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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

VDCNN based Noise Robust Speech Recognition with Combination of GMM and MFCC Features

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 speech recognition is the mechanism to interpret the meaningful description of the input speech signal, which can be used for further processing. The speech recognition models are utilized for speech-to-text, voice operated gadgets, artificial intelligence applications such as Cortana, Siri, etc. The speech recognition applications require the robust feature description along with noise removal in the given speech signal to create the interpretable data, which can be utilized for further speech recognition application. The noise robust classifications play the important role for speech recognition applications over the noise robust features, which are described to eliminate the blank speech segments, surrounding noise, stutters, etc. In this paper, the very deep convolutional neural network (VDCNN) has been proposed along with the Gaussian mixture model (GMM) and Mel-frequency cepstral coefficient (MFCC) over the aurora-4 dataset. The dataset is consisted of clean and noised speech signals, which are used in the combination to test the performance of the proposed model. The proposed model has undergone various experiments, which includes the various number of samples (200, 400, 2000, 4000 and 8000 samples) and different number of layers (4, 5 and 6). The feature descriptor of 20x3 sized matrix has been utilized for the purpose of classification, which is derived with the combination of GMM and MFCC over each of the speech samples. The word error rate has been lowered to nearly 2.24 over the variant with 8000 samples, 3.12 with 4000 samples and 8.864 with 2000 samples. This clearly depicts the impact of number of samples upon the word error rate, and higher word error rate is found efficient in comparison to the other variants.

KEYWORDS: Deep Learning, Very deep convolutional neural network (VDCNN), GMM, MFCC, Noise robust speech recognition.

I. INTRODUCTION

The speech is the most common and efficient form of exchanging information among human being as it is the skill that most of the people already have. Automatic Speech Recognition (ASR) has been an active research area for over five decades. It has always been considered critical for encouraging better human-human and human-machine interaction. But in past, ASR had not gained very much popularity since innovation at that time was not good enough to use it under real time sceneoris. Additionally, advancements like keyboard and mouse interfaces performed superior to speech in terms of efficiency and restrictions. In recent years, with the rapid development of hardware, software and cloud computing, speech technology advanced with much pace. Now we use this as our primary mean to interact with some of our gadgets. The greater part of this success is accomplished by high computational power of multi-core processors that are available today. These can train complex models of ASR through large amounts of training data. Speech Recognition causes the disabled persons to interact with machine through their voice [1]. It is centered on probing human voice as the research entity. It let the machines to automatically get the message concealed in spoken language by facilitating with speech signal processing and pattern recognition techniques. Dependability of ASR

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

structures lean on factors like number of words ASR system is required to perceive, whether it is speaker dependent or independent and with which sort of speech it will deal. Speech recognition engines take audio as the input. But these audio streams could never be in their original forms. The information exchanged is poorly tainted with background noise because of disturbances in recording environments. This noise intrudes with the recognition process and result in degraded yield. Thus, Speech engine must be able to handle then environment in which the audio was spoken [2]. First, a data base of voices is created by recording speech utterances through the microphone. The sound card of the computers mutates them to digital forms. In next step, for providing a compact representation to the digitize signal and for making the identity of sound clear, feature extraction is performed. In Feature extraction, the relevant information which can segregate between different sounds is preserved while the trivial one is discarded. The next step is of pattern recognition. In which, first the system is trained with reference samples then it is tested with unknown sample by taking into consideration the reference samples [3].

II. RELATED WORK

Majority of the speech recognition systems require priori knowledge of twists and conditions in which it needs to work. But in real world environment, we are uninformed about the conceivable conditions. Therefore to cope with this problem, in [4], Vikramjit Mitra et al. investigate impact of feature-space maximum likelihood linear regression (fMLLR) transform and deep convolutional neural networks on automatic speech recognition in real time conditions. They used feature set of gammatone filter bank features (GFBs) and normalized modulation coefficients (NMCs) along with feature set consisting mel filter bank energies (MFBs) and mel frequency cepstral coefficients (MFCCs). GFBs better approximates the sound related conduct and NMCs are good at tracking the directions of modulated speech. These features are then used for training deep autoencoder (DAE), yield of which is further utilized for training fully connected deep neural network. In second model, they replaced DNNs with time frequency convolutional neural networks (TFCNN). The execution is measured in word error rate (WER). 20% relative error reduction was noted in event of DNN and 4% reduction was seen in TFCNN.

