Speech Enhancement using Fully Convolutional UNET and Gated Convolutional Neural Network

Danish Baloch[1], Sidrah Abdullah[2], Asma Qaiser[3], Saad Ahmed[4], Faiza Nasim[5], Mehreen Kanwal[6]
Department of Computer Science, DHA Suffa University, Karachi, Pakistan[1]
Department of Computer Science and Information Technology,
NED University of Engineering & Technology, Karachi, Pakistan[2, 5]
Department of Computer Science, IQRA University, Karachi, Pakistan[3, 4]
MS Fast University, Department of Computer Science, Pakistan[6]

***Abstract*—Speech Enhancement aims to enhance audio intelligibility by reducing background noises that often degrade the quality and intelligibility of speech. This paper brings forward a deep learning approach for suppressing the background noise from the speaker's voice. Noise is a complex nonlinear function, so classical techniques such as Spectral Subtraction and Wiener filter approaches are not the best for non-stationary noise removal. The audio signal was processed in the raw audio waveform to incorporate an end-to-end speech enhancement approach. The proposed model's architecture is a 1-D Fully Convolutional Encoder-to-Decoder Gated Convolutional Neural Network (CNN). The model takes the simulated noisy signal and generates its clean representation. The proposed model is optimized on spectral and time domains. To minimize the error among time and spectral magnitudes, L1 loss is used. The model is generative, denoising English language speakers, and capable of denoising Urdu language speech when provided. In contrast, the model is trained exclusively on the English language. Experimental results show that it can generate a clean representation of a clean signal directly from a noisy signal when trained on samples of the Valentini dataset. On objective measures such as PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility), the performance evaluation of the research outcome has been conducted. This system can be used with recorded videos and as a preprocessor for voice assistants like Alexa, and Siri, sending clear and clean instructions to the device.**

***Keywords*—Speech enhancement; speech denoising; deep neural network; raw waveform; fully convolutional neural network; gated linear unit**

I. Introduction

Speech Enhancement has been a topic of interest for five decades. Speech enhancement aims to improve speech quality (reducing background noise) by various algorithms [1]. The purpose of enhancement is to enhance the intelligibility of the speech signal degraded by the noise using audio signal processing techniques. The conventional methods used for noise reduction are Spectral subtraction and the Wiener filter [2] and [3]. Still, both approaches leave musical artifacts in synthesized speech [4], need multiple sources as noise profile information, and distort the desired output.

Deep Learning approaches can overcome the pitfalls of conventional approaches because these systems can learn to map between complex nonlinear functions [5]. In addition, they have the ability to produce desirable outputs that can be used to decrease the Word Error Rate (WER) of automatic speech recognition (ASR) systems [6], boost the performance of speech-to-text systems [7], and in general, increase the intelligibility of speech which can be beneficial for any system whose performance is dependent on the intelligibility of speech. In Deep Learning, the classical approach to suppress noise through the signal is mask-based signal denoising [8], in which DNN models produce a TF mask that filters out the noise and leaves the speech. Mask-based approaches are mostly done on magnitude spectrograms of audio [9], [10]; this creates a challenge of reconstructing the audio again to the time domain once it is filtered using the predicted spectrogram mask and reconstruction of audio is heavily dependent on the phase of noisy input audio.

Another investigated approach is a mapping-based approach where a representation of a complex nonlinear noisy signal is directly mapped onto a clean signal [11], [12] and [13]. Mapping-based approaches directly map noisy signals to their clean representations. Due to the fast variation of amplitudes in raw audio waveforms, mapping-based approaches are based on STFT (short-time Fourier Transform) of audio.

