Jupiter Publications Consortium

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

AN OVERVIEW OF NEURAL NETWORK BASED NOISE ROBUST SPEECH RECOGNITION METHODS USING DWT

Dr. S.Tamil
Department of Electronics and Communication Engineering,Shadan College of Engineering and Technology HYD,T.S,INDIA
Received 15, October 2017 | Accepted 24, November 2017

ABSTRACT

The neural community classifier is being used for the quite a number purposes now-a-days for the purpose of records classification and pattern recognition. The neural community algorithm offers deep learning for the massive scale databases, the place it will become very tough otherwise to locate the matching samples. In this paper, the study of the present models for the noise robust speech attention models has been conducted for the overall performance analysis, which offers the indispensable overview of all of the analyzed models. The noise strong speech consciousness models are utilized to tackle the real-time speech environments for the handling of the noisy records for the purpose of speech or speaker recognition. In this paper, the noise strong speech recognition method has been proposed, which combines the melfrequency cepstral coefficient (MFCC) with filter financial institution technique based upon the discrete wavelet seriously change (DWT) and deep neural network classification. The proposed mannequin is expected to improve the average performance of the new noise robust speech consciousness machine in comparison with the existing models.

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

I. INTRODUCTION

A. Speech Recognition

Speech is the most environment friendly way of replacing information and expressing ideas between two human beings. Speech is a herbal way of verbal exchange because it requires no one-of-a-kind coaching as most of the humans are born with this instinct. It is regarded as the most flexible, cost effective way of conveying information. Life would be extra comfortable, if speech is used for Human Machine Interface (HCI). The other interfaces like mouse, keyboard, joystick and contact pad requires some quantity of information in the use of them. Therefore, physically challenged humans find it difficult to have interaction with computer systems or machines [1]. In ASR, Speech is given as enter to the recognizer which radically change acoustic sign into aspects form. Then it responds appropriately using the features. This will both 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 sign concerning the gender, age, accent, identification of speaker, emotion and health of speaker. Speech focus systems are divided into exclusive classes viz. Isolated Word, Connected Word, Continuous and Spontaneous Speech Recognition, Speaker Dependent, Speaker Independent models. Mel Frequency Cepstral Coefficient (MFCC),

Jupiter Publications Consortium

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

Linear Predictive Coding Coefficient (LPCC) and Probabilistic Linear Discriminate Analysis (PLDA) are some of the feature Extraction techniques [2].

B. Neural Network

Artificial Neural Networks are mathematical fashions that mimic the conduct of neurobiological networks. ANN consist set of quite interconnected processing elements known as synthetic neurons. All these neurons work in parallel to resolve a precise problem. ANN are typically adaptive in nature because there occurs a alternate in shape of community each time data is passed at some point of studying phase. The collective behavior of all neurons makes ANN suitable for pattern consciousness and pattern classification tasks. Fault tolerance, generalization, trainability, robustness, uniformity are some of the blessings of ANNs [3].