Khan Suhail Ahmad et al. [5] propelled the utilization of blend of mel frequency cepstral coefficients (MFCC) and its delta derivatives in content independent speaker recognition systems. They evaluated proposed technique using diverse sets of frame overlap and feature vector sizes. By setting frame overlap at different thresholds levels, they were able to obtain the final value which results in minimum loss of information between consecutive frames. And the adjustment in feature vector size gives most ideal precision. Delta derivatives assume a critical part in identifying speaking styles, delays and duration of a particular speaker. Instead of using triangular filter banks, Gaussian filter banks (GFBs) were employed. GFBs provide better relationship and smooth moves between two adjacent sub-bands. Moreover, to make framework computationally quicker, Principal Component Analysis (PCA) was employed. PCA extract only the relevant information, thus, reduces the feature dimensionality. Probabilistic Neural Networks (PNNs) were used as classifiers. PNNs are only a change of Back Propagation Neural Network in which exponential function is utilized in place of sigmoid function. The results demonstrated that 18 MFCC coefficients yield best outcomes at 90% frame overlap.

In [6] Dimitri Palaz et al. broadened their past work of feeding convolutional neural networks (CNNs) with raw speech signals to large vocabulary continuous speech recognition task. State-of-the-art systems typically perform the task by first transforming the data into features, usually made out of a dimensionality reduction phase and an information selection phase. Furthermore, in next stage probablility of phonemes is estimated. Nevertheless the proposed model uses sequence of raw input signals, separate them into frames and after that yield a score for each class, for each frame. There were two corpus on which proposed model was evaluated viz. TIMIT and WSJ. Initially, the features were learned using TIMIT corpus which were additionally utilized for recognizing words on WSJ. Secondly, phoneme recognition was carried out on TIMIT by learning features from WSJ. In last observations were made between CNNs and ANN (Artificial Neural Network) based system. Their investigations demonstrated that CNNs accomplished better results than ANNs which were trained using standard cepstral features as input.

In [7] Mohammed Kyari Mustafa et al. embraced techniques to minimize the required computational assets for an viable mobile based speech recognition system. Authors put use of Dynamic Multi-Layer Perceptron in proposed technique. This algorithm is fit for running on currently available mobile devices in real time. There were two databases used for this work CSLU2002 and TIDIGITS. Both of these databases include isolated spoken utterances

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

from 1 to 9. Linear Predictive Coefficients (LPC) and Mel Frequency Cepstral Coefficients (MFCC) were extricated as features from signals. Dynamic Multi-Layer Perceptron was used as classifiers because their dynamic nature allows system to deal with varying lengths of speech signals. In order to make MLP dynamic, not all of the input neurons were utilized. Only the active neurons that were indulged in any calculation were utilized. In the mean time, weights associated with the connections from these active neurons to the hidden neurons were dynamic. The size of output layer was directly equal to the number of digits to be recognized. Further experiments were conducted using deep structure of network. The end results demonstrated that even though the HMM methods outperformed dynamic MLP in terms of recognition performance but proposed model runs significantly faster than HMMs resulting in less processing time.

Inspired from past studies that articulatory data can enhance the execution of ASR frameworks, Vikramjit Mitra et al. in [8] utilized Deep Neural Networks which helps in extracting the articulatory information from the speech signals. These ariculatory trajectories were then used for continuous speech recognition task. For estimating trajectories, DNN were trained with English words corpus created by Haskins Laboratories. Execution of proposed model was evaluated using AURORA-4 database. Speech is first parameterized in the form of cepstral features. Authors explored four different cepstral features and their part in the accuracy of articualtory trajectories estimation. Initially tries were carried out using shallow networks. At that time, performance of different features contrasted essentially. However, with increase in number of hidden layers difference started to diminish. The observations demonstrated that articulatory information augmented with traditional cepstral features significantly decreased the word error rates.

In [9], Qian, Yanmin et. al. has expanded previous design of Very Deep Convolutional Neural Networks (CNNs) for strong speech recognition. The authors are inspired by the consequences of very deep CNNs in the field of computer vision, where image classification has enhanced to extraordinary degree by growing the quantity of convolutional layers in the conventional CNNs. The measurements of filters and pooling are lessened and size of input features are expanded so more number of convolutional layers can be included the framework. Diverse pooling and padding strategies are concentrated to make the framework fit for de-noising and de-resonation. Results are evaluated on Aurora-4 and AMI meeting interpretation. Best setup is accomplished at 10 convolutional layers. In the outcomes, it is inferred that input highlight maps cushioned on the two sides convey best outcomes. Additionally, it is seen that very deep CNNs with static features perform better than traditional dynamic features. The proposed arrangement of very deep CNN demonstrates enhanced word error rate with respect to LSTM-RNN acoustic models. Also, the model has minimized size and the training merging speed of network is also very quick.

