2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)
Military Institute of Science and Technology (MIST), Dhaka-1216, Bangladesh

2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) | 978-1-6654-9522-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEEICT53905.2021.9667875

A Self-Supervised Convolutional Neural Network Approach for Speech Enhancement

Nursadul Mamun[1], Sharmin Majumder[2], Khadija Akter[3]
[1,2]Department of Electrical Engineering, University of Texas at Dallas, USA
[3]Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology, Bangladesh
[1]nursad49@gmail.com, [2] jouty_eee@yahoo.com and, [3] khadija.ete@cuet.ac.bd

***Abstract*— Enhancement of speech means modification to the speech which is degraded by noise. Speech enhancement leads to improvement in the intelligibility of speech to human listeners. Deep learning techniques have drawn tremendous attention for speech enhancement in recent years which require clean speech along with noisy speech for training purpose. However, availability of clean speech signal in naturalistic scenarios is challenging. To ameliorate it, this study proposes a deep neural network- based on speech enhancement approach without the requirement of clean speech to train the model called self-supervised learning. In the proposed framework, two CNN-based speech enhancement models have been deployed for two noisy conditions (babble noise and machinery noise). This work has been accomplished on two different datasets: IEEE speech corpus distorted with real-time noise and recorded speech signals in naturalistic environment. Experimental result demonstrates that the proposed framework achieved significant improvement in both subjective and objective measures.**

Keywords—speech enhancement, deep learning, convolutional neural network, self-supervised learning, babble noise, machinery noise.

I. Introduction

Restoration of clean speech from the degraded noisy ones is referred as speech enhancement and this enhanced speech assist in the application of automatic speech recognition (ASR), speech coding, speaker recognition, and hearing aids [1]. To model the nonlinear relationship of the corruption of speech by the noise, recently convolutional neural network (CNN) based approaches have shown promising results which outperform conventional speech enhancement methods [2]. Reinforcement learning based speech enhancement [3], Speech Enhancement Using Generative Adversarial Network (SEGAN) [4], Convolutional Auto encoder [5], and Residual Convolutional Neural Network [6] which are few examples of deep learning-based speech enhancement method. Notable two recent neural network-based approach for speech enhancement are: (1) *Supervised learnin*g which requires a huge training set [7] and (2) *Self-Supervised learning*-to relax the constraints on training data [8].

Aforementioned existing supervised learning approaches require a large number of training data containing noisy speech mixtures along with their corresponding clean speech samples to train the deep learning model [9]. However, it is not always possible to obtain clean recorded version of speech signals for real environment scenarios [7]. For paradigm, the clean signal of corresponding machinery noise is troublesome to obtain. Therefore, a model that can train and test on flying mode is required to get benefit in real scenario. To overcome the drawbacks of supervised learning, a new approach for speech enhancement, self-supervised learning, is explored which does not require paired label of training data.

In this work, a self-supervised deep neural network solution is deployed to enhance speech signal in real time environment. This study proposes a speech enhancement technique without the requirement of clean speech to train the model. In the training stage of the proposed model, the input and target are the noisy signals from a dual channel microphone, respectively. In the testing stage, noisy speech coming from naturalistic environment is sent through the trained model and output is the enhanced signal. Therefore, the proposed model is independent of the clean signal. In this study, two different models are designed for two different noisy conditions which are babble noise and machinery noise. To show the robustness of the developed approach, two commonly used objective metrics performances are considered along with the scores from subjective study.

The rest of the paper is organized as follows: In section II, the introduced deep speech enhancement approach and its implementation aspects are presented as well as an overview of the examined datasets is depicted. A comprehensive set of experimentations and their results are then described in Section III. Finally, the conclusion of this work is drawn.

