# Dilated Multi-Activation Autoencoder to Improve the Performance of Sound Separation Mechanisms

Ghada Dahy[1] , Mohammed A.A.Refaey[2], Reda Alkhoribi[3], M.Shoman[4]
Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University,
Cairo, Egypt

***Abstract*—Speech enhancement is the process of improving the quality of audio relative to target speaker while suppressing other sounds. It can be used in many applications as speech recognition, mobile phone, hearing aids and also enhancing audio files resulted from separation models. In this paper, a convolutional neural network (CNN) architecture is proposed to improve the quality of target's speaker resulted from speech separation models without having any prior information about the background sounds. The proposed model consists of three main phases: Pre-Processing phase, Autoencoder phase and Retrieving Audio phase. The pre-processing phase converts audio to short time Fourier transform (STFT) domain. Autoencoder phase consists of two main modules: dilated multi-Activation encoder and dilated multi-Activation decoder. Dilated multi-Activation encoder module has a six blocks with different dilation factors and each block consists of three CNN layers where each layer has different activation function then the encoder's blocks are arranged in reverse order to construct dilated multi-activation decoder. Audio retrieving phase is used to reconstruct audio depending on feature resulted from second phase. Audio files resulted from separation models are used to build our datasets that consist of 31250 files. The proposed dilated multi-activation autoencoder improved separated audios Segmental Signal-to-Noise Ratio (SNRseg) with 33.9%, Short-time objective intelligibility (STOI) with 1.3% and reduced bark spectral distortion (BSD) with 97%.**

***Keywords—Speech de-noising; speech enhancement; speech separation; short time Fourier transform (STFT); autoencoder; dilated Convolution neural network; multi-activation functions, convolution neural network (CNN); bidirectional long short memory (BLSTM)***

## I. INTRODUCTION

The field of machine learning proposed different architectures in complex and difficult problems that were unattainable in the field of speech processing. Speech separation and enhancement are considered two of the most important problems in signal processing where machine learning field achieves better performance on dealing with them. Speech enhancement is used as a step of removing or reducing background noise to enhance the main signal of the target speaker. In the current days, researchers do a lot of effort to have a very high quality speech that relative to the target speaker.

In the field of speech analysis, speech enhancement [1] focus on having high quality sound of the target speaker by suppressing other sounds except sound of the target one [2] while speech separation [3] is considered the process of separating mixed sounds.

The main objective of speech enhancement and separation is extracting signals of interest parts from an audio according to critical rules that can be done by post production software that depends currently on deep learning field. Speech enhancement is considered the main stone of many applications as teleconferencing systems, speech recognition [4,5], hearing aids [6-8], speaker recognition [9,10] and also can be used to enhance the resulted signals from speech separation systems.

Speech understanding in noisy environments is still one of the major challenges as we need very high quality audio files with large datasets and different types of noises to train the enhancement models and this will need very high quality resources. To face the previous challenge, we built a dataset with 31250 audio files resulted from separation model that trained on a subset samples from The Oxford-BBC LRS2 Dataset.

Recently, deep learning made a great improvement in the speech enhancement field by using different structures over signal processing methods that based on supervised enhancement techniques.

The main contribution of the proposed model is implementing speech enhancement structures to improve the speech quality of the target's speaker resulted from separation models by analyzing input speech in multiple levels. The proposed model plays an important role in the digital speech signal processing according to the type of degradation and background sounds in the speech signal of the target speaker.

In this paper, we proposed a new approach for speech enhancement which achieved an interesting performance comparing against old methods. It improves the quality of speech resulted from separation model. It consists of three main phases: Pre-Processing phase, AutoEncoder phase and Retrieving Audio phase. It considers low quality audios resulted from separation models as an input and complex ratio mask relative to cleared audio as output. The dilation convolutional neural network is considered as the main stone of our proposed model as it has ability to expand the covered areas of the input audio without needing to apply pooling that maybe cause missing some information. Dilated convolutional network mainly uses dilation factor to control the size of the covered area. It also could analysis input signal to suppress background sounds without increasing kernel parameters.

The rest of this paper is organized in the following order. Section II introduces the related work; Section III introduces the proposed structures and defines the main mechanism of speech enhancement in details. Section IV describes the dataset, network parameters and demonstrates the performance of our network. Finally, we conclude the idea of the proposed model in Section V.

## II. Related Work

Neural Networks is considered one of the most common machine learning algorithms. It has been improved over time and also outperforms other algorithms and methods in speed and accuracy. There are different types of neural network as convolutional neural network (CNN) [11.12], RNN (Recurrent Neural Networks) [13], Autoencoders [14], LSTM (Long short Tern Memory) [15,16] and also WaveNet that depends on dilated convolution network [17].

Emad el al. [11] proposed multi-resolution convolution neural network that support them in separating audio files from mixed sounds. They proved that concatenating sets of convolution network with different filters could improve the performance of separation compared with deep neural network (DNN) and fully convolution neural network.

Jen-Cheng el al.[12] proposed audio visual model for speech enhancement using CNN that complete its process by analyzing visual and audio stream which processed using separated CNN models then fed into AUDIO-VISUAL Autoencoder model. Their model could outperform speech enhancement models that use audio-only and existing audio-visual enhancement models.