In paper [10], Michael L. Seltzer et al. explored the noise dealing capability of DNN-based acoustic models and find that they can coordinate state of-the-art execution on the Aurora 4 errand with no unequivocal noise remuneration. This execution is further enhanced by consolidating data about surrounding into DNN training , utilizing another technique called noise-aware training. Whenever authors joined late proposed dropout training strategy with their presented model, a 7.5% relative change over the already best distributed outcome on this task is accomplished utilizing just a solitary decoding pass. Moreover, no extra decoding complexity was observed with a standard DNN.

In [11] Dong Yu et. al. demonstrated that deep neural networks (DNNs) perform significantly superior to anything shallow networks and Gaussian mixture models (GMMs) on huge vocabulary speech recognition errands. In this paper, authors contended that the precision accomplished by the DNNs is the consequence of their capacity to extricate discriminative inner portrayals that are powerful to the many wellsprings of fluctuation in speech signals. They demonstrated that these portrayals turn out to be progressively insensitive to the bothers in the input with expanding network depth, which prompts better speech recognition execution with deeper networks. Moreover, authors demonstrated that DNNs can't extrapolate to test samples that are considerably unique from the training illustrations. In the case that the training information is adequately illustrative, interior features learned by the DNN are generally steady with deference to speaker contrasts and environmental noises. This empowers DNN-based recognizers to execute too or superior to state-of-the-art frameworks in light of GMMs or shallow networks without the requirement for express model adjustment or feature standardization.

III. DESIGN & IMPLEMENTATION

The proposed model is based upon the very deep convolutional neural network (VDCNN) for the classification of noise robust speech recognition model, which utilizes the combination of gaussian mixture model (GMM) and Mel-

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

frequency cepstral coefficient (MFCC) features. The convolutional neural network (CNN) is a special feed forward network to describe the inter-links between deep features using the neural network layers. In CNN models, the 3-dimensional layers are utilized in contrast to the 1-D layers in the ordinary neural networks. In the following figure 1, the typical structure of CNN has been explained, which depicts the sub-sampling, convolutions and full connections. The CNNs work on the basis of two native concepts involving the local connections and parameter sharing to analyze the features between the different samples. The convolutional neural network is consisted of three major layers, which are explained in the following sub-sections:

(1) Convolutional layer,
(2) Pooling layer, and
(3) Fully-connected layer.

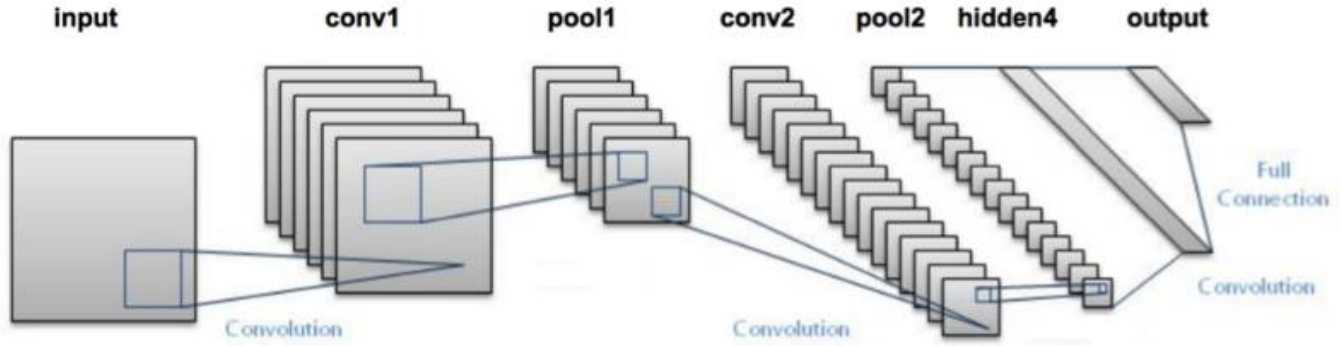

Figure1. Example of Convolutional Neural Network

A typical convolutional layer uses the parsing, processing and forwarding mechanism to process the input data. The convolution effect is known as the coiling or twisting effect, which is embedded in the convolutional neural networks to describe the repeated learning of the features over different layers in order to learn the features with greater depth. The fully-connected layering mechanism is used as per the ordinary neural network design, which is used to perform the coiling effect for feature processing and depicts the cramming based learning on the previous layers. The pooling layer describes the detailed down sampling over 2x2 data blocks in accordance with the previous layers to construct the full-stack architecture of CNN model [12].