A. Proposed Approach

Our proposed approach is a mapping-based approach in raw audio waveform (time-domain). The loss function is optimized for time and STFT of audio. This approach eliminates the requirement of reconstruction of the audio from the spectrogram output into raw audible audio waveform as in [11] and [12], rather it generates the audible enhanced speech output directly. The magnitude spectrogram of audio is incorporated inside of the loss function rather than as input to the model as in [9] and [13], which gives us leverage to do speech enhancement on raw audio waveform directly. Given an audio, our system directly generates its clean representation without any additional post-processing on the output of the model. The proposed approach focuses on enhancing the speech and suppressing the noise in audio sampled at 22.05 KHz. To achieve this U-Net architecture is used. The choice of this architecture is due to the fact that it takes audio as raw waveform without any manual feature extraction and provides output also in the raw audio waveform, which can be converted to mp3 file and can be saved on disk directly. It consists of convolutional layers and a middle layer which is a

bottleneck. The middle layer (bottleneck layer) represents the data in encoded representation which is then decoded by the Decoder architecture connected to the bottleneck layer.

The objectives of this research are two-fold:

1) Discusses the common approach in deep learning, specifically mask-based signal denoising, where deep neural network (DNN) models generate a time-frequency (TF) mask to filter out noise from the audio signal.

2) Reviews mapping-based approaches that directly map complex nonlinear noisy signals onto clean representations. It distinguishes these approaches from mask-based methods and notes that mapping-based methods typically rely on the short-time Fourier transform (STFT) of audio due to the fast variation of amplitudes in raw audio waveforms.

3) Introduces a novel mapping-based approach specifically applied to raw audio waveforms in the time domain

B. Research Distribution

Section II of this paper covers related work in speech enhancement, followed by Section III, which discusses the dataset used for this research. Section III also consists of discussion on the U-Net Architecture. Section IV covers model training, and Section V, the last section discusses results and concludes the paper.

II. Related Work

Speech enhancement research has been ongoing for the last half-century. Earlier, classical linear noise filtering approaches were used for reducing noise. Two notable examples are spectral subtraction and wiener filter approaches [2] & [3]. The former needed multiple sources and works average with static noise and below the bar with non-stationary noises. The latter had its pitfalls, such as it required two sources; one of them is a mixed signal, and the other is the background sound signal. With the rise of deep learning, these pitfalls were eliminated as deep learning made it easier to notably reduce the noise in the noisy speech samples. This approach made no assumptions regarding the statistical attributes of the signals and used a wide variety of noise types to provide a variety of noisy speech samples for training [14]. Moreover, these systems can learn to map between complex nonlinear functions. They can produce desirable outputs that can be used to decrease the Word error rate (WER) of automatic speech recognition systems (ASR). In general, it can increase intelligibility, which can benefit any system whose performance depends on speech intelligibility.

In a notable work [15], a causal model was proposed based on auto-encoder architecture. They also proposed effective data augmentation techniques, frequency band masking, and reverberation. Their results suggest that the proposed system is comparable to the SOTA (State-Of-TheArt) model across all performance measures while working directly on the raw waveform. It also discovered that up-sampling the audio before feeding it into the encoder improves accuracy, and then they downsampled the outputs by the same amount.

Another innovative approach, presented in [16] proposed a new deep learning-based framework for real-time speech enhancement on dual-microphone mobiles for close-talk scenarios. They used a masking-based approach using a computationally efficient CRN (Convolutional Recurrent Neural Network), which was trained for intra-channels and inter-channels. Their experimental results showed that their proposed approach outmatched the DNN-based and other traditional methods.

Alternatively, authors in [17] used a hybrid approach using DSP techniques and deep learning for noise suppression. The deep recurrent neural network with four hidden layers was used. The resulting lower complexity made it practical to be used in video-conferencing systems. Their results showed a significant improvement in quality from deep learning, especially for non-stationary noise types.

The authors of [7] also used a hybrid approach consisting of noise estimation and speech-to-text block. This paper's focus was on spontaneous speech in the medical domain. As the medical terms used in the area are complex, and speech recognition systems often fail to recognize those words, the idea here was to propose an algorithm that resolves this issue. Non-linear spectral subtraction for noise reduction and the Hidden Markov Model (HMM) were incorporated for converting the speech to text to reduce the word error rate.