Jinuk el al. [13] proposed model to separate audio sounds of two speakers. They used CNN followed by LSTM and recurrent neural network (RNN) then a fully connected layer is applied on the concatenation of forward LSTM-RNN and backward LSTM-RNN. They achieved better signal-distortion-ratio (SDR) and signal interference ratio (SIR) compared with deep clustering method that used in sound separation.

Yi Luo and Nima Mesgarani [15] proposed time-domain audio separation model to obtain target speaker sound with acceptable performance. Their proposed autoencoder is able to outperform LSTM that analyze log power magnitude spectrogram to complete the process of separation.

Jitong Chena and DeLiang Wang [16] proposed speech separation model based on LSTM that suitable to temporal dynamics of speech. They used Short-time objective intelligibility (STOI) to measure the performance of their model which outperformed DNN model as LSTM could capture long term context of input speech and has low latency in completing speech separation process.

Aaron et al. [17] proposed dilated CNN model called wavenet that captures the main characteristics of multiple speakers. their model has been trained on thousands of samples. They trained wavenet on different datasets as text to speech and music datasets. It could generate musical fragment that very near to realistic and has very high quality. In WaveNet, a set of dilated convolution layers are stacked together to generate a very large receptive fields with a few layers. Recently, wavenets are used in time domain. It can be used in songs separation from background noise and enhancing speech quality.

Dario et al. [18] proposed speech de-noising technique that able to improve speech signals quality with background-noise. Their proposed model depends on dilated convolution. Their proposed model used different dilation factor in each layer changes in the following order 1, 2, ..., 256, 512. Their pattern is repeated three times. Their model has capability to de-noise an entire audio file in one-shot. They generated their datasets from two sources: speech data from the Voice Bank corpus and environmental sounds from Diverse Environments Multichannel Acoustic Noise Database (DEMAND). Their Perceptual tests proved that their model's outperform the Wiener filtering results. They proved that their system able to learn multi-scale hierarchical representations from raw audio instead of magnitude spectrograms.

Yi Luo and Nima Mesgarani [19] proposed Time-domain Audio Separation Network (TasNet). Mainly, their network structure used in speech separation and also can be used in speech enhancement. It consists of four processes: a preprocessing normalization, an encoder to retrieve the mixture weight, a separation and a decoder to construct waveform. The first process in their network is used to predict the weights that relative to each source in the mixture weight. The weight detection is defined as masks that predict the contribution of each speaker in the mixture weight as T-F masks that are used in short time Fourier transform (STFT) models. In the Separation Network, a deep LSTM network is used to estimate the source masks. They used WSJ0-2mix dataset to evaluate their system on two-speaker speech separation problem. The mixtures relative to their dataset are generated randomly speakers from Wall Street Journal (WSJ0) dataset. Their proposed TasNet system could outperforms systems that use a T-F representation as an input to complete the process of learning. Comparing with STFT system, their experiments proved that the TasNet system was six times faster and could achieve very high quality speech separation performance and also it can be used in speech enhancement.

Craig Macartney and Tillman Weyde [20] proposed speech enhancement model that used Wave-U-Net architecture for speech enhancement. Their experiments proved that the Wave-U-Net could improve several metrics as Perceptual Evaluation of Speech Quality (PESQ), the rating of background distortion (CBAK), the rating of speech distortion (CSIG), segmental signal-to-noise ratio (SSNR) and the predicting rating overall quality (COVL) over the state-of-the-art. They showed that reducing number of hidden layers is more suitable for speech enhancement.

Yong Xu et al. [21] proposed a speech enhancement regression model depending on deep neural networks (DNNs) that consists of multiple-dense layers. Their proposed model could achieve better performance in case of different objective measures. They held a subjective evaluation measure to check the quality of DNN model with 10 listeners, about 76.35% of the listener prefer using DNN results to complete the process of enhancement.

Andreas Jansson et al. [22] proposed a novel architecture of the U-Net architecture that used to complete the source separation task. They benefit from skip connections paths in building encoder-decoders of u-net. They held quantitative evaluation and subjective assessment to evaluate the performance of their model. Their experiments proved that their proposed architecture able to achieves state-of-the-art performance.

## III. Proposed Model

In this paper, we proposed a model based on dilation convolution neural network that able to analysis and discover local pattern of the target speaker in STFT. The main objective of the proposed model is suppressing other human's sounds except target speakers. It mainly enhances the quality of speech resulted from separation models [23].

Fig. 1 depicts the structure of the proposed model. It consists of three main phases: Pre-Processing phase, Autoencoder phase and Retrieving Audio phase. It takes low quality audios resulted from separation models as an input and scaled complex ratio mask of the cleared audios with same size of the input using (1,2) as output.

$$Mask_{(t,f)} = \frac{Target\ sound_{(t,f)}}{Mixed\ Sound_{(t,f)}} \quad (1)$$

$$Scaled\ Mask_{(t,f)} = 10\,\frac{1-e^{-0.1*Mask}}{1+e^{-0.1*Mask}} \quad (2)$$

where, t and f are considered as indexes of time and frequency.

In the Pre-Processing phase, audio files are converted to STFT to generate 304X256X2 real and imaginary numbers that taken as an input to the Autoencoder phase and saved as numbers to make the process of training faster.

