ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 6, June 2017

A Review on Neural Network based Noise Robust Speech Recognition Methods

Kamlesh Kaur[1], Dr. Pooja[2], Dr. Pankaj Mohindru[3]

M.Tech Student, Department of ECE, UCOE, Punjabi University, Patiala, Punjab, India[1]

Assistant Professor, Department of ECE, UCOE, Punjabi University, Patiala, Punjab, India[2]

Assistant Professor, Department of ECE, UCOE, Punjabi University, Patiala, Punjab, India[3]

ABSTRACT: The neural network classifier is being used for the various applications now-a-days for the purpose of data classification and pattern recognition. The neural network algorithm offers deep learning for the large scale databases, where it becomes very difficult otherwise to find the matching samples. In this paper, the study of the existing models for the noise robust speech recognition models has been conducted for the performance analysis, which gives the critical overview of all of the analyzed models. The noise robust speech recognition models are utilized to tackle the real-time speech environments for the handling of the noisy data for the purpose of speech or speaker recognition. In this paper, the noise robust speech recognition method has been proposed, which combines the mel-frequency cepstral coefficient (MFCC) with filter bank method based upon the discrete wavelet transform (DWT) and deep neural network classification. The proposed model is expected to improve the overall performance of the new noise robust speech recognition system in comparison with the existing models.

KEYWORDS: Noise Robust Speech Recognition, Neural Network, Filter bank, Discrete wavelet transform (DWT)

I. INTRODUCTION

A. *Speech Recognition*

Speech is the most efficient way of exchanging information and expressing ideas between two human beings. Speech is a natural way of communication because it requires no special training as most of the humans are born with this instinct. It is considered as the most flexible, economical way of conveying information. Life would be more comfortable, if speech is used for Human Machine Interface (HCI). The other interfaces like mouse, keyboard, joystick and touch pad requires some amount of expertise in using them. Therefore, physically challenged people find it difficult to interact with computers or machines [1]. In ASR, Speech is given as input to the recognizer which transform acoustic signal into features form. Then it responds appropriately using the features. This will either generate a transcript or will perform some control action. The transcription is generated with the help of acoustic model and language model. We can retrieve a lot of information from speech signal regarding the gender, age, accent, identity of speaker, emotion and health of speaker. Speech recognition systems are divided into different categories viz. Isolated Word, Connected Word, Continuous and Spontaneous Speech Recognition, Speaker Dependent, Speaker Independent models. Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding Coefficient (LPCC) and Probabilistic Linear Discriminate Analysis (PLDA) are some of the feature Extraction techniques [2].

B. *Neural Nework*

Artificial Neural Networks are mathematical models that mimic the behavior of neurobiological networks. ANN consist set of highly interconnected processing elements called artificial neurons. All these neurons work in parallel to solve a particular problem. ANN are generally adaptive in nature because there occurs a change in structure of network whenever information is passed during learning phase. The collective behavior of all neurons makes ANN suitable for

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

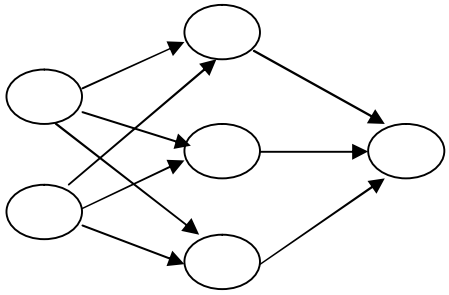
International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 6, June 2017

pattern recognition and pattern classification tasks. Fault tolerance, generalization, trainability, robustness, uniformity are some of the advantages of ANNs [3].



Input Layer Hidden Layer Output Layer

Figure 1. Neural Network

The neurons in ANN are arranged in layer structure. Some researchers worked on randomly connected neurons, but not much success was achieved. Layers are grouping of neurons. There are three types of layers – input layer, one or more hidden layers and output layer. Neurons of one layer are connected to the neurons of other layer. But there is no interconnection between neurons of a single layer. Initially weights are chosen randomly. Then training or learning phase begins. There are two approaches used for training – Supervised and Unsupervised Learning. In Supervised learning, network is provided with inputs and desired outputs. After processing inputs, observed outputs are compared with desired outputs. Error is calculated and propagated back to the input side. This causes the system to change its weights. In unsupervised learning, network is not provided with desired output. Network has to decide on its own how to group the input data. It is also called self organization [4].