II. Proposed Deep Neural Network

Figure 1 illustrates the block diagram for the proposed self-supervised deep neural approach. This figure is divided into two parts: training and testing. To train the model, two uncorrelated noises are added with the same speech signal. These two signals are captured in two channels of a mid-side microphone. Therefore, the correlation between two channels were found approximately 0.8. The signals first go through a Voice Activity Detection (VAD) to clip the non-speech parts from the audio signals. Here, preprocessing block includes

978-1-6654-9522-6/21/$31.00 ©2021 IEEE

resampling, reshaping and normalization etc. The input noisy speech was originally collected at 48 KHz and then resampled at 16 KHz. To reduce the latency, the original input signal is framed for 20 ms as the signal is quasi-stationary within 32 ms. A Hanning window of length 20 ms (with an overlap of 10 ms) is applied to the input signal. The two normalized speech signals are fed to the CNN model as input and target. After the training of the CNN model, the model is tested with new noisy speech signals and the predicted signal is the de-noised enhanced audio speech.

A. Audio Corpus

To implement the proposed self-supervised approach, this study used speech signals from standard database and speech signals collected from two different naturalistic environment (Babble noise and machinery noise), described in the following section.

Offline Database: Speech signal from IEEE data corpus were used to train and test the CNN model under different noise and different SNR conditions. The dataset consists of 1440 speech samples collected in university of Boston and divided into 72 lists (10 sentences for each set) and uttered by one male and one female speakers. Each speech sentence is about approximately 3 seconds long on average. The speakers for this experiment are from two American English regions of the Pacific Northwest (PN) and the Northern Cities (NC) reading the IEEE "Harvard" sentences.

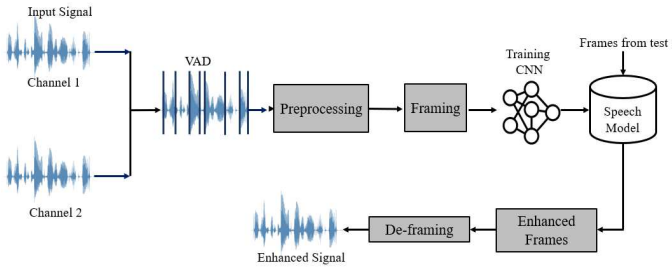


Fig. 1. Illustration of the proposed framework for the speech enhancement using self-supervised deep learning method.

The speech sample was originally recorded with a sampling frequency of 25 kHz and down sampled to 16 KHz in this study. The dataset is randomly divided into a training set and a testing set without any overlap. The 90% randomly selected samples were used to train the CNN model where rest of the speech were used to validate the trained model.

Real Time Database: This self-supervised model was designed for two types of noises: babble noise and machinery noise. Therefore, speech signals uttered in these two noisy environments are used as real time database in this work.

Babble Noisy Data: Human subjects with normal hearing face difficulties in conversation while they are in cafeteria. Therefore, this study considered babble noise recorded in the cafeteria located at the student union of University of Texas at Dallas (UTD). For this study, data have been collected in two sessions. In session I, speech signals (120 samples) from IEEE database were uttered for 15 mins in babble noisy environment (UTD student union) by 2 subjects (one male and one female) and recorded using a mid-side microphone (STEREO TASCAM TM-ST1). The noisy speech signal was recorded using MATLAB at a sampling frequency of 48 kHz and then down sampled to 16 kHz for the study. A part of the real time noisy speech signal corrupted with babble noise with their corresponding spectrogram is presented in Fig. 2 for both channels.

In session II, only babble noise was recorded from the student union for an hour at a sampling frequency of 48 kHz and resampled to 16 kHz. The main goal of this was to distort IEEE database in offline with the real time babble noise which is required for the training session.

Machinery Noisy Data: Machinery noises were recorded in two different scenarios, university local bus and cluster room. Similar to the babble noise, sentences from IEEE data were uttered in a cluster room situated in the UTD-CRSS lab at a sampling frequency of 48 kHz and down sampled at 16 kHz. In addition, cluster noise was recorded for one hour in the same environment. University bus noise was also recorded as machinery noise for one hour.

B. Architecture of CNN Model

Two CNN models are designed in this study for removing two types of noises from the speech signal: Babble noise adapted model and Machinery noise adapted model. Both models conduct speech enhancement for different SNR values. The performance of the models is then evaluated using two objective metrics which are short-time objective measure (STOI) and PESQ. To quantify the effectiveness of the proposed architecture subjective studies have been accomplished for the enhanced speech signal from both models.