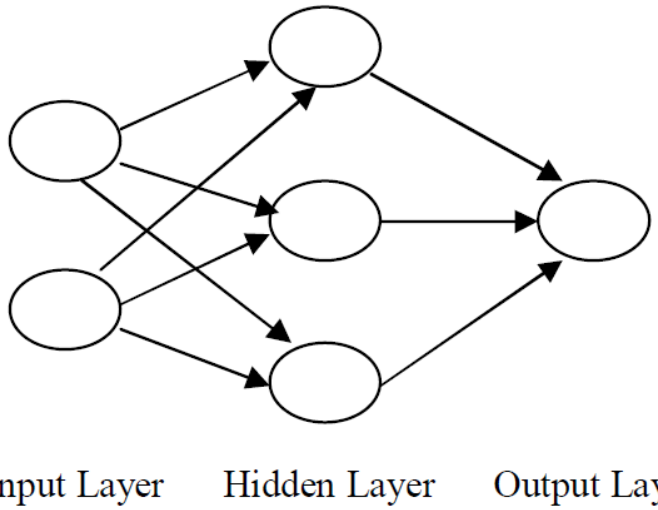


Figure 1. Neural Network

The neurons in ANN are arranged in layer structure. Some researchers labored on randomly connected neurons, but now not a lot 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 coaching or getting to know section begins. There are two strategies used for training – Supervised and Unsupervised Learning. In Supervised learning, network is furnished with inputs and preferred outputs. After processing inputs, located outputs are compared with preferred outputs. Error is calculated and propagated again to the enter side. This causes the device to exchange its weights. In unsupervised learning, network is no longer furnished with desired output. Network has to decide on its own how to crew the enter data. It is also called self employer [4].

II. CASE STUDY

In the undertaking work [5], a regular phoneme professional machine is carried out that can examine to represent and represent phonemes. As the technique of implementation is quite complex, so full scale implementation is now not possible. Speech waveform consists of countless phonemes units. Memory is required for storing the patterns related with these phonemes units so that Neural Network can predict the output not solely from the modern input but additionally from the previous inputs fed into the network.

Jupiter Publications Consortium

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

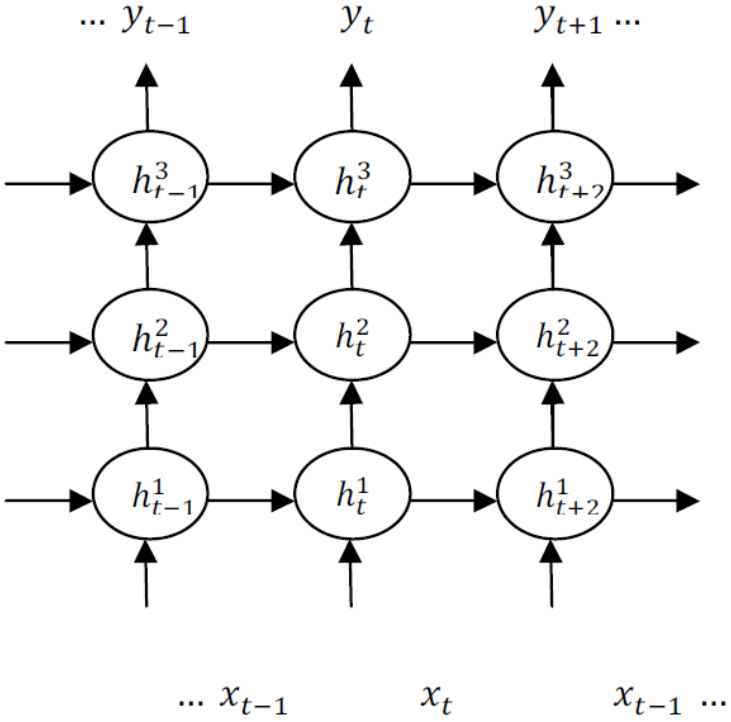


Figure 2. Deep Recurrent Neural Network

Thus a recurrent neural community (RNN) is employed in the project work to acquire required reminiscence depth. RNN have remarks connections that enable positive layer's output to depend on the past outputs. Here, outputs of hidden neurons are first weighted then again given as enter 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 additionally called country layer. These networks are trained the usage of lower 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 body corresponds to the one column of phoneme vector. For implementing, Gradient Descent Backpropagation algorithm in MATLAB traingdx activation feature is used for all neurons. Both learning charge and momentum are adaptable. For testing, special constructions have been 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 specialist would be capable to recognize a specific phoneme. At any time t, outputs of all experts are combined to mark out the most probably phonemes. Hypothesis is carried out stochastically for finding a nearest matching word which turns out as the remaining output.

