IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 5, May 2021 ||

DOI:10.15680/IJIRSET.2021.1005009

An Audio Speech Enhancement Using Deep Convolutional Neural Network

R.Saktheeswaran, V.N.Sankarananth, B.Vishal, S.Balasubramanian

U.G. Student, Department of Electronics and Communication Engineering, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India

U.G. Student, Department of Electronics and Communication Engineering, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India

U.G. Student, Department of Electronics and Communication Engineering, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India

Associate Professor, Department of Electronics and Communication Engineering, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India

ABSTRACT: The project is about Audio speech enhancement using deep convolutional neural networks. The objective of enhancement is improvement in intelligibility and overall perceptual quality of degraded speech signal. There are three main steps preprocessing, training and the testing. Here we take audio signal and preprocessing is done where the audio signal is converted to 50frames/second and STFT is taken and combine it with noise signal which is given as input to training network. Then the network is trained using various layers of deep convolutional neural networks. Then test sample (audio signal + noise) is given as input to the network in which we obtain the enhanced audio signal as output. We have carried out the analysis for the enhancement using various noise signals such as factory noise, babble noise, destroy engine. we have been instrumental in achieving good SNR ratio of the output enhanced signal to the input noise signal.

KEYWORDS: Features Extraction, Predictor, Target, Estimate, Convolutional Neural Network

I. INTRODUCTION

The primary goal of speech enhancement (SE) is to increase the quality of noisy speech signals by reducing the noise components of noise-corrupted signal to attain a good performance. Speech Enhancement has been used in various speech- related applications, such as automatic speech recognition, speaker recognition, speech coding, hearing aids, and cochlear implants. In the past few years, numerous Speech Enhancement methods have been proposed and has proven to provide an improved sound quality. One notable approach, spectral restoration, estimates a gain function (based on the statistics of noise and speech components), which is then used to suppress noise components in the frequency domain to obtain a clean speech spectrum from a noisy speech signal. Another class of approaches adopts a nonlinear model to map noisy speech signals to clean ones. In recent years, SE methods based on deep learning have been proposed, such as denoising auto-encoders. SE methods using deep neural networks (DNN) generally has better efficiency than old SE models. Approaches that utilize long short-term memory models have also been confirmed to exhibit efficient SE and related speech signal processing efficiency .

II. METHODOLOGY

Audio has many different ways to be represented, going from raw time series to time- frequency de-composition. Among time-frequency decompositions, for audio processing the spectrogram helps to be a use full tool for representation. In addition CNN- based model has been shown to obtain good results in enhancing speech owing to its strength in handling image like 2-D time-frequency representations of noisy speech.

They consist of 2D images representing sequences of STFT with time and frequency as axes, and brightness representing the strength of a frequency component at each time frame. As such, they appear a natural domain to apply

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

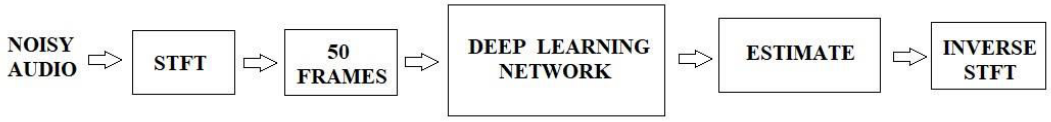
|| Volume 10, Issue 5, May 2021 ||

DOI:10.15680/IJIRSET.2021.1005009

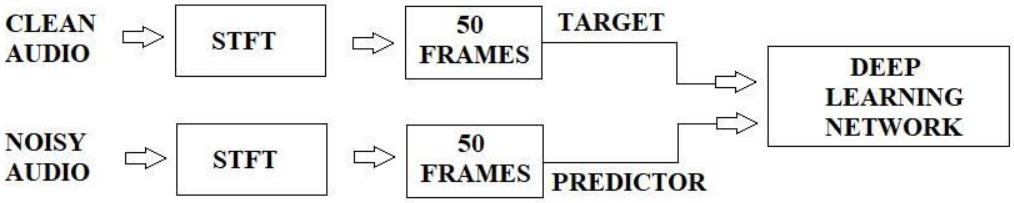
the CNN architecture for images directly to sound. Between magnitude and phase spectrograms, magnitude spectrograms contain most the information of the signal . In this project, we will use magnitude spectrum for representing the sound in order to predict the noise model to be subtracted to a noisy voice spectrum

The predictor and target network signals are the magnitude spectra of the noisy and clean audio signals, respectively. The network's output is the magnitude spectrum of the signal without noise . The regression network uses the predictor input to reduce the mean square error between its output and the input target. The denoised audio is then converted back to the time domain using the output magnitude spectrum and the phase of the noisy signal.