II. CASE STUDY

In the project work [5], a general phoneme expert system is implemented that can learn to characterize and represent phonemes. As the procedure of implementation is quite complex, so full scale implementation is not possible. Speech waveform consists of several phonemes units. Memory is required for storing the patterns associated with these phonemes units so that Neural Network can predict the output not only from the current input but also from the previous inputs fed into the network.

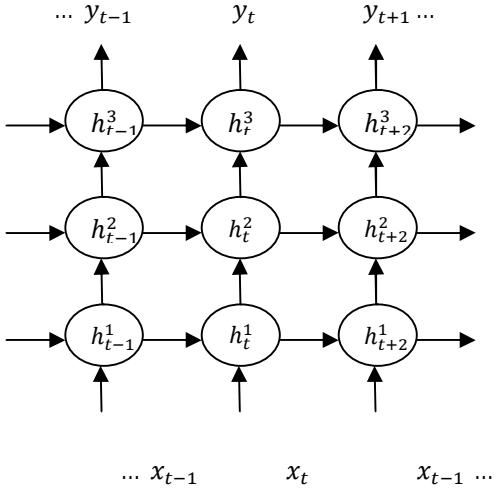


Figure 2. Deep Recurrent Neural Network

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 6, June 2017

Thus a recurrent neural network (RNN) is employed in the project work to achieve required memory depth. RNN have feedback connections that allow certain layer's output to depend on the past outputs. Here, outputs of hidden neurons are first weighted then again given as input to the hidden layer neurons. This adds memory to the network. Elman type RNNs are used in the work. Elman networks are three layer perceptrons where connections are from hidden layer to context layer also called state layer. These networks are trained using back propagation algorithm. During 1 time frame of a speech waveform 15 Mel Frequency Cepstral Coefficients are used as an input to the RNN. One time frame corresponds to the one column of phoneme vector. For implementing, Gradient Descent Backpropagation algorithm in MATLAB traingdx activation function is used for all neurons. Both learning rate and momentum are adaptable. For testing, different structures were used with at least 1 hidden layer and 1 output layer. Output layer has only 1 neuron. In Speech Recognition implementation, mixture of expert systems is used with no gating mechanism. After training, each expert would be able to recognize a specific phoneme. At any time t, outputs of all experts are combined to mark out the most probable phonemes. Hypothesis is carried out stochastically for finding a nearest matching word which turns out as the final output.

III. RELATED WORK

In [6], Qian, Yanmin et. al. has extended former architecture of Very Deep Convolutional Neural Networks (CNNs) for robust speech recognition. The authors are inspired by the results of very deep CNNs in the field of computer vision, where image classification has improved to great extent by growing the number of convolutional layers in the conventional CNNs. The dimensions of filters and pooling are reduced and size of input features are extended so that more number of convolutional layers can be added in the architecture. Different pooling and padding strategies are studied to make the system capable of de-noising and de-reverberation. Results are evaluated on Aurora-4 and AMI meeting transcription. Best configuration is achieved at 10 convolutional layers. In the results, it is concluded that input feature maps padded on both sides deliver best results. Moreover, it is seen that very deep CNNs with static features perform better than traditional dynamic features. The proposed system of very deep CNN shows improved word error rate relative to LSTM-RNN acoustic models. Also, the model has compact size and the training convergence speed of network is also very fast.