Autoencoder phase consists of two modules, Dilated MultiActivation Encoder and Dilated MultiActivation decoder as in Fig. 2. Dilated MultiActivation Encoder module consists of six blocks with different dilation factors and filters to analysis and track relation between audio patterns. Encoder module able to cover more information from patterns depending on dilation factor without pooling which reduces the size of the matrix that are taken as input to the next convolution layer, this will cause missing some audio features and maybe has bad effect on the model performance. Each block consists of three layers, first one is CNN with dilation factor that different in each block to expand the covered area of the input features, second one takes the resulted features from the first layer as an input then apply CNN with the same dilation factor, number of filters of the first layer and uses activation function called activation 1 and finally, the third layer takes the resulted features of first layer as input then applies CNN with activation function named activation 2 where the filter size in second and third layers is s3*S4 as in Fig. 3. Each block has different activations, RELU and element wise multiplication of resulted data from nonlinear output of activation1 and activation2 as experiments results [18] proved that the non-linearity activation work better than RELU activation in dealing with audio signals. Finally, the element wise multiplication of the second and third layers is taken as input to the next block. In the Dilated MultiActivation decoder, the six blocks of the encoder are reversed to build the decoder module. There are different number of filters in each block with kernel size s1*s2 starting with reduction factor changes according to this range [N1, N2, N3, N4, N5, N6] and dilation factor changes according to the following range [D1, D2, D3, D4, D5, D6]. We used dilated convolutional layer in our proposed model as it could expand the covered area in the input audio without pooling. We controlled the covered area in the input by changing the value of the dilation factor. Dilated autoencoder support us in dealing with speech and discover relation between background sounds and target sounds without needing to increase kernel parameters or reducing size of the input features.

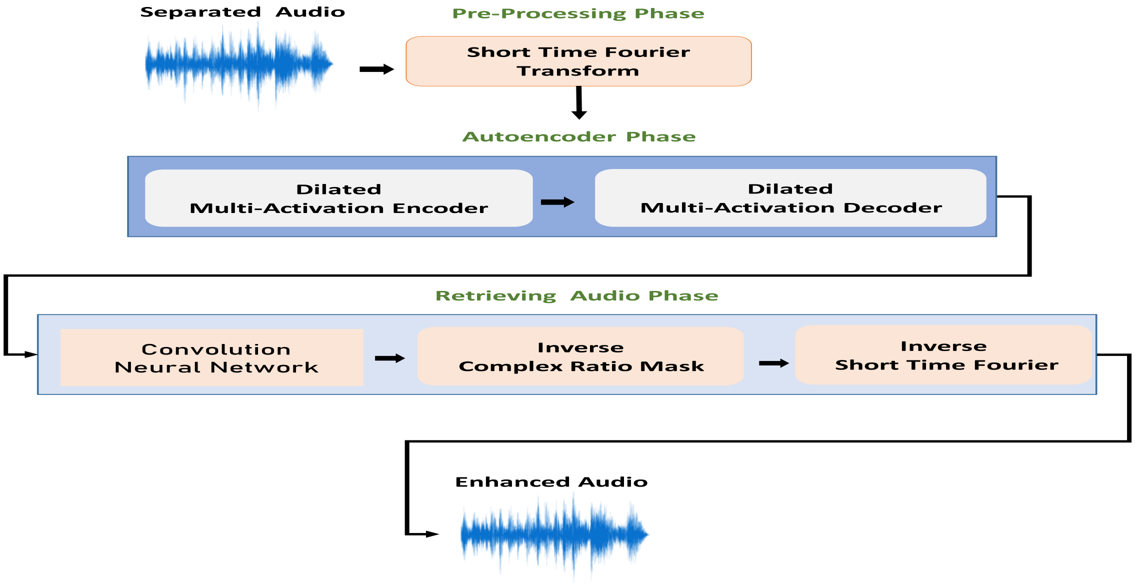
Separated Audio
Pre-Processing Phase
Short Time Fourier Transform
Autoencoder Phase
Dilated Multi-Activation Encoder
Dilated Multi-Activation Decoder
Retrieving Audio Phase
Convolution Neural Network
Inverse Complex Ratio Mask
Inverse Short Time Fourier
Enhanced Audio

Fig. 1. The proposed model of speech enhancement

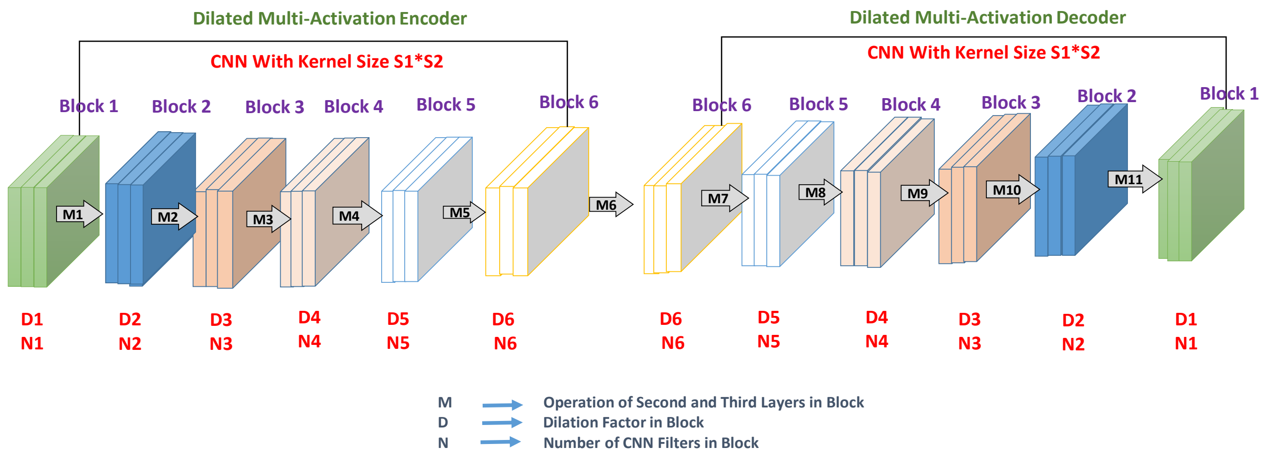
Dilated Multi-Activation Encoder
CNN With Kernel Size S1*S2
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
M1
M2
M3
M4
M5
M6
D1 N1
D2 N2
D3 N3
D4 N4
D5 N5
D6 N6
Dilated Multi-Activation Decoder
CNN With Kernel Size S1*S2
Block 6
Block 5
Block 4
Block 3
Block 2
Block 1
M7
M8
M9
M10
M11
D6 N6
D5 N5
D4 N4
D3 N3
D2 N2
D1 N1
M → Operation of Second and Third Layers in Block
D → Dilation Factor in Block
N → Number of CNN Filters in Block