Transform the audio signal to the frequency domain using the Short-Time Fourier transform (STFT), with a window length of 257 samples, an overlap of 37.5%, and a Hamming window. The predictor input consists of 5 noisy STFT frames, and then each STFT output estimate is computed based on the current noisy STFT and the added with ±2 noisy STFT frames.



(a)



(b)

Figure 1: Block Diagrams (a) Training Process (b) Testing Process

DATA COLLECTION

To create the datasets for training, we gathered English speech examples with clean voices and environmental noises from different sources. The clean voices were mainly gathered from grid corpus. GRID is a large multitasked audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. For this project, we focused on 3 classes of environmental noise: babble noise, factory noise, Destroyer Engine noise. To create the datasets for training/validation/testing, audios were sampled at 16 kHz and extracted each windows of 32 millisecond.

FEATUREEXTRACTION

We resampled the audio signal to 16 kHz, and used only a mono channel for further processing. Speech signals were converted into the frequency domain and processed into a sequence of frames using the short- time Fourier transform. Each frame contained a window of 32 milliseconds, and the window overlap ratio was 37.5%. Therefore, there were 50 frames per second. We concatenated ±2 frames to the central frame as context windows. Accordingly, audio features had dimensions of 257 × 5 at each time step

USING THE ADCNN MODEL FOR SPEECHENHANCEMENT

An ADCNN network comprised of Convolutional layer, pooling layer and fully connected layers. A 2-D convolutional layer applies sliding filters to the input. The layer convolves the input by moving the filters along matrix both horizontally and vertically and computing the dot product of the weights and the input, and then adding a bias term. Convolutional layers typically consist of fewer parameters than fully connected layers in a fully connected layer all the neurons are connected to the activation points previous layer. A fully connected layer multiplies the input by a weight matrix and then adds a bias vector. The dimensions of the weight matrix and bias vector are determined by the number of neurons which is present in the layer and the number of activations points from the previous layer.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 5, May 2021 ||

DOI:10.15680/IJIRSET.2021.1005009

Define layers of ADCNN network consists of two Convolutional layers with filter width of 12 and 5. Number of filters 10 and 4 respectively Convolutional layer are followed by Relu and Pooling layer .The pooling layer consists of filter width of 2 and Pooling layer is between two Convolutional layers and followed by three fully connected layer consisting of 600,400,257 neurons respectively

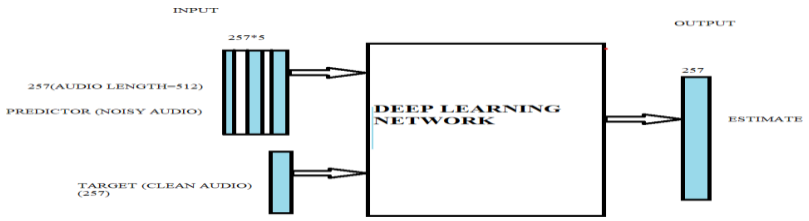
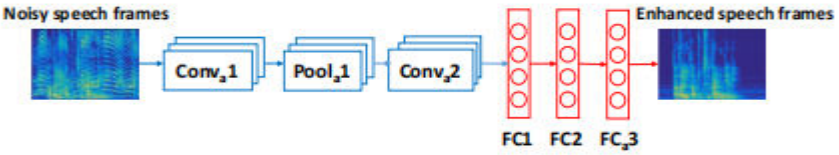


Figure 2: ADCNN model overview



.