III. RELATED WORK

In [6], Qian, Yanmin et. al. has extended former structure of Very Deep Convolutional Neural Networks (CNNs) for strong speech recognition. The authors are inspired by using the effects of very deep CNNs in the discipline of pc vision, the place photograph classification has extended to excellent extent via growing the variety of convolutional layers in the traditional CNNs. The dimensions of filters and pooling are decreased and measurement of input features are extended so that more range of convolutional layers can be introduced in the architecture. Different pooling and padding strategies are studied to make the machine capable of de-noising and de-reverberation. Results are evaluated on Aurora-4 and AMI

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

assembly transcription. Best configuration is achieved at 10 convolutional layers. In the results, it is concluded that input characteristic maps padded on each sides deliver excellent results. Moreover, it is viewed that very deep CNNs with static points operate better than traditional dynamic features. The proposed device of very deep CNN shows expanded phrase error price relative to LSTM-RNN acoustic models.

Also, the mannequin has compact dimension and the coaching convergence velocity of network is additionally very fast. In [7], Xu, Yong et. al. has developed a supervised approach for improving speech by way of capacity of a deep neural community based totally function. The feature presents a mapping from noisy speech alerts to clean speech signals. A massive dataset is first designed which consists large range of noises pertaining to actual world situations. The realized mannequin is particularly succesful of suppressing the non-stationary noise. The mannequin can efficiently work in actual world environments where speech statistics is intensely contaminated by noisy signals.

For imposing a enormously non linear regression function, many stages of non linearities are delivered in the feed-forward neural network. Furthermore, the MMSE optimized DNNs suffer from the hassle of over smoothing, which is resolved in this paper the usage of a technique called Global Variance (GV) Equalization. Moreover, dropout coaching is adopted to make the machine succesful for dealing with unseen noises. Weninger, Felix et. al. has labored upon the enhancement of speech body for the noise-robust speech attention functions [8]. The writer a mannequin based totally on LSTM which is skilled discriminatively. The mannequin has given very efficient consequences with CHiME corpus when used in the front give up processing for speech separation.

Furthermore, some feature primarily based adjustments are also proposed to combine the proposed mannequin with ASR lower back end. RNNs could supply higher effects than DNNs if the vanishing temporal gradient trouble of RNN would be eliminated or reduced. So in the work, LSTM activation feature is used rather of the traditional sigmoid activation characteristic which helps in fixing the temporal gradient problem. At the 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 ultimate end result shows a high correlation between the SDR and WER which is in contradiction to the previous studies.

In [9], a sturdy speech cognizance gadget is developed the use of Long Short-Term Memory Recurrent Neural Networks (RNNs) as an acoustic model. Here the writer has deployed Long Short-Term Memory RNNS due to the fact these networks take the benefit of self learnt quantity of temporal context. But the LSTM RNNs approach requires GMM acoustic mannequin in the multi-stream framework. Moreover, there is loss of modeling strength of community at some stage in predication of phonemes. The proposed mannequin in this paper is succesful of overcoming these two drawbacks. Work is carried out on 2nd CHiME Speech Separation and Recognition challenge. LSTM RNN is used in aggregate with NN/HMM. The LSTM are used in bidirectional mode. The HMM states are supplied as training goals to the community then the output 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 mannequin function higher than the mannequin in which LSTM are used for predicting phonemes and the phonemes are similarly transformed to kingdom likelihoods in GMM-LSTM double move setup.

Jupiter Publications Consortium

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

In [10] Weng, Chao et. al., Weng, Chao et. al. has worked toward the development of deep neural network for speech recognition. In the work, conventional feedforward deep neural networks are modified by means of adding full recurrent connection to one of the hidden layers of architecture. The layer has recurrent connection with itself. In addition, Back Propagation via Time (BPTT) algorithm is used to replace the weights of recurrent layers. This algorithm is a little modified to make utility of SGD body by means of frame possible. The minibatch SGD consequences in discount of sizes of matrices and it becomes effortless to shop them on GPU. Minibatch SGD makes the proposed model greater environment friendly and effective. Experimental outcomes were evaluated on two databases CHiME and Aurora-4. The results DNN-BPTT mannequin achieves state of the artwork performance besides any model adaptation, a couple of passes. As the RNN performance in the field of speech awareness was not very pleasing, therefore in 2013 Graves, Alex et. al. [11] proposed a mannequin which augmented deep Long Short-term Memory RNNs with an quit to stop coaching approach called Connectionist Temporal Classification.