To train the proposed model, the noisy dataset developed for two different channels (one hour for each channel) are used as target and predictor. The architecture of the proposed model consists of 5 convolutional layers (4 hidden layers and 1 output layer). Each of the CNN layer (except output) consists of 80 kernels with size of [30 1]. The output layer consists of 1 kernel with size of [1 1]. The Adam optimization algorithm is used to train the network and the size of the mini batch and initial learning rate is set to 100 and 0.004, respectively. Finally, 'tanh' and 'linear' activation function is used for the intermediate and output layers, respectively. The root mean square error (RMSE) and loss were recorded to optimize the weight for the network architecture.

To test the model, two different types of data are used in this study. For first part of the testing, 40 speech signals (uttered by 1 male and 1 female) from IEEE database (different from training samples) are distorted with the Babble and machinery noise for different SNRs.
The preprocess signal is then sent frame-by-frame to the trained model. The enhanced signal is then reconstructed

from the enhanced frame which is the output of the CNN network. The models is then evaluated using two different objective metrics, STOI (for intelligibility) and PESQ (for quality). For the second part of the testing, the noisy speech signal recorded from the naturalistic environment are sent to the selected (best model selected in the first part) CNN network. Finally, de-noised speech signal is reconstructed in the similar procedure of the first part.

TABLE I. Objective Speech Intelligibility Scores

Noise Type	Metric	Noisy		Enhanced	
		5 dB	8 dB	5 dB	8 dB
Babble	STOI	0.817	0.8675	0.876	0.9013
	PESQ	1.879	2.071	2.14	2.357
Machinery	STOI	0.8183	0.8701	0.8732	0.9013
	PESQ	1.7446	1.9609	2.1772	2.4382

C. Objective Metrics

Several speech intelligibility and quality metrics have been proposed in recent years to quantify the speech intelligibility in noisy environment [10, 11, 12, 13]. To quantify the speech intelligibility and quality score for the noisy and enhanced speech, this study used one objective speech intelligibility metric, short-time objective measure (STOI) and one objective speech quality metric, PESQ.

Short-time objective measure (STOI): Taal and colleagues developed the short-time objective measure (STOI) [10]. STOI is a function of clean and degraded speech that decomposes both speech signals into DFT-based, 1/3 octave bands. The STOI score varies from 0 to 1, where 1 represents the highest and 0 represents the lowest intelligibility, respectively.

Perceptual Evaluation of Speech Quality (PESQ): PESQ is a full reference metric that compares two signals before and after passing through a communications channel to predict speech quality. The signals are time aligned, followed by a quality calculation based on a psychophysical representation. Quality is scored in a range of -0.5 to 4.

III. Experimental Results and Analysis

This section provides simulation results of the proposed two CNN architectures considering for babble and machinery noises using two different datasets. The performance evaluation using two different objective metrics is presented for the two CNN architectures. Also, the subjective evaluation of the two developed models regarding for two noisy conditions is demonstrated.

A. Speech Enhancement for Babble and Machinery Noise

Figure 2 depicts the spectrograms for noisy and enhanced speech signal for the developed CNN model regarding babble and machinery noise. The time domain signal and spectrogram show that, the proposed model effectively remove the unwanted noise from the noisy speech signal without scarifying the quality and intelligibility of the speech signal.

B. Objective Speech Intelligibility and Quality Measurement

To examine the performance of the proposed system for babble noise, STOI and PESQ metrics are used to calculate the objective speech intelligibility and quality, respectively and presented in Table I. The result shows that, the objective speech intelligibility score for enhanced signal is higher than that for the noisy signal for all individual samples.