In [7], Xu, Yong et. al. has developed a supervised method for enhancing speech by means of a deep neural network based function. The function provides a mapping from noisy speech signals to clean speech signals. A huge dataset is first designed which consists wide range of noises pertaining to real world situations. The realized model is highly capable of suppressing the non- stationary noise. The model can effectively work in real world environments where speech data is intensely contaminated by noisy signals. For implementing a highly non linear regression function, many stages of non linearities are added in the feed-forward neural network. Furthermore, the MMSE optimized DNNs suffer from the problem of over smoothing, which is resolved in this paper using a technique called Global Variance (GV) Equalization. Moreover, dropout training is adopted to make the system capable for dealing with unseen noises.

Weninger, Felix et. al. has worked upon the enhancement of speech frame for the noise-robust speech recognition applications [8]. The author presented a model based on LSTM which is trained discriminatively. The model has given very efficient results with CHiME corpus when used in front end processing for speech separation. Furthermore, some feature based adjustments are also proposed to integrate the proposed model with ASR back end. RNNs could deliver better results than DNNs if the vanishing temporal gradient problem of RNN would be eliminated or reduced. So in the work, LSTM activation function is used instead of the conventional sigmoid activation function which helps in solving the temporal gradient problem. At front end, speech separation is evaluated in Signal to Distortion Ration (SDR) on two channel system. At the back end, speech separation is measured in Word Error Rate (WER). The final result shows a high correlation between the SDR and WER which is in contradiction to the previous studies.

In [9], a robust speech recognition system is developed using Long Short-Term Memory Recurrent Neural Networks (RNNs) as an acoustic model. Here the author has deployed Long Short-Term Memory RNNS because these networks take the advantage of self learnt amount of temporal context. But the LSTM RNNs approach requires GMM acoustic model in the multi-stream framework. Moreover, there is loss of modeling power of network during predication of phonemes. The proposed model in this paper is capable of overcoming these two drawbacks. Work is carried out on 2nd CHiME Speech Separation and Recognition challenge. LSTM RNN is used in combination with NN/HMM. The LSTM are used in bidirectional mode. The HMM states are provided as training targets to the network then the output

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 6, June 2017

predictions are converted into state likelihoods. Further, these state likelihoods are used in hybrid setup for decoding purpose. Experimental results show that LSTM with hybrid acoustic model perform better than the model in which LSTM are used for predicting phonemes and the phonemes are further converted to state likelihoods in GMM-LSTM double stream setup.

In [10] Weng, Chao et. al., Weng, Chao et. al. has worked towards the development of deep neural network for speech recognition. In the work, conventional feedforward deep neural networks are modified by adding full recurrent connection to one of the hidden layers of architecture. The layer has recurrent connection with itself. In addition, Back Propagation through Time (BPTT) algorithm is used to update the weights of recurrent layers. This algorithm is a little modified to make application of SGD frame by frame possible. The minibatch SGD results in reduction of sizes of matrices and it becomes easy to store them on GPU. Minibatch SGD makes the proposed model more efficient and effective. Experimental results were evaluated on two databases CHiME and Aurora-4. The results DNN-BPTT model achieves state of the art performance without any model adaptation, multiple passes.

As the RNN performance in the field of speech recognition was not very pleasing, therefore in 2013 Graves, Alex et. al. [11] proposed a model which augmented deep Long Short-term Memory RNNs with an end to end training method called Connectionist Temporal Classification. This hybrid approach has provided very prolific results in cursive handwriting recognition. When the alignment of input output is not known then RNNs trained with CTC can be used for sequence labeling. The depth added in LSTM by author is inspired by the past results where more number of feed forward layers in convolutional neural networks returned better results. Thus, several recurrent hidden layers are framed up on each other in deep Long Short-term Memory architecture of RNN. The end to end training was modified so that RNNs can learn direct mapping of acoustic sequences to phonetic sequences. Work was carried out on TIMIT database using nine RNNs. All RNNS were trained using Stochastic Gradient Descent (SGD). The depth introduced in RNNs dropped the CTC error rate from 23.9% to 18.4%. It is concluded that bidirectional LSTMs are more advantageous than unidirectional LSTMs.