This hybrid approach has provided very prolific outcomes in cursive handwriting recognition. When the alignment of input output is now not recognized then RNNs skilled with CTC can be used for sequence labeling. The depth introduced in LSTM by means of creator is inspired through the past results where more wide variety of feed forward layers in convolutional neural networks again better results. Thus, numerous recurrent hidden layers are framed up on every different in deep Long Short-term Memory architecture of RNN. The end to cease education used to be modified so that RNNs can research direct mapping of acoustic sequences to phonetic sequences. Work was once carried out on TIMIT database using nine RNNs. All RNNS have been educated using Stochastic Gradient Descent (SGD). The depth added in RNNs dropped the CTC error fee from 23.9% to 18.4%. It is concluded that bidirectional LSTMs are greater superb 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

Jupiter Publications Consortium

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

			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

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

Jupiter Publications Consortium

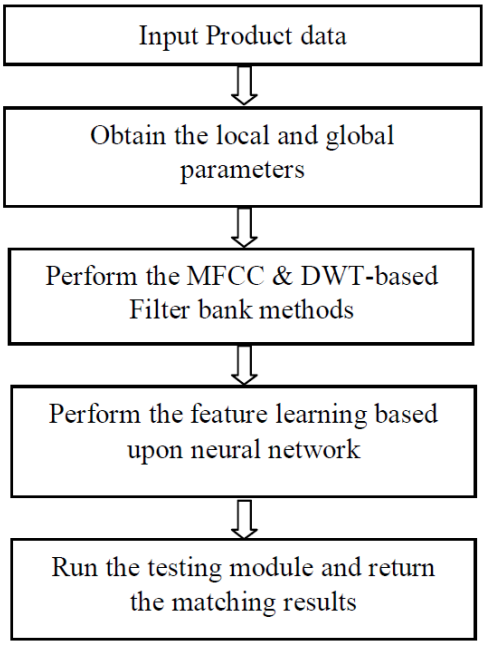
Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

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

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

Vol. 3 Iss.2, pp. 38-45, 24th NOV, 2017
ISSN Print : 2454-1435 © 2017, IJRMMAE
ISSN Online : 2454-1443 © 2017, IJRMMAE
http://www.ijrmmae.in

International Journal of Research in Mechanical, Mechatronics and Automobile Engineering (IJRMMAE)

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] Radha, V., and C. Vimala. "A review on speech recognition challenges and approaches." doaj. org 2.1 (2012): 1-7.
[2] Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." International Conference on Machine Learning. 2016.
[3] Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.
[4] Maind, Sonali B., and Priyanka Wankar. "Research paper on basic of artificial neural network." International Journal on Recent and Innovation Trends in Computing and Communication 2.1 (2014): 96-100.
[5] Graves, Alex, and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." International Conference on Machine Learning. 2014.
[6] Qian, Yanmin, et al. "Very deep convolutional neural networks for noise robust speech recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.12 (2016): 2263-2276.
[7] Xu, Yong, et al. "A regression approach to speech enhancement based on deep neural networks." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23.1 (2015): 7-19.
[8] Weninger, Felix, 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. Springer, Cham, 2015.
[9] Geiger Jürgen T., Zixing Zhang, Felix Weninger, Björn Schuller, Gerhard Rigoll, "Robust speech recognition using long short-term memory
[10] Recurrent neural networks for hybrid acoustic modeling," International Speech Communication Association, pp. 631-635, 2014.
[11] Weng, Chao, et al. "Recurrent deep neural networks for robust speech recognition." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.
[12] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013.