This paper [18] discussed the classical approaches for noise reduction by using filters. It also discusses stationary and non-stationary noise and its subtypes. This approach [19] combined a short-time Fourier transform (STFT) and a learned analysis and synthesis basis in a stacked-network method with less than one million parameters for real-time noise suppression. [20] an improved approach to their previous research was proposed, where the Deep Denoising Autoencoder (DAE) is trained on only clean speeches. In this paper, they trained DAE on pairs of noisy signals and clean output using a mapping-based approach stack AE approach where AE is stacked to form DAE to estimate the noise from the noisy signal. This paper [21] explores a greedy layer-wise pretraining strategy to train a DAE for speech restoration and then applies that restored speech for noisy robust speech recognition.

III. Methodology

A. Dataset

Datasets consist of pairs of corresponding audios sampled at 22.05 KHz stored in WAV format in a Linux environment. The dataset is created using two publicly available datasets. From one dataset noisy environment audio samples are obtained and from another dataset, clean human speech audio samples are obtained later these two samples of datasets are mixed together using simple arithmetic addition in order to create noisy simulated environments and their corresponding clean speech pairs.

Noise samples are from the DEMAND dataset to generate simulated noisy environments [22]. The DEMAND dataset is recorded with an array of sixteen microphones with an original sampling rate of 48kHz. It is publicly available in 48kHz or a downsampled version of 16kHz. In this paper, the

48khz version is downsampled to 22.05kHz utilizing the librosa module of Python and later used in the dataset. Three noise profiles (noise environments) are chosen from the DEMAND dataset namely DKITCHEN, PRESTO, and OMEETING. DKITCHEN includes recordings of kitchen noises while cooking. At the same time, PRESTO consists of a set of noise recordings taken from the university restaurant during lunchtime, and OMEETING, consisting of meeting room sounds during discussions from the microphone array. At first, all 16 channels of DKITCHEN, PRESTO, and OMEETING are mixed together respectively in order to create a single noise profile for each environment.

Next, the first eight channels of PRESTO and OMEETING are mixed together. This was done because it is observed that in the case of PRESTO and OMEETING with all sixteen channels added together, the noise profile was overruling the speech components in raw audio and also in spectrograms. Our proposed model takes 2.97sec windows of inputs, so eight sections of length equal to 2.97sec are used from each noise profile. Eight sections are used because for each speaker eight utterances are chosen. Later these noise sections are mixed with each speaker.

For clean speech representations, eight unique utterances of 47 notable speakers from the Valentini dataset are used [23]. Then mixed the noisy environment samples and the clean samples from the Valentini dataset, to create simulated noisy environment signals, and their correspondence clean speech representation. These pairs of audio are converted to pickle format and saved on the disk.

There are a total of five batches of data, each with 376 utterances of 47 unique speakers. Batch-1 represents the DKITCHEN (sixteen channels mixed) noise condition, Batch-2 represents the PRESTO (sixteen channels mixed) noise condition, Batch-3 represents the OMEETING (sixteen channels mixed) noise condition, Batch-4 represents PRESTO (eight channels mixed) noise condition and Batch-5 represents OMEETING (eight channels mixed) noise condition.

B. Notations and Problem Settings

Targeting speech enhancement in mono-aural signals, where $x \in R^T$ is the given signal composed of additive background noise as $n \in R^T$ and clean speech $y \in R^T$. so that $x = y + n$. The length T is of a fixed duration, which equals *65536* samples when audio is sampled at 22.05 KHz. Our main objective is to find a function *f* through the non-linear architecture of the neural network that reduces the enhancement function to $f(x) \approx y$.

In this problem set, the function *f* is the neural network architecture, producing the clean speech *y* at its output layer.

C. UNET Architecture

As presented in Fig. 1, the adopted neural network architecture is a one-dimensional UNET encoder-to-decoder architecture with skip connections [24] and gated linear units. The input shape of the model is equal to the number of 65536 samples when audio is sampled at 22.05 KHz; the output shape is also the same as this is an end-to-end approach.

Gated Linear Units [25] are incorporated in the encoding and decoding blocks of the model; there are no GLU in the bottleneck section of the model; convolution layers throughout the model are one-dimensional.

A detailed overview of the proposed Fully Convolutional Gated Encoder-to-Decoder architecture is shown in Fig. 2. The proposed model has three sections Encoder Section, Bottleneck section, and Decoder section, which is in the end is connected to the output layer producing clean speech output. Each section of the model and their connection with each other is discussed as follows.