Algorithm 1: Very Deep Convolutional Neural Network for Noise Robust Speech Recognition

1. Load the dataset to runtime memory
2. Prepare the training dataset
    a. Acquire the m number of samples from the noisy data → nD
    b. Acquire the n number of samples from the clean data → nS
    c. Combine both subsets to prepare the training dataset, TrainDat ← concatenate(nD, nS)
    d. Load the noisy data labels
    e. Extract the m number of labels according to the indices assigned by noisy data matrix → LDn
    f. Load the clean data labels
    g. Extract the n number of labels according to the indices assigned by clean data matrix → LDc
    h. Combine the label subsets and prepare the training data label set, TrainLabel ← concatenate(LDn, LDc)
3. Prepare the testing dataset
    a. Acquire the q number of samples from the noisy data → nDtest
    b. Acquire the t number of samples from the clean data → nStest
    c. Combine both subsets to prepare the training dataset, TestDat ← concatenate(nDtest, nStest)
    d. Load the noisy data labels

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

e. Extract the q number of labels according to the indices assigned by noisy data matrix → n_test
f. Load the clean data labels
g. Extract the t number of labels according to the indices assigned by clean data matrix → c_test
h. Combine the label subsets and prepare the training data label set, TestLabel ← concatenate(n_test, c_test)

4. Define number of clusters → Nc
5. Define number of iterations → Ni
6. Run the iteration over each column of training data to extract the features, count *v*
   a. Extract current column, cMv ← TrainDat(v)
   b. Reshape the current data vector matrix, cMm ← Reshape(cMv, noColumns, noRows)
   c. Apply GMM over the data vector using Ni and Nc parameters → gM
   d. Apply MFCC over the GMM output → gMm
   e. Update the training feature matrix, TFM(v) ← gMm
7. Run the iteration over each column of testing data to extract the features, count *v*
   a. Extract current column, cMv ← TestDat(v)
   b. Reshape the current data vector matrix, cMm ← Reshape(cMv, noColumns, noRows)
   c. Apply GMM over the data vector using Ni and Nc parameters → gM
   d. Apply MFCC over the GMM output → gMm
   e. Update the testing feature matrix, TSFM(v) ← gMm
8. Apply feature scaling over the training feature matrix →sTFM
9. Apply feature scaling over the testing feature matrix →sTSFM
10. Fix missing values in the training feature matrix → siTFM
11. Fix missing values in the testing feature matrix → siTSFM
12. Apply very deep convolutional neural network (VDCNN) model
    a. Assign the maximum number of iterations
    b. Assign the batch size to process the data in segments
    c. Assign the step size to describe the gradient based variation in data over each iteration
    d. Assign the minimum percentage of samples to control sample elimination
    e. Declare the method of computation for classifier training
    f. Input the number of layers for CNN
    g. Create the convolutional neural network (CNN) → net
    h. Train the CNN model with training data and input arguments, dnn ← train(net, siTFM, TrainLabel)
    i. Test the CNN model with testing data to estimate the predictions, preds ←test(dnn, siTSFM, siTFM)
    j. Return the neural network results
13. Verify the overall performance by comparing the classification predictions and TestLabel
14. Compute and return the final results

IV. RESULTS & DISSCUSSION

The very deep convolutional neural network (VDCNN) model has been utilized among the proposed model for noise robust speech recognition. The Gaussian mixture model (GMM) and Mel-frequency cepstral coefficient (MFCC) has been combined for the feature description over the speech samples, which describes the feature in 20x3 matrix. The following table 1 shows the results upon the VDCNN with 4-layers and 200 samples. The 4-layer VDCNN model has been undergone 10 testing rounds, where it achieves the average word error rate (WER) percentage at 18%, which is significantly higher than existing models. Also, the 4-layer model does not meet the very deep learning paradigm.

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

Table 1: VDCNN with 4-layers and 200 samples

Iteration Number	RMSE	CER	WER
1	0.21	1.78	8
2	0.21	1.71	12
3	0.22	1.82	20
4	0.21	1.73	28
5	0.33	2.17	20
6	0.22	1.8	24
7	0.2	1.7	12
8	0.21	1.84	20
9	0.22	1.82	8
10	0.21	1.76	28

The following table 2 shows the results upon the VDCNN with 5-layers and 400 samples. The 5-layer VDCNN model has been also undergone 10 testing rounds like the previous deviation, where it achieves the average word error rate (WER) percentage at 20%. This model is found with higher than the all of the classification models of noise robust speech recognition, which shows the worst performance among all of the models.

