Multi-Block U-Net for Wind Noise Reduction in Hearing Aids

Arth J. Shah, Manish S. Suthar, and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar, India
https://sites.google.com/site/speechlabdaiict/
202101154@daiict.ac.in, manishkmrsuthar@gmail.com,
hemant_patil@daiict.ac.in

Abstract. With advancement in technology, many changes can be observed in various technological equipments. With advancement in hearing aids technology, hearing disabled peoples are benefited abundantly. In this study, we aim to improve hearing aid technology, by proposing an advanced solution to one of the major problem of wind noise disturbance. In particular, we design a three stage multi-block U-Net model to suppress the degradation with high-quality reproduction of sound. We analyzed time-frequency domain-based audio representation analysis, and trained model on realistic noise for better hearing aids user experience. The properties of wind noise, affecting the signal quality have been discussed deeply in addition to effect of wind noise for hearing aids users. Fully-convoluted two different blocks of U-Net were used in order to generate the proposed model, which outperforms existing models when evaluated with different performance metrics. The importance of deep learning methodologies in combination with hearing aids chips, and importance of realistic data for training an balanced model is also demonstrated in this study.

Keywords: Hearing Aids, Wind Noise Reduction, U-Net, End-to-End Denoising, Speech Enhancement.

1 Introduction

After the invention in 19^{th} century, audio recording technologies and devices have been upgrading continuously, and so is hearing aids technology. With advancement in technology, hearing aids users have been benefited abundantly by improved quality of speech, availability of different features, and functions in hearing aids. According to recently released reports of World Health Organization (WHO), 1.5 billion people (nearly 20 % of humans) have hearing loss, among whom 439 million have disabling hearing loss [1]. These 1.5 billion numbers can grow upto 2.5 billion people by the end of year 2050 as stated by reports [2]. While more than 400 million people worldwide could benefit from hearing aid use alone, only 17 % get to use these devices. An hearing disabled, individual faces many issues in communication, in particular, malfunctioning of hearing aids. One

of such commonly faced issue by hearing aids users, is Traveling Speech Degradation (TSD), which refers to low speech quality received by hearing aids due to noise in scene of hearing aid user. Due to such unwanted noise interruption, the quality of speech degrades abundantly, resulting into difficulty for hearing to the subject. Such noise is mainly due to high velocity of wind (i.e., wind noise), and a low frequency vehicle as well as background noise. In this paper, aim is to capture the pattern of audio waveform and decrease the background noise by employing advanced deep learning methods. Many such methods have also been explored in this field, however, they fail to capture properties of speech individual, resulting into miscasting of original speech wave.

Hearing Aids users, experience a variety of unwanted sounds from multiple sources, among which one of the main problem being encountered is wind noise while traveling or while outstation. The quality of speech also degrades, due to hiss and clicks, resulting into low quality of speech along with wind noise. Both hiss and clicks can disrupt the clarity of the sound, making it difficult for users to focus on important sounds like speech. Other factors also such as, listening fatigue, background noise, and distort speech sounds also make it harder to understand conversation. When this unwanted sounds mix with wind noise, the quality of speech is at its poorest version of interpretation. Moreover, it also helps us to suppress the impulsive events. Denoising of speech signals have been previously tackled with various technologies, such as wavelets [3], spectral subtraction [4], Linear Prediction [5], and Wiener filtering [6], which were successful only for stationary noises, and resulted not giving better performance on non-stationary noise. Wind noise, is non-stationary because of its characteristics, such as amplitude and frequency, that changes over time. Wind speed and direction can vary, leading to fluctuations in the intensity and spectral properties of the noise it produces, and thereby resulting into speech signal degradation in hearing aids.

2 Related Work

In [7], the researchers employs wind noise attenuation algorithm (WNAA) to capture properties of speech signal properties in order to decrease Signal-to-Noise Ratio (SNR) level of speech signal. They claim the wind noise to be bounded between low-mid frequency range, i.e., in regions $\leq$ 3 kHz. This claim may be an important clue for conducting further research on similar topic. They also illustrated and classified the effects on background noise due to direction of microphones. Study reported in [8] provides literature about directional and unidirectional microphones in behind-the-ear (BTE) hearing aids. In [9], the authors present one of the most interesting works on real-time wind noise cancellation and reduction using an fully signal and speech processing techniques. They employ Fast Fourier Transform (FFT)-based system, resulting into reduction of wind noise with just 32 *ms* of speech / audio frame. Also their system is known that the tolerable group delay function for mild hearing loss should be below