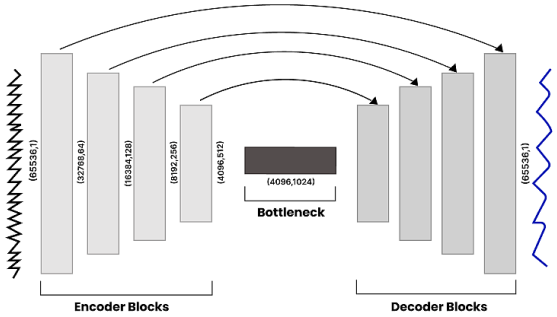

Fig. 1. Proposed fully convolutional gated encoder-to-decoder architecture with bottleneck for speech enhancement. Input and output shapes are the same.

1) Encoder section: The encoder section compresses the data of the input from a higher dimension to a lower dimension while reducing the noise from the data. The input of the encoder is the raw audio, and its output is a compressed data format that represents the input raw audio. Leaky ReLU is used as an activation function so that the model can learn to flow the gradient from most neurons. Most of the input data consists of negative values and the range of input data is [-1,1]. ReLU was also used as an activation function in order to obtain a sparse network but with ReLU, a dead ReLU problem was observed as input data is in the range of [-1, 1]. One-dimensional convolutional layers are used with batch normalization, Leaky Relu activation, and Gated linear Units are applied. The encoder section contains two layers of convolution than the max pooling layer.

2) Bottleneck section: The tanh activation function is used in the bottleneck section to apply a non-linear transformation on the signal at the most encoded layer. The input of the bottleneck section is the output of the encoder section, there are no Gated linear units applied at the bottleneck section, and it is also fully convolutive as the encoder section. In the bottleneck section, the one-dimensional convolutive layer is used with batch normalization, and the tanh activation function is applied. The bottleneck section contains two convolution layers only.

3) Decoder section: Transpose one-dimensional convolution is applied to convert back the original shape of the data. The decode section consists of one-dimensional convolution transpose along with gated linear units and

concatenation layers which serve as skip connections so that the model uses the features learned in the encoder section of the model.

4) Output layer: The output layer consists of a single channel focusing on mono-aural speech enhancement; Tanh activation is applied at the output layer, which provides the final denoised signal. In the output layer, a single one-dimensional convolutional layer is used which acts as a dense layer with a shape of (65536,1). The output layer is connected to the decoder layer. And it gives the same data shape as the input layer.

D. Objective

Mean Absolute Error (L1 Loss) is used as a loss function to minimize the error between the predicted signal $\boldsymbol{y'}$ and the clean signal $\boldsymbol{y}$. L1 loss is incorporated over the time domain of signals to reduce the loss over the sequential time domain and to minimize the loss over the spectral domain, L1 loss is used over on STFT (short-time Fourier transform) of the signals.

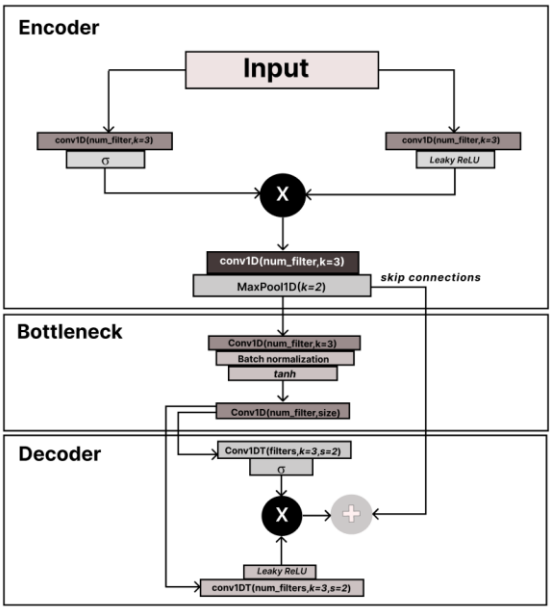

Fig. 2. Detailed overview of each separate block of the model and their connections with each other. The input and output shapes are the same.