Fig. 2. Dilated multi-activation autoencoder

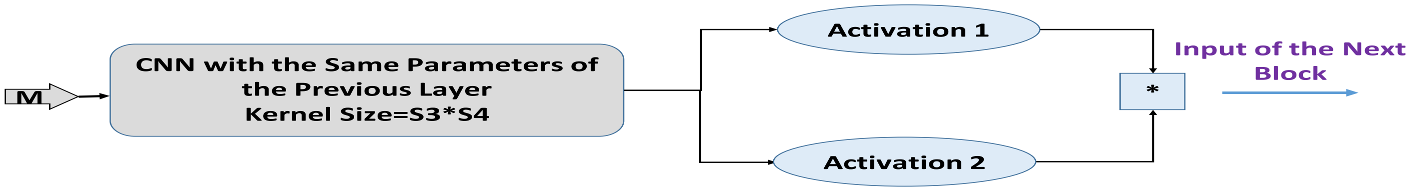
M
CNN with the Same Parameters of the Previous Layer Kernel Size=S3*S4
Activation 1
Activation 2
*
Input of the Next Block

Fig. 3. Second and third layers of autoencoder's blocks

In the retrieving audio phase, CNN layer is applied on the resulted features from decoder, then inverse complex ratio mask is calculated on the resulted features from CNN [13] after rescaling it by using (3) and finally inverse short time Fourier transform (ISTFT) is used to retrieve the enhanced version of separated audios.

$$S(Resulted\ Mask) = \frac{1}{0.1}\log(\frac{10-Resulted\ Mask}{10+Resulted\ Mask} \qquad (3)$$

where, S is the scaled complex ratio mask.

## IV. Experiments and Results

We introduce a small-scale audio dataset containing speech audio with no interfering background signals taking from separated audio of our previous work [23]. The audio segments have length between 3 and 10 seconds The enhanced audio dataset which is used to train our network contains 31250 audio files resulted from separation process [23] that are splitted into 28126 for training and 3124 for testing. Fig. 4 shows Mixed signal of two speakers in time domain and corresponding spectrogram and amplitude spectrum. It also shows magnitude spectrum and spectrogram of speaker 1's sound and speaker 2's sound before mixing them

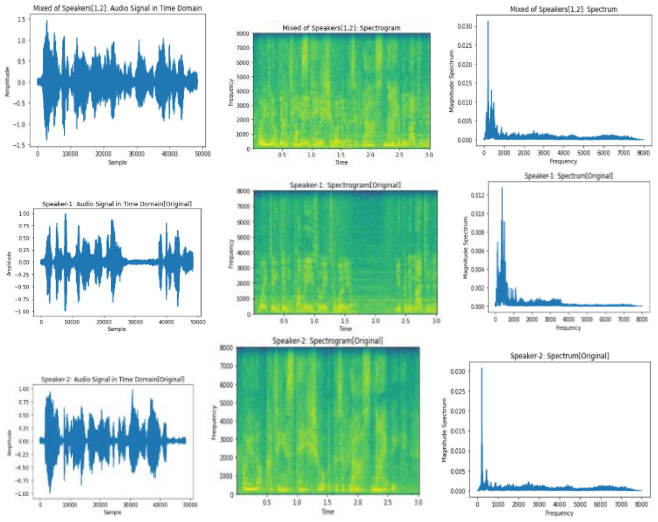

Fig. 4. Original signal in time domain, spectrogram and magnitude spectrum [mixed signal, speaker 1 and speaker2]

We implemented Dilated MultiActivation AutoEncoder in keras. In case of training, we use a batch size of 2 samples and Adam optimizer for 14062 batches with a learning rate $1X10^{-5}$. Tables I and II indicate Hyper-parameters that used to complete the process of training. The nonlinearity activation function is a combination of Tanh and Sigmoid. The number of filters in the retrieving phase is two. Number of filters in the autoencider's blocks starting from 256 to 4 with decreasing factor equal half and dilation factor starting from 64 to 1 with decreasing factor half. We used STFT as a pre-processioning phase on the input audio where STFT is calculated using a Hann window with length 25ms, hop length of 10ms, and FFT size of 512.

