from both male and female speakers are considered for training. A total of 100 sentences have been collected with different types of noise signals as mentioned earlier. All utterances are sampled with a sampling rate of 8 KHz. Fig.4 shows the spectrogram of the clean speech signal. Fig.5 is the noisy signal (Babble noise of SNR 10dB). The clean signal is considered as the target signal. The noisy signal is applied to the ADALINE and DNN. The enhanced signal of the ADALINE obtained as the error signal as shown in Fig.6.

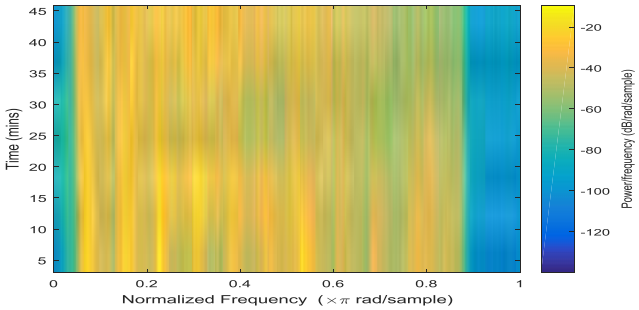

Fig.4. Clean Signal collected from database

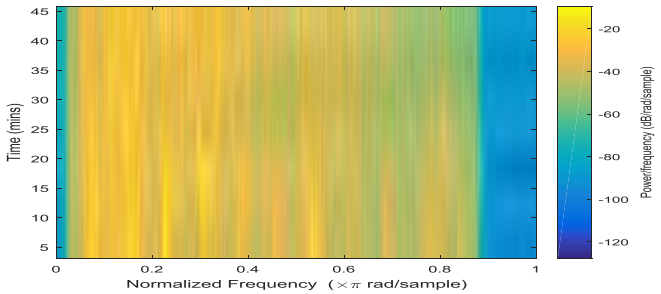

Fig.5. Noisy Signal (Babble Noise of SNR 5dB)

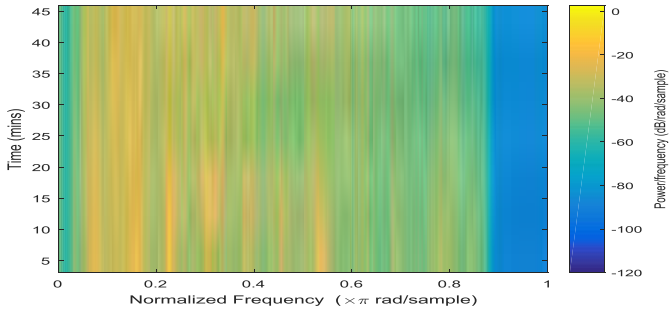

Fig.6. Enhanced signal of ADALINE

All 1024 utterances are considered to build the training set. One sample of clean signal and all the noisy data sets are employed to train the DNN model. A total 2400 frames are measured to train the DNN model. Another 200 arbitrarily selected clean and noisy utterances from the database are considered for testing phase for each combination of noise levels. These signals are estimated and evaluated for mismatch conditions of testing phase. Two hidden layers are considered with 36 frames expansion. Total 1024 hidden units are there in each hidden layer. For pretraining the DNN, the learning rate is 0.05 and for fine tuning, the learning rate is 0.001. 80 epochs are considered for mini batch size of 100. The enhancement using DNN_1 (DNN_* represents the hidden layer number) and DNN_2 are shown in Fig.7 and Fig.8 respectively.

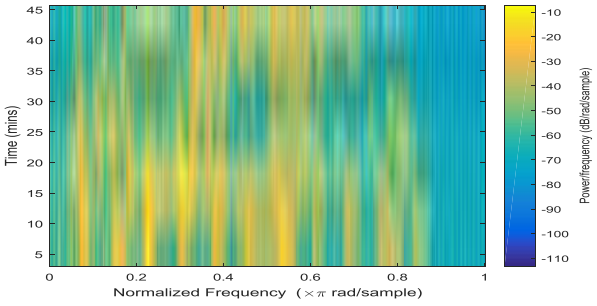

Fig.7. Enhanced signal using DNN_1

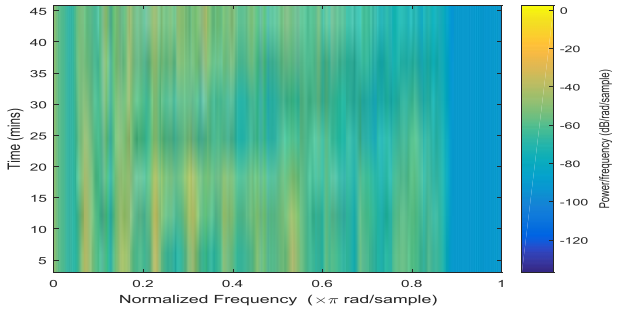

Fig.8. Enhanced signal using DNN_2

In this work, Perceptual Evaluation of Speech Quality (PESQ) and Signal-to-Noise-Ratio (SNR) are measured and evaluated for the quality of the speech signal. The listening test is also performed by different persons to verify the test results. Table I shows the SNR as well as the improvement of SNR of ADALINE and DNN. The maximum SNR improvement is 2.87 dB achieved in DNN_2. Table II shows the PESQ measures of the different noise levels. DNN_2 provides a maximum PESQ of 3.67 for 15dB of Babble noise. When the number of hidden layer increases, the better enhanced signal is obtained in the DNN.