Table 2: VDCNN with 5-layers and 400 samples

Iteration Number	RMSE	CER	WER
1	0.22	1.8	24
2	0.91	1.7	28
3	0.22	1.91	24
4	0.83	1.82	24
5	0.21	1.75	12
6	0.89	1.71	12
7	0.91	1.72	36
8	0.9	1.81	16
9	0.9	1.75	8
10	0.9	1.8	16

The following table 3 shows the results upon the VDCNN with 6-layers and 2000 samples. The 6-layer VDCNN model has been also undergone 10 testing rounds as the latter models. The proposed model has been recorded with average accuracy of nearly 7.5% WER, which shows the robust performance among the speech recognition models. This model is recorded significantly lower than the existing models, where the difference has been recorded between 9.28% and 10.96%.

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

Table 3: VDCNN with 6-layers and 2000 samples

Iteration Number	RMSE	CER	WER
1	0.28	3.69	8.33
2	0.66	3.72	13.33
3	0.61	3.87	8.33
4	0.56	3.82	8.33
5	0.64	3.93	6.67
6	0.52	3.82	3.33
7	0.72	4.03	6.67
8	0.61	3.75	6.67
9	0.63	3.91	6.67
10	0.71	3.95	6.67

The following table 4 shows the results upon the VDCNN with 6-layers and 4000 samples. This model has been also tested in 10 random rounds to access the accuracy of the proposed model in recognizing the speech over the standard aurora-4 dataset. The WER has been recorded between 1.6 and 4.8% for the word recognition in the speech samples. Also the character error recognition (CER) has been also recorded between 7.2% and 9.11%.

Table 4: VDCNN with 6-layers and 4000 samples

Iteration Number	RMSE	CER	WER
1	0.86	9.11	4
2	0.88	7.83	1.6
3	0.78	7.2	2.4
4	0.3	7.75	1.6
5	0.46	7.66	2.4
6	0.87	7.86	4.8
7	0.22	7.56	4
8	0.25	7.79	2.4
9	0.86	8.08	3.2
10	0.87	7.82	1.6

The following table 5 shows the results upon the VDCNN with 6-layers and 8000 samples. This model has been also tested in 10 random rounds to access the accuracy of the proposed model in recognizing the speech over the standard aurora-4 dataset. The WER has been recorded between 0.8% and 2.4% for the word recognition in the speech samples. Also the character error recognition (CER) has been also recorded between 14.72% and 17.85%. The proposed model based upon 6-layered VDCNN has been found most efficient among all of the classification models for the speech recognition.

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

Table 5: VDCNN with 6-layers and 8000 samples

Iteration Number	RMSE	CER	WER
1	0.21	15.27	0.8
2	0.86	17.1	1.2
3	0.88	16.15	1.6
4	0.21	14.92	1.6
5	0.27	14.72	1.6
6	0.87	15.62	2
7	0.5	14.77	1.6
8	0.87	15.01	2.4
9	0.25	15.53	1.2
10	0.84	17.85	2.4

The comparison of the proposed model has been conducted with various existing noise robust speech recognition models, which includes the very deep learning (VDL), convolutional neural networks (CNN), time-extension, frequency-extension and full extension (combing frequency and time) models. The proposed model is based upon the very deep convolutional neural network (VDCNN) has been arranged with different number of layers (4, 5 and 6) with different number of samples varying between 200 and 8000. The noisy and clean speech samples have been obtained from the respective datasets for the performance analysis. The proposed VDCNN (vd6) model has been recorded with 8.864%, 3.12% and 2.24% WER, which gives the average value of 4.74%. The proposed vd6 with 4.74% average WER model has been found efficient than the existing vd6 and vd10 models, which are recorded with 10.34% and 9.78% respectively. The proposed vd6 model with 4.74% WER has been also found efficient than time-extension, frequency extension and full extension models with 9.84%, 10.02% and 9.28% WER respectively. Also, the existing convolutional neural network (CNN) based models with 10.64% and 10.96% WER are outperformed by proposed vd6 (VDCNN) model with 4.74% WER.