TABLE I. TRAINING HYPER-PARAMETERS OF DILATED MULTIACTIVATIO AUTOENCODER MODEL

| | **Value** |
|---|---|
| **Learning Rate** | $1X10^{-5}$ |
| **Optimizer** | Adam |
| **S1*S2** | 15*15 |
| **S3*S4** | 3*3 |
| **Activation 1** | Tanh |
| **Activation 2** | Sigmoid |
| **CNN Filters of Retrieving audio phase** | Number of filters[2] |

TABLE II. TRAINING HYPER-PARAMETERS OF MULTI-ACTIVATION ENCODER AND DECODER BLOCKS

| | **Number of Filters** | **Dilation Factor** |
|---|---|---|
| **Block 1** | N1 =256 | D1=64 |
| **Block 2** | N2 =128 | D2=32 |
| **Block 3** | N3 =64 | D3=16 |
| **Block 4** | N4 =32 | D4=4 |
| **Block 5** | N5 =8 | D5=2 |
| **Block 6** | N6 =4 | D6 =1 |

The proposed model takes STFT of separated audios resulted from separation model [23] as input and its output is enhanced version of the target speaker by removing background sounds of other speakers. Fig. 5 shows spectrograms and spectrums of two speaker resulted from separation model [23] before enhancement process where the separation models take mixed signal as input and generated separated sound relative to each speaker by using facial embedding of each speaker as a guide to complete the separation.

To test the performance of our model, we held four experiments, three from literature papers and the last one depending on our proposed model. All experiments use the same dataset resulted from the audio separation models.

In the first experiment, advanced version from DNN [21] is trained to reduce background sounds. It consists of a pipeline of dense layers [24]. The pipeline consists of three dense layers followed by dropout and the previous architecture is repeated two times. Fig. 6 indicates the magnitude spectrum and spectrogram of the enhanced audio after taking the low quality speeches resulted from separation models as input.

In the second experiment, we train u-net [22] to suppress background sounds. Firstly, three convolution layers with filters [128,128,1] is applied on the STFT values to prepare U-net input [25]. It is considered as deep convolutional autoencoder with skip connection between encoder and decoder parts. They adopted their network to de-noise voices and we used the same architecture to enhance audio signal resulted from separation model. After training, u-net could have better spectrogram and magnitude spectrum comparing with results of separation model as in Fig. 7. In the third experiment, we trained Wavenet [18] architecture to enhance speech of the target speaker. Fig. 8 shows the enhanced magnitude spectrum and spectrogram of the target speakers after applying wavenet architecture on separated audio it has better results comparing with DNN as it depends mainly on dilated convolutional neural network. In the last experiment, our proposed multi-activation AuteEncoder is trained to complete the process of enhancement, Fig. 9 shows the enhanced magnitude spectrum and spectrogram resulted from our proposed model that is very near to original spectrogram and magnitude spectrum.

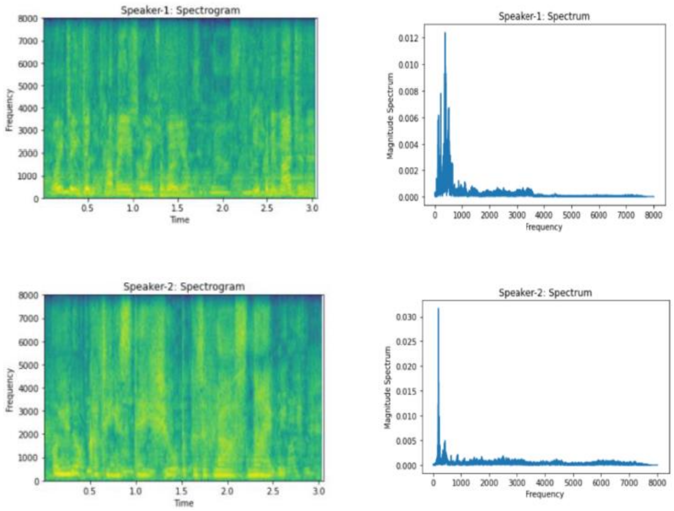

Fig. 5. Speech separation of two speakers [spectrogram and spectrum]

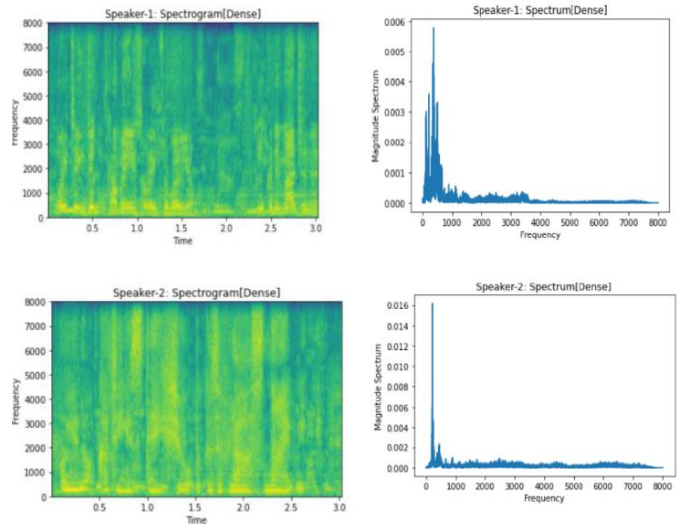

Fig. 6. Speech enhancement using DNN [spectrogram and spectrum]