Lastly, these L1 losses are added together to optimize the model in both the time domain (on amplitude vectors) and the spectral domain (on STFT).

A minimal epsilon value = $1e^{-10}$ is used to omit the undefined log error.

$$L_{stft}\,(y, y') = L_{mag}\,(y, y')$$

$$L_{mag}\,(y, y') \Rightarrow \frac{1}{T}|| \, log\,|STFT(y)| + \varepsilon - log|STFT(y')| + \varepsilon||_1$$

And,

$$L_{time}\,(y, y') = ||y - y'||_1$$

Overall, we wish to minimize the following:

$$L_1 \Rightarrow L_{mag}\,(y, y') + L_{time}\,(y, y')$$

IV. Training

Proposed model is trained on our custom dataset, 100 epochs for each of the three batches. The best model is saved for each of the three batches while monitoring the lowest L1 loss over the spectral domain, as discussed in Section 3.3.

Table I reports the training parameters, using the Adam optimizer, which has a learning rate of 1^{e-4}, with a momentum of β1=0.9 and a denominator momentum of β2=0.999.

TABLE I. Tuning Hyper-Parameters of Adam Optimizer Used for Training the UNET Model

Optimizer Tuning hyper-parameters	
optimizer	Adam
lr (learning rate)	1^{e-4}
β1 (momentum)	0.9
β2 (denominator)	0.999

The audio samples used for training are sampled at 22.05 KHz. The first 282 samples are used for the training set and the remaining 94 samples are used for the validation set. A batch size of 4 is used, dividing the data into 71 batches, each consisting of four unique utterances of the same speaker and the next batch containing the. The model optimizes the loss function and adjusts the weights. The next batch includes four different unique utterances of the previous speaker. After that, the next batch focuses on different speakers. The random order of data was not utilized. In each training session with varying noise conditions, the utterances' order remains consistent with their corresponding pairs in the noisy environment.

Gated linear units are used in order to control the information flow inside the model. It is observed that Gated liner units act as voice activity detection layers. Where clean speech is present, the neurons have greater weight and lower weight where noise is present. The output layer acts as a normalization layer for the denoised output.

V. Result

The performance of the model is calculated over cumulative loss of Mean Absolute Error of predicted and actual audio samples in both the time domain which is WAV MAE and also is spectral domain which is STFT MAE.

It is observed that the addition of STFT MAE in loss function improved the audible quality of speech in noisy environments and preserved the speech components of audio by making them sound less distorted.

Firstly, performance is evaluated over sixteen channels of respective noise environments that were present in the DEMAND dataset [22], this is reported in Table II. Then performance is evaluated over only eight channels of respective noise environments. It is observed that in the case of 16 channels of PRESTO and OMEETING audile quality of

audio was distorted, the reason behind this is that in the case of PRESTO and OMEETING noise environment babble noise is present and when it gets mixed with human speech it is hard for model to differentiate between noise and speech. We reported the performance of PRESTO and OMEETING with eight channels mixed together in Table III and a significant drop in Loss is observed while preserving the speech quality with less distortion.

TABLE II. PERFORMANCE OVER SIXTEEN CHANNEL MIX

Noise Environment All Sixteen Channel Mixed	Train			Test		
	LOSS	***WAV MAE***	***STFT MAE***	***LOSS***	***WAV MAE***	***STFT MAE***
DKITCHEN	0.755	0.011	0.744	0.777	0.012	**0.764**
PRESTO	0.734	0.017	0.716	0.811	0.018	**0.792**
OMEETING	0.537	0.010	0.527	0.573	0.010	**0.563**

This table represents the cumulative loss, the wav loss over signals, and the STFT loss of the signals. This metric represents the performance of the model where the lowest STFT MAE is observed in each noise condition. All sixteen channels of noise are added together in all of three noise profile cases.