Table 6: Comparative analysis of proposed model against existing models

Model	TxF	L	A	B	C	D	AVG
CNN	11x40	2	4.11	7	6.33	16.09	10.64
vd6	11x40	6	3.94	6.86	6.33	15.56	10.34
time-ext	17x40	8	3.72	6.57	5.83	14.79	9.84
freq-ext	11x64	10	3.79	6.51	6.26	15.19	10.02
vd10	17x64	10	4.13	6.62	5.92	14.53	9.78
full-ext	21x64	10	4.04	6.23	5.4	13.86	9.28
CNN2	17x64	2	4.2	7.36	6.84	16.36	10.96
Proposed vd4-VDCNN	20x3	4	8	12	20	28	14.4
Proposed vd5-VDCNN	20x3	5	24	28	24	24	21
Proposed vd6-VDCNN (2000)	20x3	6	8.33	13.33	8.33	8.33	8.864
Proposed vd6-VDCNN (4000)	20x3	6	4	1.6	2.4	1.6	3.12
Proposed vd6-VDCNN (8000)	20x3	6	0.8	1.2	1.6	1.6	2.24

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

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

(An UGC Approved Journal)

Website: www.ijareeie.com

Vol. 6, Issue 8, August 2017

V. CONCLUSION

The noise robust speech recognition model is entirely based upon the neural network classification model using very deep convolutional neural network (VDCNN) with GMM and MFCC features. The proposed model has been designed with 6 layers under the deep learning model, which is tested over four differently drawn test datasets categorized A, B, C and D. The word error rate (WER) percentage has been estimated over the randomly drawn dataset variants, which has been computed for average value over all four samples. The proposed model has been tested against the various noise robust speech recognition models of convoluional neural network, very deep (6-layer) model, very deep (10-layers) model, very deep convolutional neural network with 4, 5 and 6 layers with different number of samples. The proposed model of vd6-VDCNN has been recorded with WER percentage of 8.864, 3.12 and 2.24 with 2000, 4000 and 8000 samples respectively. The overall average of WER percentage for vd6-VDCNN model is found at 4.7413, which is significantly higher than existing CNN (10.64), vd6 (10.34), time-ext (9.84), freq-ext (10.02), vd10 (9.78), full-ext (9.28) and CNN2 (10.96), which shows the robustness of the proposed model.

REFERENCES

[1] Yu, Dong, and Li Deng. Automatic speech recognition: A deep learning approach, Springer, 2014
[2] Manoj Kumar Sharma and Omendri Kumari, “Speech Recognition: A Review”, National Conference on Cloud Computing & Big Data, pp. 62-71. 2017
[3] Bhoomika Dave, D.S. Pipalia, “Speech Recognition: A Review”, IJAERD, vol. 1, Issue 12, pp. 230-236, 2014
[4] Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Espy-Wilson and Elliot Saltzman, “Articulatory features from deep neural network and their role in speech recognition”, IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 3017-3021, 2014
[5] Khan Suhail Ahmed, Anil S. Thosar, Jagannath H. Nirmal and Vinay S. Pande, “A unique approach in Text independent speaker recognition using MFCC feature sets and probabilisitic neural networks”, IEEE International Conference on Advances in Pattern Recognition, pp. 1-6, 2015
[6] Dimitri Palaz, Mathew Magimai. Doss and Ronan Collobert, “Convolutional Neural Network based continuous speech recognition using raw speech signal”, IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 4295-4299, 2015
[7] Mohammed Kyari Mustafa, Tony Allen and Kofi Appiah, “A comparative review of dynamic neural network and hidden markov model methods for mobile on-device speech recognition”, Neural Computing and Applications, pp. 1-9, 2017
[8] Vikramjit Mitra, Horacio Franco, Chris Bartels, Julien Van Hout, Martin Graciarena and Dimitra Vergyri, “Speech recognition in unseen and noisy channel conditions”, IEEE International Conference on Acoustics, Speech and Signal processing, pp. 5215-5219, 2017
[9] 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.
[10] Michael L. Seltzer, Dong Yu and Yongqiang Wang, “An investigation of deep neural networks for noise robust speech recognition”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7398-7402, 2013
[11] Yu Dong, Michael L. Seltzer, Jinyu Li, Jui-Ting Huang, Frank Seide, "Feature learning in deep neural networks-studies on speech recognition tasks." *arXiv preprint arXiv:1301.3605*, pp. 1-9, 2013
[12] O'Shea, Keiron, and Ryan Nash, "An introduction to convolutional neural networks", arXiv preprint arXiv:1511.08458, 2015