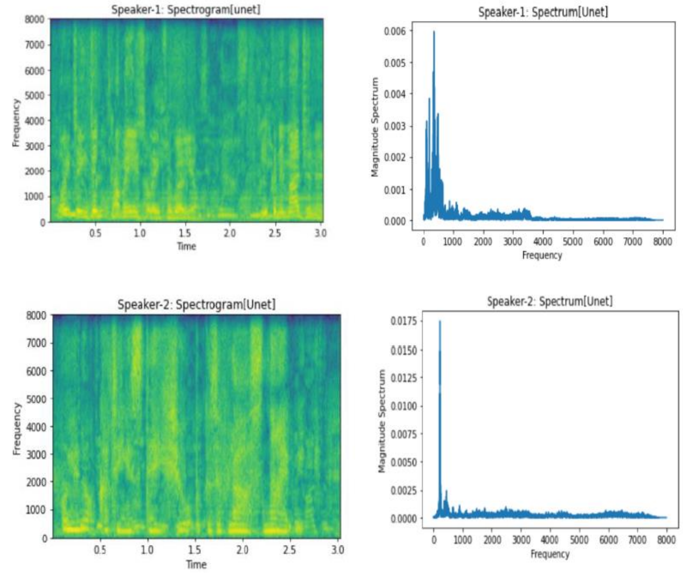

Fig. 7. Speech enhancement using U-net [spectrogram and spectrum]

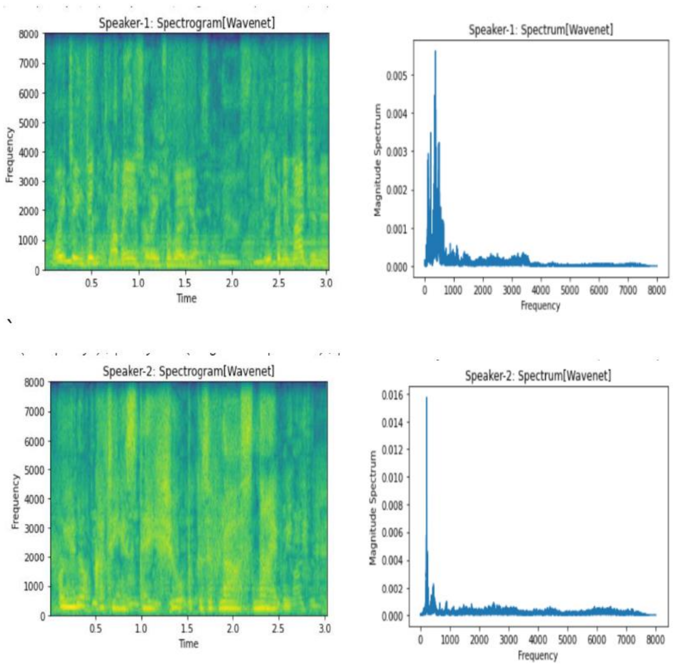

Fig. 8. Speech enhancement using wavenet [spectrogram and spectrum]

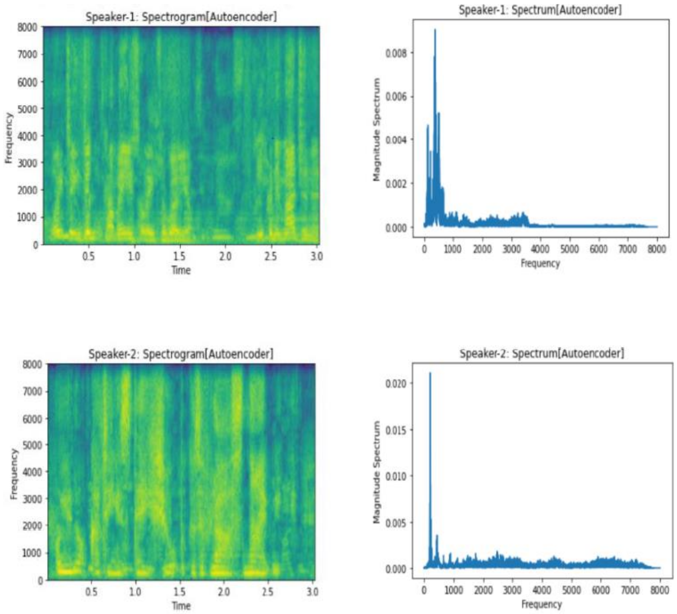

Fig. 9. Speech enhancement using autoencoder [spectrogram and spectrum]

Fig. 5-9 show that the proposed model has best spectrum and spectrogram compared with all experiments that is very near to the original version as it is mainly depends on dilated convolution with different dilation factor in each block to analysis input signal in different level and it does not use pooling in its architecture so there is no missing information during the training process

Fig. 10 illustrates the average testing loss during the process of learning. It is clear that dilated multi-activation Autoencoder experiment has best testing loss comparing with DNN, U-NET and Wavenet.

DNN has loss equal to 0.0991 after the second epoch, U-net [experiment 2]. has loss equal to 0.099 that is very near to DNN Wavenet [experiment 3] has loss equal to 0.095 that less than DNN and U-NET. Dilated multi-activation autoencoder has the minimum loss compared with DNN, U-NET and Wavenet that equal to 0.89.