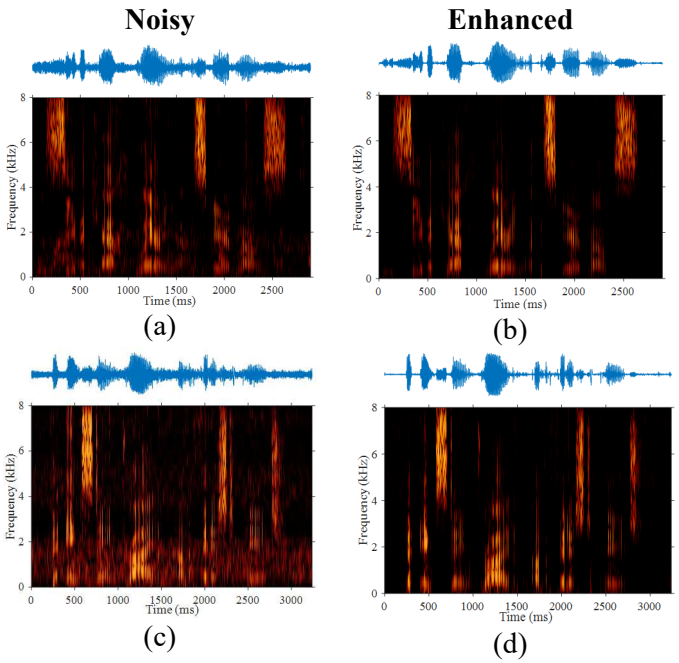


Fig. 2. Top panel: The spectrogram of the signal with babble noise and corresponding enhanced speech. Bottom panel: The spectrogram of the signal with machinery noise and corresponding enhanced signal.

The average STOI score for noisy and enhanced speech is found as 0.817 and 0.876, respectively. Similarly, the objective speech quality score for noisy and enhanced speech was found as 1.879 and 2.14, respectively.

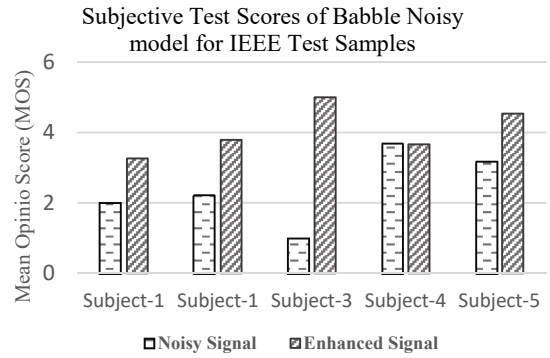
C. Subjective Evaluation

This section presents the subjective evaluation of the proposed model considering to the babble noise using two different datasets.

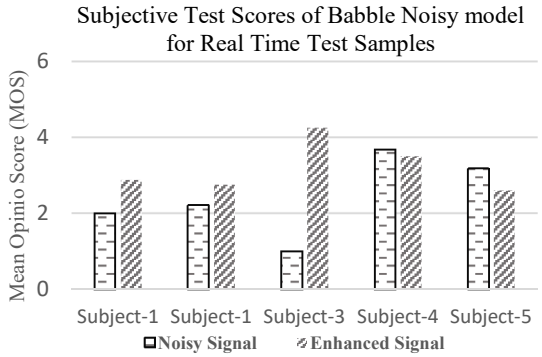
Figure 3 presents the subjective score for the proposed speech enhancement algorithm. Speech signals are used from two different sets of data. For first study, 15 IEEE sentences were distorted with the real time babble noise and presented to the different subjects. In the second phase of the study, eight speech signals recorded in real time environment (babble noise) were used. Total five subjects were participated in this study where four of them were male and one of them was female with the age between 25 and 30 years. In both phase, noisy and enhanced speech signals were presented to the subject and asked to rate both signal within a parameter of 1 to 5, indicating 1 as lowest and 5 as the highest quality. The mean opinion score (MOS) over the dataset was then

calculated and presented for five different subjects. Figure 3 (a) represents the MOS score for IEEE test samples and 3 (b) represents the MOS score for real time test samples for both noisy and enhanced speech. In general, the MOS score for enhanced speech is higher than that of noisy speech over all subjects. In addition, subject 3 preferred enhanced speech with highest accuracy where subject 4 has equal probability of selecting enhanced and noisy signal.

The result of paired preference test is also presented in Fig. 4 using both IEEE sentences and real time speech corpus. The result shows that, subjects preferred enhanced signal over 90% case when its speech data comes from IEEE database where, it becomes 78% when it was tested for real time speech samples.