Figure 3: Architecture of the ADCNN model

TRAINING THE ADCNN MODEL

To train the AVDCNN model, we first prepare noisy-clean speech pairs we have the amplitudes of noisy (X) and clean (Y) spectra for each time step, we obtain the output of the audio network as,

$$A_i = Conv_a2\left(Pool_a1\left(Conv_a1(X_i)\right)\right), i = 1 \dots K$$

where K is the number of training samples. \
Fully-connected network is computed as:

$$\hat{Y}_i = FC_a3\left(FC2\left(FC1(F_i)\right)\right), i = 1 \dots K$$

Next, specify the training options for the network. Set Max Epochs to 15 so that the network makes 15 passes through the training data. Set Mini Batch Size of 64 so that the network looks at 64 training signals at a time. Specify Plots as "training-progress" to generate plots that show the training progress as the number of iterations increases. Set Verbose to false to disable printing the table output that corresponds to the data shown in the plot into the command line window. Specify Shuffle as "every-epoch" to shuffle the training sequences at the beginning of each epoch. Specify Learn Rate Schedule to "piecewise" to reduce the learning rate by a factor (0.9) every time a certain number of epochs has passed. Set Validation Data to the validation predictors and targets. Set Validation Frequency such that the validation mean square error is computed once per epoch.

LAYER NAME	KERNAL	ACTIVATION FUNCTION	NUMBER OF NUERONS OR FILTERS
CONV 1	12 x 2	LINEAR	10
POOL 1	2 x 1		
CONV 2	5 x 1	LINEAR	4
FC 1		LINEAR	600
FC 2		LINEAR	400
FC 3		LINEAR	257

Table 1: Configuration of the ADCNN

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 5, May 2021 ||

DOI:10.15680/IJIRSET.2021.1005009

TEST THE ADCNN MODEL

After training the ADCNN Network read test data set using network. In this testing stage,we will use speech with added noise not used in the training stage. Compute the denoised speech signals. ISTFT performs the inverse STFT. Use the phase of the noisy STFT frames to reconstruct the time-domain signal. In the testing phase, the amplitudes of noisy speech signals are fed into the trained AVDCNN model to obtain the amplitudes of enhanced speech signals as outputs. Similar to spectral restoration approaches, the phases of the noisy speech are borrowed as the phases for the enhanced speech. Then, the ADCNN- enhanced amplitudes and phase information are utilized to bring out the enhanced speech

IV. EXPERIMENT AND RESULTS

Dataset's collection involves the process of collecting different audio speech signals from GRID CORPUS website. Collected Audio speeches consists of audios with length of 3 seconds in mp3 format and contains different speakers such as male and female. The main steps involved in this process is (1) Pre-Processing (2) Training (3) Testing. These datasets are then preprocessed and converted from Mp3 to 2DFrames. These audio signals are converted to2D frames (257x150) because CNN network accepts only images and frames as input. The datasets are separated as predictor and target. Predictor which is the noisy speech with various SNR levels and target is the clear speech. After preprocessing is done, the training and testing process takes place and the results are obtained. The CNN network is trained for obtaining the enhanced speech in which the predictor and target is given as input. The output from the CNN network is the estimate which consists of the enhanced speech of the noisy input speech. This estimate consists frames of 257x1x150. We take inverse short time flourier transform (ISTFT) for the estimate in order to get the audio speech for obtaining the frames