Paper Details	Title	Technique Proposed	Merits	Demerits
Qian Yanmin, Mengxiao Bi, Tian Tan, and Kai Yu [6], IEEE/ACM TASLP, 2016	Very deep convolutional neural networks for noise robust speech recognition	More number of convolutional layers are added to tradition Covolutional Neural Networks by varying sizes of filter, pooling layers	1. This technique has reduced the overall error to 10% in comparison with traditional CNN models on AMI database. 2. 17% WER improvement has been recorded this model, which has been reduced to nearly 8%.	1. No blank detection and removal method has been utilized, for example: MFCC. 2. Hierarchical feature description model, combining MFCC with DWT (Filter Bank) for higher accuracy and reduced WER.
Xu Yong, Jun Du, Li-Rong Dai, Chin-Hui Lee [7], IEEE/ACM TASLP, 2015	A regression approach to speech enhancement based on deep neural networks	DNN based function is purposed for mapping noisy signals to clean signals	1. Efficiently handles the noisy speech data. 2. Does not generate the annoying artifacts (specifically	1. Dense training data can further improve the overall results of this model. 2. Gammatone filter bank or Cochleagram

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 6, June 2017

			musical artifacts)	feature based filter bank method can further improve the accuracy
Weninger Felix, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux et.al. [8], International Conference on Latent Variable Analysis and Signal Separation, 2015	Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR	Long Short-term Memory with RNN for the automatic speech recognition.	1. WER has been recorded at 13.76%, which shows adequate performance. 2. Raw signal decomposition-based processing is learnt to remove the inherent redundancy.	1. Frequency component analysis can be incorporated in the form of power spectrum or other frequency oriented feature for higher accuracy. 2. Harmonic echo separation can be incorporated to reduce the complexity.
Geige Jürgen T., Zixing Zhang, Felix Weninger, Björn Schuller, Gerhard Rigoll [9], International Speech Communication Association, 2014	Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modeling	LSTM-RNN with HMM for speech utterance recognition	1. Performs better than the models based upon phonemes 2. Hybrid double layered model incorporates state prediction model.	1. Generative pre-training with detailed and versatile data can further improve the accuracy 2. DNN with LSTM can further improve the results instead of LSTM-RNN.
Weng Chao, Dong Yu, Shinji Watanabe, Biing Hwang Fred Juang [10], IEEE International Conference on Acoustics, Speech and Signal Processing, 2014	Recurrent deep neural networks for robust speech recognition	In Deep Recurrent Neural Network fully connected layers are added to some of the hidden layers and network is trained using modified Back Propagation through time algorithm	1. Shows adequate accuracy to be determined as “state-of-art” system	1. Use of back-propagation makes it more training data dependent. And make it incapable of recognizing the samples with new variations

Table I. Literature Table

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

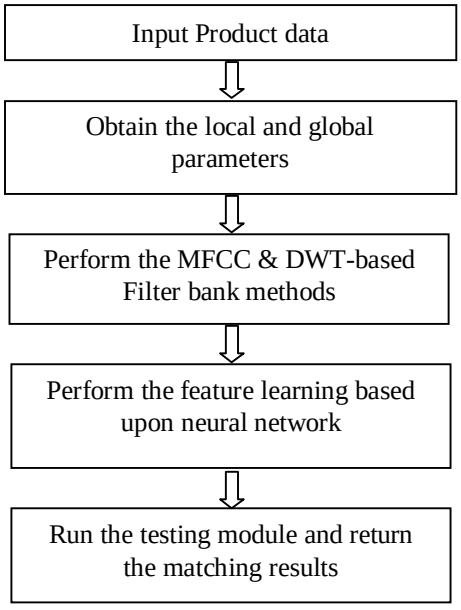
Vol. 6, Issue 6, June 2017

IV. FINDINGS OF LITERATURE SURVEY

The existing model is based upon the filter bank model, which extracts the target features in the multiple frequencies. The tri-frequency feature defines the various aspects of the target sample, which are further utilized to find the words in the given speech data. These frequency components are resulted by the filter bank method based upon the frequency based feature description method, which defines the angular and directional features of the speech sample, which matches the samples with higher accuracy. Further, the deep neural network has been utilized in the existing model to discover the matching samples. The very deep convolution neural network has been utilized for the purpose of speech sample classification, which utilizes the frequency oriented features. The existing model does not incorporate the speech region localization, which accounts the blank regions and eliminate them from the feature bank, which is extracted by using the filter bank in the case of existing model. The feature extraction in existing model does not apply the mel-frequency cepstral coefficient (MFCC) over the speech signal to acquire the speech part and to eliminate the blank regions out of the final feature, which is expected to improve the overall classification accuracy.