about 5 ms. The authors of [9], employ system that prevents spatial information for binaural hearing aids (BHA), by cancelling low delay wind noise. They employ Short-Time-Fourier-Transform (STFT)-based technique and estimated perceptual evaluation of speech quality (PESQ) scores for their model. Their success motivated us to employ a similar system based on spectrogram denoising. By advancement in technology, much progress have been made in field of machine learning (ML), and deep learning (DL). Motivated by this progress, we employ an system based on advanced deep learning model U-Net, which is a convolutional-based neural network, originally proposed for image segmentation task [10].
Inspired by historic recordings denoising [11], we in this study employ an system similar to them, however the novelty lies in the structure of U-Net blocks. Authors of [11], employed particular type of blocks (I-Blocks), which resulted in low restoration of high frequency bins while denoising. We employ multiple blocks, namely, J-Block and K-Block, inorder to restore this high frequency bins while reducing effect of wind noise in speech signal. We also perform, various types of analysis based on the proposed model, that validates the capability and usability of model. We employed an end-to-end system that takes noisy audio file as input, and gives output of clean denoised audio file. For robust model, we employed 6 types of different wind noise at various SNR levels. The proposed system provides the following novelty:

- Deep understanding of wind noise effect on hearing aids.
- U-Net Multiblock system.
- STFT-based analysis for wind noise reduction.

The rest of the paper is organized as follow: Section 2 provides information and properties of wind noise in hearing aids. Section 3 gives details of proposed U-Net model, and methodology employed. Dataset and performance metrics detailed information are proposed in Section 4. Experimental results, their discussion, and different SNR related noise experiments are discussed in detail in Section 5. Section 6 summarizes and concludes the paper along with potential future research directions.

3 Wind Noise

This Section describes wind noise, its characteristics, and its relation with hearing aids. Hearing aids users face a variety of acoustic environments in day-to-day life. Classification of these environments for hearing aids have also been attempted in the literature. This classification task for hearing aids is also known as Acoustic Scene Classification (ASC), which is one of the primitive and basic task for speech enhancement in hearing aids. This study focus on primary and one of the most challenging task of speech enhancement in hearing aids. *Wind noise* in this study refers to air pressure creating blockage at microphones in hearing aids resulting into undesired sound at the output as hearing aids. While traveling on an vehicle (specially 2 - wheeler, i.e., bike), in deserts (where velocity of wind is

much higher), in front of air conditioner (where air directly strikes hearing aids), and while running, are most common unavoidable circumstances, where wind noise in hearing aids increases abundantly and hence, reducing speech quality of speakers. In such scenario, hearing aids user face unacceptable circumstances when the adjacent speakers' voice is affected abundantly due to high power wind noise. This results in malfunctioning of hearing aids and may cause severe fatal accidents to the users and thus, degrade quality of life. For such observations, a bunch of hearing aids users were examined in presence of wind noise, resulting into increase in hearing deficiency at that moment [7]. According to reports conducted by ORCA-US, 42 % users reported negative feedback with hearing aids in presence of wind noise [12]. Motivated by this, we propose use of an alternate model into hearing aids as an solution to this problem. While physical movement of hearing aids user, there exists two types of wind, i.e., true wind and head wind, whose vector sum results in apparent wind as mentioned in Eq. (1). In particular,

$$W_{aperrent} = W_{true} + W_{head}. \tag{1}$$

For example, if a hearing aids user is running at speed of 2 m/s opposite to 2 m/s wind, apparent wind could be calculated as 4 m/s, however, if the runner and wind are in the same direction with the same velocity, apparent velocity becomes 0 m/s.

3.1 Wind Noise Creation

Low velocity of air flowing around an object, result into parallel moving (also referred to as laminar flow), and air with higher velocity cannot go around object laminar. This phenomena results in changing of direction of airflow or returning of air into the same direction as proposition generating spatial pressure difference between layers. Such partial pressure difference are refereed as turbulence, resulting into pressure variation on hearing aid microphones due to velocity variations caused by irregular airflow as shown in Fig. 1 (a). The disturbance in hearing aids output due to this partial difference is known as "wind noise" [13]. Sometimes pressure exerted by wind noise moves the diaphragm to microphone amplifier, resulting into noise distortion. Red rays on Fig. 1 (b), refers to airflow directions, which creates loud wind noise as compared to blue region resulting into quiet wind noise w.r.t. microphone position. U-Net architecture is particularly suitable for tasks like audio denoising due to its unique characteristics that facilitate efficient feature extraction and reconstruction [14].