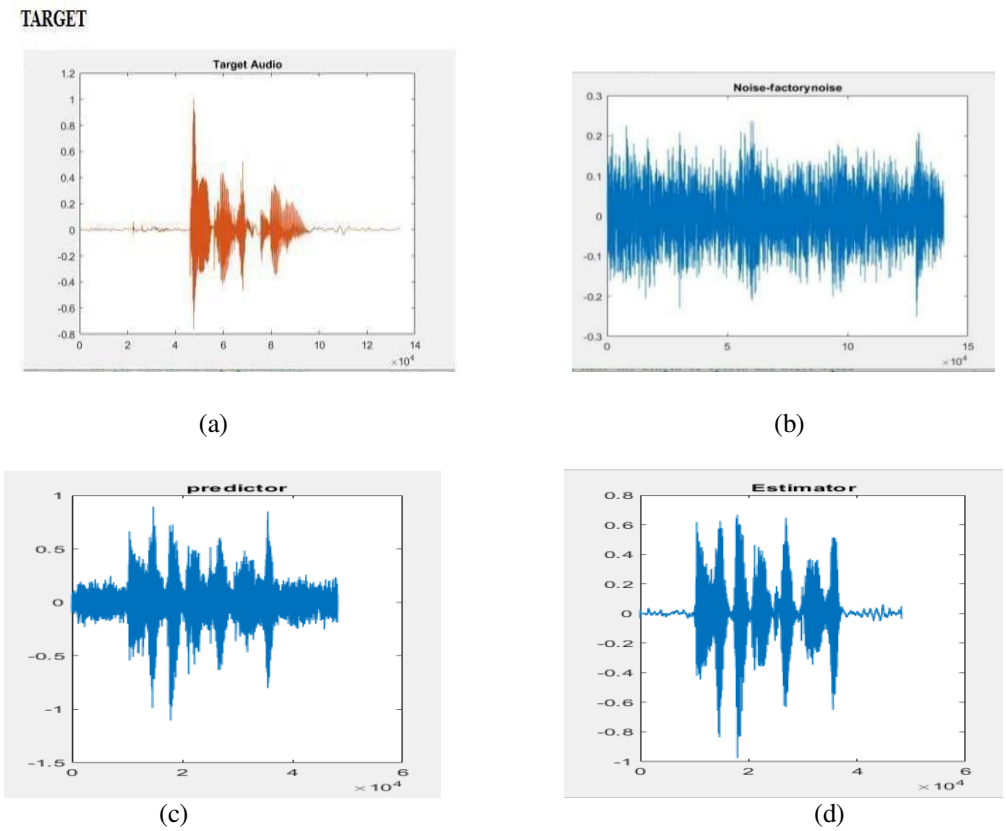


Figure 4: (a) Target Audio, **(b)** Factory Noise, **(c)** Predictor Audio, **(d)** Estimator Audio

MEASUREMENT

After getting audio speech from frames estimate to find the enhancement of speech signal we measure the SNR of audio speech obtained from frames estimate and compare with SNR of input noisy speech in order to find SNR improvement. Since the input noisy speech (predictor) consists of 3 different noisy speech babble noise, destroyer

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 5, May 2021 ||

DOI:10.15680/IJIRSET.2021.1005009

Engine, Factory noise. we have calculated various SNR for all noisy speech and for audio enhanced speech.

SL. NO	S N R	OUTPUT SNR	SNR Improvement in dB	OUTPUT SNR	SNR Improvement in dB	OUTPUT SNR	SNR Improvement in dB
		BABBLE NOISE	BABBLE NOISE	Destroyer Engine Noise	Destroyer Engine Noise	Factory Noise	Factory Noise
1	-5	3.8932	8.8932	3.9195	8.9195	3.5248	8.5248
2	-4	4.0289	8.0289	4.4109	8.4109	3.8617	7.8617
3	-3	4.4604	7.4604	4.6557	7.655	4.0958	7.0
4	-2	5.0197	7.0197	5.1831	7.1831	4.5333	6.5333
5	-1	5.2954	6.2954	5.8000	6.8000	4.9107	5.9107
6	0	5.8954	5.8954	6.2165	6.2165	5.2915	5.2915
7	1	6.1628	5.1628	6.2990	5.2990	5.7082	4.7082
8	2	6.3116	4.3116	6.6895	4.6895	6.2570	4.2570
9	3	6.6454	3.6454	6.8770	3.8770	6.3673	3.3673
10	4	7.0287	3.0287	7.5352	3.5352	6.7445	2.5352
11	5	7.5859	2.5859	7.8937	2.8937	7.1183	2.1183
Average			5.6127		5.9526		5.2822

Table 2: Various levels of SNR

After calculating various SNR for all noisy speech and for audio enhanced speech we have plotted graph for three different noisy speech corresponding enhanced speech. As input SNR for three different noisy speech babble noise, Destroyer Engine noise, Factory Noise is plotted in x axis and corresponding enhanced speech for three noisy speech is plotted in y axis. The graph is shown below