The number of trainable parameters relative to each experiments is summarized in Fig. 11. It is cleared that the dilated autoencoder has an acceptable number of parameters that very near to Wavenet, u-net and less than DNN. Although DNN consists of high number of dense layers and has very high parameters, it has low performance comparing with our proposed model. Wavenet has the minimum number of parameters but its performance is lower than our proposed model.

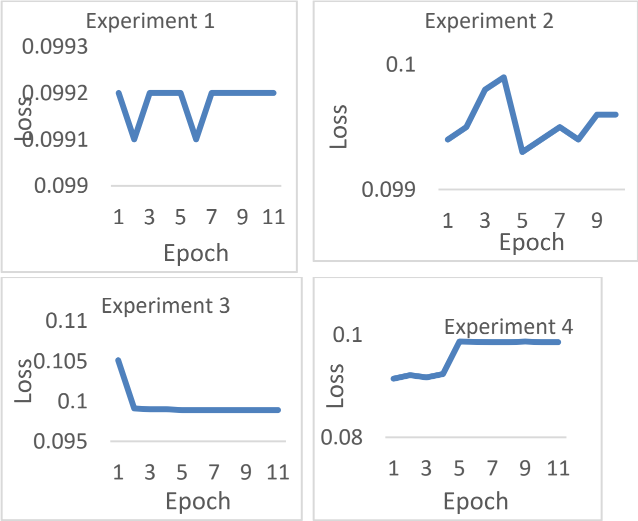

Fig. 10. Average testing loss of enhancement models

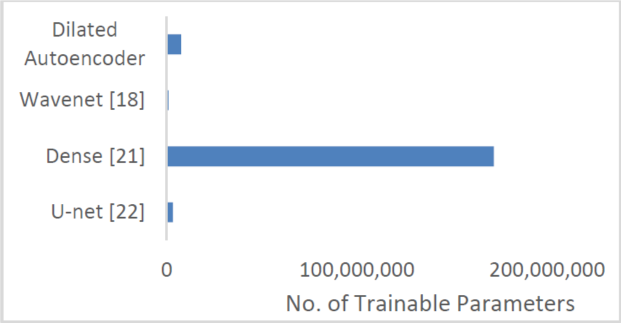

Fig. 11. Trainable parameters of enhancement models

To evaluate objective performance of the proposed model, supporter Python Speech Enhancement Performance Measures (Quality and Intelligibility) project [26] is used. Four speech objective evaluation metrics are calculated to measure speech quality [Segmental Sign al-to-Noise Ratio (SNRseg), PESQ, bark spectral distortion (BSD) and STOI].

Table III and Fig. 12 and 13 indicated that the proposed autoencoder outperform most of enhancement models in SNRseg, STOI and BSD.

It improved SNRseg of separation audio with 33.9%, STOI with 1.3% and reduced BSD with 97%, it is cleared that our proposed model has the best SNRseg, STOI and BSD comparing with audio resulted from separation and other enhancing models. Table III proved that audio files resulted from separation models has high noise and also high distortion comparing with enhanced version resulted from our proposed model that trained on 31250 audio files divided into 28126 for training and 3124 for testing.

TABLE III. OBJECTIVE EVALUATION MEASURE OF SPEECH ENHANCEMENT MODELS

| **Model/Measure** | **SNRseg** | **Stoi** | **Pesq** | **BSD** |
|---|---|---|---|---|
| Separation model[23] | 3.63 | 0.80 | 2.17 | 4,329.90 |
| U-net[22] | 2.87 | 0.80 | 2.17 | 218.90 |
| DNN [21] | 2.81 | 0.80 | 2.13 | 231.69 |
| WaveNet[18] | 2.71 | 0.80 | 2.15 | 212.62 |
| Dilated AutoEncoder | 4.86 | 0.81 | 1.95 | 128.89 |

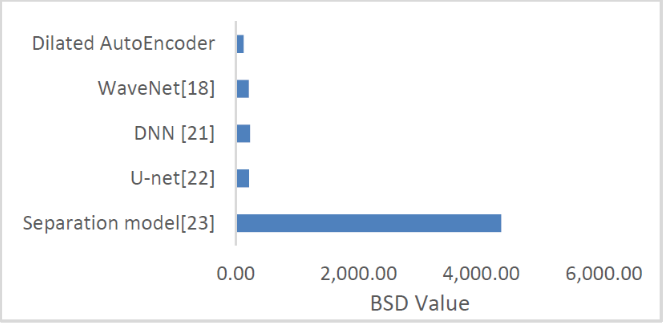

Fig. 12. Bark spectral distortion

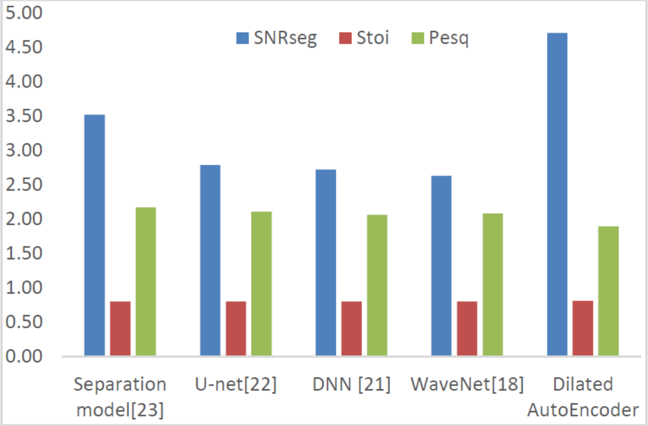

Fig. 13. Objective evaluation measures [SNRseg, Stoi, Pesq]