TABLE III. PERFORMANCE OVER EIGHT CHANNEL MIX

Noise Environment First Eight Channel Mixed	Train			Test		
	LOSS	***WAV MAE***	***STFT MAE***	***LOSS***	***WAV MAE***	***STFT MAE***
PRESTO	0.665	0.012	0.653	0.744	0.013	**0.730**
OMEETING	0.494	0.008	0.486	0.510	0.009	**0.500**

This table represents the cumulative loss, the wav loss over signals, and the STFT loss of the signals. This metric represents the performance of the model where the lowest STFT MAE is observed in each noise condition. Eight channels of noise are added together in all of the two noise profile cases.

A significant drop in STFT Loss is observed in Table III. When the first eight channels are mixed, and the output audio quality is better than the previous setting in the case of PRESTO and OMEETING. Reducing the number of channels in DKITCHEN is not tested because the audible intelligibility of the clean speeches was satisfactory, this shows that the model is giving better results on the DKITCHEN environment despite higher STFT MAE as reported in Table II.

On Objective measures, intelligibility, and speech quality is measured as reported in Tables IV and V. The metrics used are PESQ [26] and STOI [27]. Baseline PESQ and STOI are calculated over denoised signals by using the noisereduce library of Python. Then it is compared with PESQ and STOI of denoised signals of the model.

TABLE IV. PERFORMANCE ON OBJECTIVE MEASURES USING PESQ AND STOI (SIXTEEN CHANNELS MIXED)

Noise Environment All Sixteen Channel Mixed	Test Set	
	PESQ wb	***STOI***
Baseline_DKITCHEN	1.16	0.83
DKITCHEN	**1.60**	**0.85**
Baseline_PRESTO	1.30	0.66
PRESTO	**1.31**	**0.73**
Baseline_OMEETING	1.30	0.82
OMEETING	**2.46**	**0.88**

Objective measures of enhanced speech on PESQ and STOI. Sixteen channels of noise are added together in all three noise profile cases.

TABLE V. PERFORMANCE ON OBJECTIVE MEASURES USING PESQ AND STOI (EIGHT CHANNELS MIXED)

Noise Environment First Eight Channel Mixed	Test Set	
	PESQ wb	***STOI***
Baseline_PRESTO	1.24	0.74
PRESTO	**1.38**	**0.81**
Baseline_OMEETING	1.35	0.85
OMEETING	**3.24**	**0.90**

Objective measures of enhanced speech on PESQ and STOI. Eight channels of noise are added together in all of the two noise profile cases.

In Tables IV and V, our performance of the model in objective measures using PESQ and STOI is reported. Improvement in the audible quality of denoised speech is observed when the human voice is more prominent (audible) in signal than in the noise environment. There is a significant improvement in both objective measures over the baseline model, baseline measurement is carried with PESQ and STOI on our test set using spectral gating.

VI. DISCUSSION

In summary, this research addresses the challenges in speech enhancement by examining a mapping-based approach on raw audio waveforms using the U-Net architecture. The study identifies limitations in conventional methods like Spectral subtraction and the Wiener filter, prompting an exploration of deep learning solutions.

The proposed approach optimizes the loss function for both time and short-time Fourier transform (STFT) of audio, enabling the direct generation of clean audio representations without additional post-processing. The research systematically evaluates mask-based and mapping-based deep learning approaches, revealing the effectiveness of the latter in various noise environments through a comprehensive metric.

Objective measures, including PESQ and STOI, indicate notable improvements in audible quality and intelligibility of denoised speech compared to baseline models. The research findings have practical implications for applications such as automatic speech recognition and speech-to-text systems. The mapping-based approach on raw audio waveform emerges as a viable strategy for addressing the inherent challenges in speech enhancement, offering tangible advancements in audio quality assessment.

VII. CONCLUSION

The proposed approach can be scaled by incorporating all the noise profiles into one single dataset and training a model in a single pass over the entire dataset, the hypothesis is that the model may generalize better to each noise condition, and a reduction in Loss is observed. It is observed that the intelligibility of output speech samples is improved when STFT is included with the Loss Function, as discussed in the objective of this paper. In the future, the proposed approach can be scaled with the use of multiple-resolution STFT loss as used in [28, 29].