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 5, May 2021 ||

DOI:10.15680/IJIRSET.2021.1005009

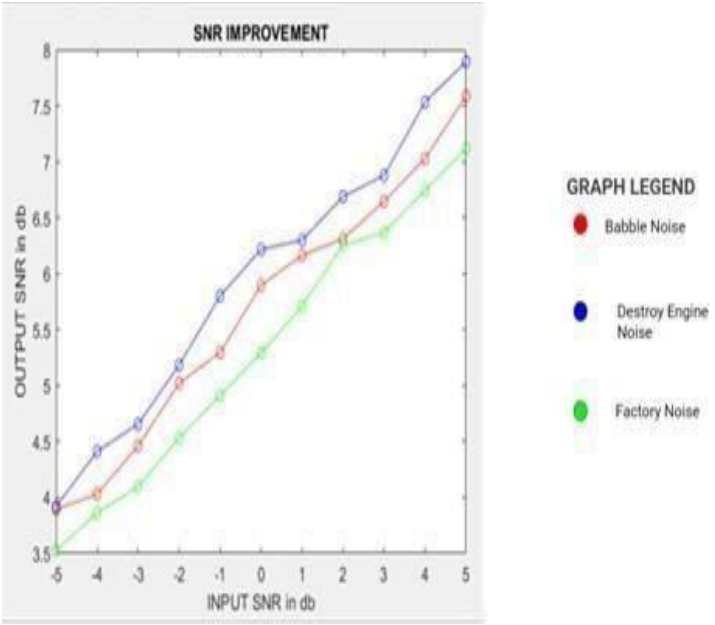


Figure 5:SNR comparison

V. CONCLUSION

In this project, we have proposed a novel CNN-based audio system learning for speech enhancement, called ADCNN. The model utilizes individual networks to process input data with different modalities. We trained the model in an end-to-end manner. The experimental results obtained using the proposed architecture show that its performance for the SE task is efficient, here we have taken audio speeches with three different noise speech such as babble, factory and destroyer engine noise with various SNR levels from -5 to 5. Here were able to get SNR levels averaging from 5 to 6 dB confirming the effectiveness of audio information into the SE process. . We have been instrumental in achieving good SNR ratio of the output enhanced signal to the input speech mixed with noise speech.

REFERENCES

1. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2), Apr 1979
2. Guo-Hong Ding, Taiyi Huang, and Bo Xu. Suppression of additive noise using a power spectral density mmse estimator IEEE Signal Processing Letters, 11(6):585–588, June 2004.
3. Tobias Goehring, Federico Bolner, Jessica JM Monaghan, Bas van Dijk, Andrzej Zarowski, and Stefan Bleeck. Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. Hearing research, 344:183–194, 2017.
4. Szu-Wei Fu, Yu Tsao, and Xugang Lu. Snr-aware convolutional neural network modeling for speech enhancement. In INTERSPEECH, pages 3768–3772, 2016Szu-Wei Fu, Yu Tsao, and Xugang Lu. Snr-aware convolutional neural network modeling for speech enhancement. In INTERSPEECH, pages 3768–3772, 2016
5. Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters, 21(1):65–68, 2014.
6. S. -W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform -based speech enhancement by fully convolutional networks,” arXiv: 1703.02205, 2017.
7. S Lakshmikanth, KR Nataraj, and KR Rekha. Noise cancellation in speechsignal processing: A review. International Journal of Advanced Research in Computer and Communication Engineering, (1), 2014.
8. O. Ronneberger, P. Fischer, and T. Brox, “U -Net: convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015
9. Rezayee, and S. Gazor, “An adaptive KLT approach for speech enhancement,” IEEE Transactions on Speech and Audio Processing, vol. 9, no. 2, pp. 87–95, 2001.
10. D. Liu, P. Smaragdis, and M. Kim, “Experiments on deep learning for speech denoising,” in Proc. INTERSPEECH, 2014, pp. 2685– 2689.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details