## V. CONCLUSION

We proposed dilated multi-activation autoencoder to enhance the performance of audio sounds resulted from separation models. It consists of two main modules, dilated multi-Activation encoder and dilated multi-Activation decoder where dilated multi-Activation encoder module has six convolutional neural network blocks with activation functions and different dilation factors. To build the structure of the decoder, the six blocks of the encoder are arranged in reverse order. Dataset consist of 31250 files splitted into training and testing sets where training set consists of 28126 files and testing set consists of 3124. The proposed model improved SNRseg of separated audios with 33.9%, STOI with 1.3% and reduced BSD with 97%. In the future, we will try to improve the performance of the proposed model by adding face embedding features relative to target speaker.

## REFERENCES

[1] Amol Chaudhari; S. B. Dhonde,'A review on speech enhancement technique', Proceedings in 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 8-11 January 2015

[2] Serim Park and Jin Won Lee,' A Fully Convolutional Neural Network for Speech Enhancement', Proceedings in Interspeech, Dublin, Ireland August 2017.

[3] DeLiang Wang and Jitong Chen.' Supervised Speech Separation Based on Deep Learning: An Overview,IEEE/ACM Transactions on Audio, Speech, and Language Processing , Volume: 26, Issue: 10, October 2018.

[4] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, Robust Automatic Speech Recognition: A Bridge to Practical Applications, 1st ed. Academic Press, 2015.

[5] B. Li, Y. Tsao, and K. C. Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition," in Proc. INTERSPEECH, 2013, pp. 3002–3006

[6] T. Venema, Compression for Clinicians, 2nd ed. Thomson Delmar Learning, 2006, chapter. 7.

[7] H. Levitt, "Noise reduction in hearing aids: an overview," J. Rehab. Res. Dev., 2001, vol. 38, no. 1, pp.111–121.

[8] A. Chern, Y. H. Lai, Y.-P. Chang, Yu Tsao, R. Y. Chang, and H.-W. Chang, "A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom," IEEE Access, 2017.

[9] A. El-Solh, A. Cuhadar, and R. A. Goubran. "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in Proc. ISMW, 2007, pp. 235–239.

[10] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in Proc. ICSLP, vol. 2, 1996, pp. 929–932.

[11] Emad M. Grais, Hagen Wierstorf, Dominic Ward, and Mark D. Plumbley, 'MULTI-RESOLUTION FULLY CONVOLUTIONAL NEURAL NETWORKS FOR MONAURAL AUDIO SOURCE SEPARATION', Proceedings in 13th International conference on Latent Variable Analysis and Independent Component Analysis, Grenoble, France, 21-23 February 2017

[12] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao 'Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks, IEEE Transactions on Emerging Topics in Computational Intelligence, Volume 2, Number 1, February 2018

[13] Jinuk Park, Jaeseok Kim, Heejin Choi, Minsoo Hahn,'Convolutional Recurrent Neural Network Based Deep Clustering for 2-speaker Separation', Proceeding in the 2018 2nd International Conference on Mechatronics Systems and Control Engineering, Amsterdam, Netherlands, February 2018

[14] Emad M. Grais and Mark D. Plumbley,' SINGLE CHANNEL AUDIO SOURCE SEPARATION USING CONVOLUTIONAL DENOISING, Proceedings in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14-16 November 2017

[15] Yi Luo and Nima Mesgarani,' Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network', Proceeding in Interspeech, , Hyderabad ,India, 2-6 September 2018.

[16] Jitong Che, and DeLiang Wang,' Long Short-Term Memory for Speaker Generalization in Supervised Speech Separation', The Journal of the **Acoustical Society of America Volume 41, Issue 4705, 2017**

[17] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu,' WaveNet: A Generative Model for Raw Audio', arXiv:1609.03499v2 [cs.SD,] 19 Sep 2016

[18] Dario Rethage, Jordi Pons, and Xavier Serra, A WAVENET FOR SPEECH DENOISING, Proceedings in IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15-20, 2018.

[19] Yi Luo and Nima Mesgarani,' Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation', IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume 27 Issue 8, Page 1256-1266 August 2019

[20] Craig Macartney and Tillman Weyde,' Improved Speech Enhancement with the Wave-U-Net', arXiv:1811.11307v1 [cs.SD] 27 Nov 2018

[21] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, Fellow,' An Experimental Study on Speech Enhancement Based on Deep Neural Networks', IEEE SIGNAL PROCESSING LETTERS, VOL. 21, NO. 1, JANUARY 2014

[22] Andreas Jansson, Eric Humphrey, Nicola Montecchio , Rachel Bittner , Aparna Kumar ,and Tillman Weyde,' SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL NETWORKS', Proceedings in the 18th ISMIR Conference, Suzhou, China, October 23-27, 2017

[23] Ghada Dahy, Mohammed A.A. Refaey, Reda Alkhoribi, M. Shoman.' A speech separation system in video sequence using dilated inception network and U-Net, Egyptian Informatics Journal, Vol 23, Issue 4, December 2022.

[24] https://github.com/boozyguo/ClearWave last accessed: 5/10/2022

[25] https://github.com/vbelz/Speech-enhancement last accessed:10/9/2022

[26] https://github.com/schmiph2/pysepm last accessed: 1/10/2022