V. METHODOLOGY

In research, the noise robust speech recognition algorithm has been used to classify the speech samples based upon the various features, which includes the hybrid feature descriptive method by combining the mel-frequency cepstral coefficient (MFCC) and DWT-based filter bank method for the speech feature preparations. The enlisted entity in the speech recognition paradigm undergoes the deep analytical feature analysis algorithm on the basis of the various speech feature and factors. The ASR algorithm evaluates the physiological and frequency based factors to decide the position of the given object in the listings. The following flowchart explains the things with better elaboration.



We have implemented the new recommendation system which is designed for the digital libraries. In this research work, we have used MATLAB programming for the handling of the data in the local AROURA database. MATLAB model has been developed in the three phases:

1. Acquire the database
2. Hybrid Feature description
3. Learning and testing module

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 6, June 2017

VI. CONCLUSION

In this paper, the various aspects of the neural networks, speech recognition, feature descriptors and various other schematic methods have been studied, which eventually holds the vital information for the deep understanding of speech classification model architectures. The speech recognition model using the various neural network models are found highly accurate them than speech recognition models with other probabilistic classification models. Hence, the combination of the very deep convolution neural network has been proposed along with the MFCC with DWT-based filter bank method for the realization of the noise robust speech recognition models. The proposed model is expected to resolve the issue related to the précised feature description to attain the highly accurate noise robust speech recognition environment.

REFERENCES

[1] Vimala. C, Dr. V.Radha, "A Review on Speech Recognition Challenges and Approaches," World of Computer Science and Information Technology, vol.2, no.1, pp. 1-7, 2012.
[2] Sanjay A. Valaki, Harikrishan B. Jethava, "A Survey on Feature Extraction and Classification Techniques for Speech Recognition," International Journal of Advance Research and Innovative Ideas in Education, vol.2, no. 6, pp. 830-837, 2016.
[3] Dhavale Dhanashri, S.B. Dhonde, "Speech Recognition using Neural Network : A Review," International Journal of Multidisciplinary Research and Development, vol.2, no. 6, pp. 226-229, 2015.
[4] Ms. Sonali. B. Mind, Ms. Priyanka Wankar, "Research Paper on Basic of Artificial Neural Network." International Journal on Recent and Innovation Trends in Computing and Communcation, vol.2, no.1, pp. 96-100, 2014.
[5] Graves, Alex, Navdeep Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," International Conference on Machine Learning, vol. 32, pp. 1764-1772, 2014.
[6] Qian Yanmin, Mengxiao Bi, Tian Tan, and Kai Yu. "Very deep convolutional neural networks for noise robust speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2263-2276, 2016.
[7] Xu Yong, Jun Du, Li-Rong Dai, Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech and Language Processing , vol. 23, no. 1, pp. 7-19, 2015.
[8] Weninger Felix, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux et.al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," International Conference on Latent Variable Analysis and Signal Separation, pp. 91-99, Springer International Publishing, 2015.
[9] Geiger Jürgen T., Zixing Zhang, Felix Weninger, Björn Schuller, Gerhard Rigoll, "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modeling," International Speech Communication Association, pp. 631-635, 2014.
[10] Weng Chao, Dong Yu, Shinji Watanabe, Biing Hwang Fred Juang, "Recurrent deep neural networks for robust speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5532-5536, 2014.
[11] Graves Alex, Abdel-Rahman Mohamed, Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," IEEE International Conference on Acoustics, Speech and Signal processing, pp. 6645-6649, 2013.