3.2 Characteristics of Wind Noise

Studies in the literature proved that due to high force of wind noise, wind noise has high levels as 116 dB for some BTE hearing aids at low wind speed of 11 m/s, due to high force on hearing aids microphone by wind. Characteristics of the wind noise can be obtained by two major factors, namely, wind speed and wind direction. Also wind noise level is known to be proportional to square of

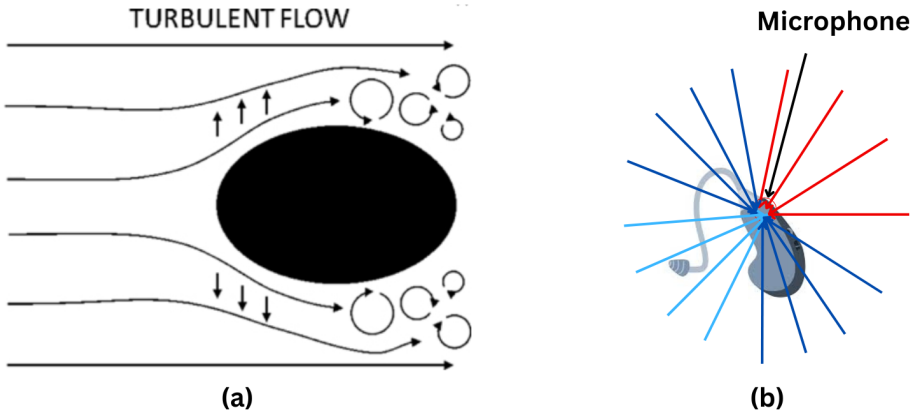

Fig. 1. (a) Turbulance created by wind [7], and (b) hearing aids noise effect w.r.t. microphone.

wind speed [15], [16]. In real life measurements, as wind speed increases, the noise level can be increased much more than theoretical analysis. As discussed in sub-Section 3.1, when wind noise faces directly to hearing aids microphone (red direction), the noise level is recorded to be more as compared to the wind passing through side portions (blue direction). Opposite direction (sky blue) rays state almost negligible impact of wind noise on hearing aids. Wind noise possesses many spectral characteristics as defined below.

- Low frequency energy concentration (below 300 Hz).
- Increase in spread of noise energy at high frequencies.
- Unique turbulence creation at each measurement point, resulting into rapid decrease in correlation between two points.
- Dual microphone hearing aids result in difference in turbulence creation at individual location.

Many mechanical changes and approaches have been explored in order to reduce wind noise effect on hearing aids, among which placing an cover over hearing aids microphone to laminate wind flow seems to have immense potential [8]. Many other approaches have also been explored on the same problem [7], [17]–[21].

3.3 Noise Degradation

Noise comes with various degradation in SNR levels, for this study, we choose various levels of SNR including severe noisy condition to make model more robust to different noise environment, variable wind speed, and speech scene and thus, increase the practical suitability of our work. In moving vehicle or any other moving condition, the velocity never remains constant, and so doesn't wind speed. For such robust analysis, we selected 6 different noise analysis at different SNR levels (varying from -10 to 10 dB), in order to make model robust as much as possible.

3.4 Spectrographic Analysis

We feed STFT spectrogram as an input to U-Net model. Fig. 2 (a) represents clean audio spectrogram (STFT), and (b) shows noisy spectrogram of wind noise signal, which we aim to denoise. Fig. 2 (c), (d), (e), (f), and (g) represents noisy spectogram, spectrogram obtained from spectral gatting, deepfilternet3, Nsnet2_denoiser, two stage unet, and multistage-mutiblock unet of audio file. We can further observe that the spectrogram as shown in Fig. 2 (g), have more clear harmonics (black box, and white box) as compared to other obtained spectrograms. It can be also observed that in process of obtaining clean spectrogram from highly noisy spectrogram, we are able to predict the spectrogram which resembles almost like clean spectrogram). However, high frequency region having low resolution and blunt harmonics can be observed within the highlighted area in Fig. 2 (g), indicating minor loss of few properties while gaining the clean audio signal from noisy audio signal. Significant difference can be observed between noisy spectrogram and predicted spectrogram in other approaches (red boxes), indicating the better working of proposed model. However, training of model can be done on higher number of epoch and different variety of noise, inorder to obtain a spectrogram nearly similar to clean spectrogram. After denoising, We employed an standard and openly available speech enhenancement model, namely, Resemble Enhenancement Model (REM)[1]. The audio files obtained were also less audible in other models due to loss of important data-points, however, they were more efficient than the original noisy audio file as we can observe their spectrographic difference in Fig. 2.