TABLE I
SNR IMPROVEMENT OF BABBLE NOISE WITH DIFFERENT NOISE LEVELS OF DIFFERENT METHODS

SNR before Enhancement (dB)		0 dB	5 dB	10 dB	15 dB
SNR after Enhancement (dB)	ADALINE	1.32	6.36	11.12	15.92
Improvement in SNR (dB)		1.32	6.36	11.12	15.92
SNR after Enhancement (dB)	DNN_1	2.43	7.02	11.98	16.26
Improvement in SNR (dB)		2.43	2.02	1.98	1.26
SNR after Enhancement (dB)	DNN_2	2.87	7.83	12.55	17.31
Improvement in SNR (dB)		2.83	2.87	2.55	2.31

TABLE III
PESQ SCORE OF BABBLE NOISE WITH DIFFERENT NOISE LEVELS OF DIFFERENT METHODS

	ADALINE	DNN_1	DNN_2
0 dB	1.23	2.34	3.11
5 dB	1.45	2.06	3.46
10 dB	1.89	2.51	2.98
15 dB	1.93	1.78	3.67

V. CONCLUSION

ADALINE and DNN are used to enhance the noisy speech signal in this work. ADALINE is considered as the basic Neural Network implemented for speech enhancement. The DNN is used for different hidden layers that can prove the validity of speech enhancement in the field of data mining. The better performance result is obtained using ADALINE model, whereas the DNN model outperforms the ADALINE. Though the time consumption is more in DNN, speech enhancement is better. In the future, the weights of the ADALINE model can be varied and other transforms can be applied to extract the features and observe the performance.

REFERENCES

[1] P. Loizou, Speech Enhancement: Theory and Practice. CRC Press, 2007.

[2] S.Haykin, Adaptive Filter Theory, Prentice Hall, Upper Saddle River, 3rd Edition, 1996.

[3] R.Ram, M.N.Mohanty, Performance Analysis of Adaptive Algorithms for Speech Enhancement Applications, Indian Journal of Science and Technology 9(44), 2016.

[4] S.Vihari, A.S.Murthy, P.Soni, D.C.Naik, Comparison of Speech Enhancement Algorithms, Procedia Computer Science 89, pp. 666 – 676, 2016.

[5] L.B.Fah, A.Hussain & S.A.Samad, Speech Enhancement by Noise Cancellation Using Neural Network, IEEE Conf., 2000.

[6] R.Ram, M.N.Mohanty, Fractional DCT ADALINE Method for Speech Enhancement, Int. Conf. on Machine Learning & Computational Intelligence, 2017. (Accepted)

[7] A.Prieto, B.Prieto, E.M.Ortigosa, E.Ros, F. Pelayo, J.Ortega, I. Rojas, Neural Networks: An Overview of Early Research, Current Frameworks and New Challenges, Neurocomputing 214, pp.242–268, 2016.

[8] T. Kounovsky and J. Malek, Single Channel Speech Enhancement Using Convolutional Neural Network, *IEEE* International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM),2017, pp.1-5.

[9] M.Kolbaek, Z.H.Tan, J.Jensen, Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(1), 2017.

[10] Y.Xu, J.Du, L.R.Dai, and C.H.Lee, An Experimental Study on Speech Enhancement Based on Deep Neural Networks, IEEE Signal Processing Letters 21(1), 2014.

[11] Y.Xu, J. Du, L.R.Dai, C.H.Lee, A Regression Approach to Speech Enhancement Based on Deep Neural Networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(1), 2015.

[12] R.Li, Y.Liu, Y.Shi, L.Dong, W.Cui, ILMSAF based Speech Enhancement with DNN and Noise Classification, Speech Communication 85, pp.53–70, 2016.

[13] Y.Li, S.Kang, Deep Neural Network-Based Linear Predictive Parameter Estimations for Speech Enhancement, IET Signal Process.11 (4), pp.469-476, 2017.

[14] T.Goehring, F.Bolner, J.Monaghan, B.Dijk, A.Zarowski, S.Bleeck, Speech Enhancement Based on Neural Networks Improves Speech Intelligibility in Noise for Cochlear Implant Users, Hearing Research 344, pp.183-194, 2017.

[15] Y.Koizumi, K.Niwa, Y.Hioka, K.Kobayashi and Y.Haneda, DNN-Based Source Enhancement Self-Optimized By

Reinforcement Learning Using Sound Quality Measurements, IEEE Conf., 2017.

[16] R. Ram, M. N.Mohanty, Deep Neural Network based Speech Enhancement. Int. Conf. On Cognitive Informatics & Soft Computing, 2017. (Accepted)

[17] J. C. Hou, S. S. Wang, Y. H. Lai, J. C. Lin, Y Tsao, H. W. Chang, H. M. Wang, Audio-Visual Speech Enhancement using Deep Neural Networks, Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, IEEE, 2016.