REFERENCES

[1] Yuliani, A. R., Amri, M. F., Suryawati, E., Ramdan, A., & Pardede, H. F. (2021). Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi, 21(1), 19-26.

[2] Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction." IEEE Transactions on acoustics, speech, and signal processing 27, no. 2 (1979): 113-120.

[3] Lim, J. S. "Enhancement and bandwidth compression of noisy speech." Proc. IEEE 67, no. 12 (1962): 1689-1697.

[4] Scalart, Pascal. "Speech enhancement based on a priori signal to noise estimation." In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 2, pp. 629-632. IEEE, 1996.

[5] Xu, Yong, 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, no. 1 (2013): 65-68.

[6] Subramanian, Aswin Shanmugam, Xiaofei Wang, Murali Karthick Baskar, Shinji Watanabe, Toru Taniguchi, Dung Tran, and Yuya Fujita. "Speech enhancement using end-to-end speech recognition objectives." In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 234-238. IEEE, 2019.

[7] Gnanamanickam, J., Natarajan, Y., & KR, S. P. (2021). A hybrid speech enhancement algorithm for voice assistance application. Sensors, 21(21), 7025.

[8] Soni, M. H., Shah, N., & Patil, H. A. (2018, April). Time-frequency masking-based speech enhancement using generative adversarial network. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5039-5043). IEEE.

[9] Takeuchi, Daiki, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. "Effect of spectrogram resolution on deep-neural-network-based speech enhancement." Acoustical Science and Technology 41, no. 5 (2020): 769-775.

[10] Liu, Kuan-Yi, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung. "Speech enhancement based on the integration of fully convolutional network, temporal lowpass filtering and spectrogram masking." In Proceedings of the 31st Conference on Computational Linguistics and Speech Processing (ROCLING 2019), pp. 226-240. 2019.

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

[12] Liu, Chang-Le, Sze-Wei Fu, You-Jin Li, Jen-Wei Huang, Hsin-Min Wang, and Yu Tsao. "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1888-1900.

[13] Tan, Ke, and DeLiang Wang. "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6865-6869. IEEE, 2019.

[14] Wang, Y., Han, J., Zhang, T., & Qing, D. (2021). Speech enhancement from fused features based on deep neural network and gated recurrent unit network. EURASIP Journal on Advances in Signal Processing, 2021, 1-19.

[15] Defossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. arXiv preprint arXiv:2006.12847.

[16] Tan, K., Zhang, X., & Wang, D. (2019, May). Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5751-5755). IEEE.

[17] Valin, J. M. (2018, August). A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In 2018 IEEE 20th international workshop on multimedia signal processing (MMSP) (pp. 1-5). IEEE.

[18] Kaur, J., Baghla, S., & Kumar, S. (2015). A review: Audio noise reduction and various techniques. Int. J. of Advances in Sci. Engn. and Techn., 3(3), 132-135.

[19] Westhausen, N. L., & Meyer, B. T. (2020). Dual-signal transformation lstm network for real-time noise suppression. arXiv preprint arXiv:2005.07551.

[20] Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013, August). Speech enhancement based on deep denoising autoencoder. In Interspeech (Vol. 2013, pp. 436-440).

[21] Lu, X., Matsuda, S., Hori, C., & Kashioka, H. (2012). Speech restoration based on deep learning autoencoder with layer-wised pretraining. In Thirteenth Annual Conference of the International Speech Communication Association.

[22] J. Thiemann, N. Ito and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," 9 June 2013. [Online]. https://zenodo.org/record/1227121#.Y_RhTHZBzIU [Accessed 30 Jan 2022]

[23] Valentini-Botinhao, C. (2017). Noisy speech database for training speech enhancement algorithms and tts models. University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR).

[24] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234-241. Springer International Publishing, 2015.

[25] Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017, July). Language modeling with gated convolutional networks. In International conference on machine learning (pp. 933-941). PMLR.

[26] Rix, Antony W., John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), vol. 2, pp. 749-752. IEEE, 2001.

[27] 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.

[28] Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim. "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199-6203. IEEE, 2020.

[29] Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim. "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation." arXiv preprint arXiv:1904.04472 (2019).