(a)



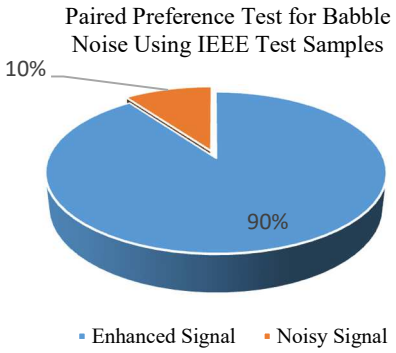
(b)

Fig. 3: Mean opinion score (MOS) for different subjects using noisy and enhanced speech for babble noise.

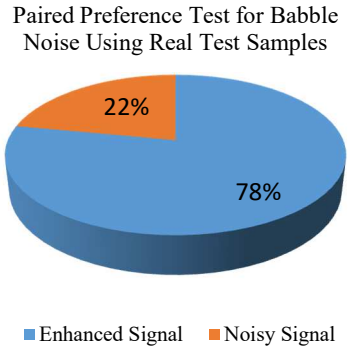
Therefore, it is suggested that the proposed CNN based speech enhancement technique effectively enhance the noisy speech under adverse conditions.

CONCLUSION

The aim of this work is to design a CNN model for the speech enhancement task without access of any clean signal in naturalistic environment. Accomplished experimentations under babble and machinery noisy conditions using two different datasets including IEEE speech corpus distorted with real time noise and recorded speech signals from naturalistic environment have been evaluated for the performance analysis of the proposed method. In addition to quantify the effectiveness of the developed model, two objective speech intelligibility and quality metrics have been conducted here. The subjective evaluation (paired preference test) for the enhanced speech was also performed. Observing in all conditions, it can be concluded that the proposed CNN model is viable in handling the reconstruction of enhanced speech from the degraded noisy signal without any requirement of clean speech signals.



(a)



(b)

Fig. 4: Paired preference test using noisy and enhanced speech for babble noise. (a) Sample from IEEE database (b) Real time audio signal.

REFERENCES

[1] P. C. Loizou, Speech enhancement: theory and practice: CRC press, 2007.

[2] Mamun, N., S. Khorram, and J. H. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipients," arXiv preprint arXiv:1907.02526, 2019.

[3] Saleem, Nasir, Muhammad Irfan, Xuhui Chen, and Muhammad Ali. "Deep Neural Network based Supervised Speech Enhancement in Speech-Babble Noise." In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp. 871-874. IEEE, 2018.

[4] Shen, Yih-Liang, Chao-Yuan Huang, Syu-Siang Wang, Yu Tsao, Hsin-Min Wang, and Tai-Shih Chi. "Reinforcement learning based speech enhancement for robust speech recognition." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6750-6754. IEEE, 2019.

[5] Tanaka, H., Sugiura, Y., Yasui, N., Shimamura, T., & Miyazaki, R. (2019). Performance improvement of speech enhancement network by multitask learning including noise information. *IEICE Technical Report; IEICE Tech. Rep.*, *119*(334), 31-36.

[6] Park, S. R., & Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.

[7] Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee. "A regression approach to speech enhancement based on deep neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, no. 1 (2014): 7-19.

[8] Mashiana, H. S., Salaria, A., & Kaur, K. (2019, November). Speech Enhancement Using Residual Convolutional Neural Network. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1193-1196). IEEE.

[9] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in Interspeech, 2016, pp. 3743-3747.

[10] Taal, Cees H., Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech." IEEE Transactions on Audio, Speech, and Language Processing 19, no. 7 (2011): 2125-2136.

[11] Mamun, Nursadul, Wissam A. Jassim, and Muhammad SA Zilany. "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.4 (2015): 760-773.

[12] Mamun, N., et al. "An intrusive method for estimating speech intelligibility from noisy and distorted signals." *The Journal of the Acoustical Society of America* 150.3 (2021): 1762-1762.

[13] Rix, Antony W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE, 2001.