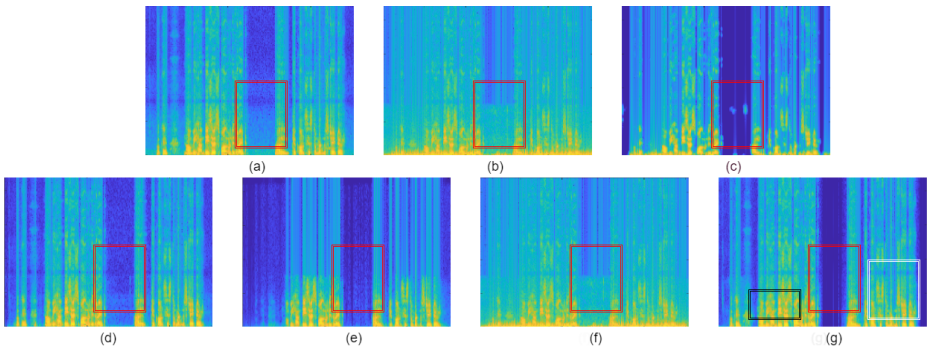

Fig. 2. (a), (b), (c), (d), (e), (f), and (g) represents clean spectogram, noisy spectogram, spectrogram obtained from spectral gatting, deepfilternet3, Nsnet2_denoiser, two stage unet, and multistage-mutiblock unet of audio file, respectively.

[1] resemble-ai "https://github.com/resemble-ai/resemble-enhance?tab=readme-ov-file"

4 Proposed Approach

This Section describes the model (spectrogram-based fully convoluted) employed in this study to minimize the impact of wind noise on the speech signal. Much progress have been made in the feild of image restoration and denoising in recent days, among whom, [22], demostrates a multistage image restoration architecture of U-Net. To that asspect, we employ an supervised attention module (SAM) based three stage multiblock U-Net architecture for wind noise reduction, as displayed in Fig. 3. Two phase approach have also been motivated by an recent study [11], which also employees almost similar architecture for historical recordings restoration. In the first phase of two phase method consists of U-Net subnetworks. However on the first stage, the inputs and training objective differs. The aim of this stage is to identify and localize the residual noise emerging after denoising an speech signal. The second stage works as an main classifier, which also ensures smooth denoising, and refines the noisy speech using extracted features from first stage. This will decrease the impact of wind noise over speech signal.

Authors of this study decided to explore Short-Time Fourier Transform (STFT) spectrogram for U-Net module's input. STFT is useful when analyzing non-stationary signal having different spectral content over time. It involves segmenting an audio signal into small time intervals and performing Fourier Transform (FT) on those segments. Unlike the standard FT, which provides a global frequency view, the STFT captures local frequency content that varies with time. Each FT in the STFT provides simultaneous time and frequency information. It highlights how different frequencies contribute to the signal over short time windows. This is particularly useful for signals like speech and music, where frequency content changes dynamically. Inorder to capture dynamic change in speech signal, and local frequency components, we extracted STFT spectrograms from 44.1 kHz audio signal, thereafter reading real and imaginary part of signal as real valued signal. Frame size and frame length were chosen as 2048 samples, and 512 samples respectively. In the initial layers, we enhance the network's frequency information by including frequency-positional embedding [11] as 10 additional channels in the input data.In each phase, the 12-channel input coming off of the previous process is passed through a shallow feature extractor, comprised of a convolutional layer followed by the Exponential Linear Unit (ELU) [23] as shown in Fig. 1. $F_{in,1}$ extracted at the first layer is then fed directly into U-Net sub-network; whereas for the second phase, we concatenate input features ($F_{in,2}$) with the ones from a set of other features produced by an additive component in the attention mask for each position. Only optimal features emerges from second stage of neural network, due to SAM module. The U-Net output features Fout,1 are used to create the estimated residual noise signal N via a 3×3 convolutional layer. The first stage output $Y1$ is estimated as $Y1 = X + N$ where X is the input spectrumgram. The attention-guided features F_{SAM} presented in Fig. 1 are determined by the attention masks M produced from $Y1$ through a 1x1 convolution and a sigmoid function. Lastly, we take the

features output from the second U-Net ($F_{out,2}$) then apply a 3×3 conv layer to generate the denoised output $Y2$.

To supervise the model, we minimize the mean absolute error of both outputs at each stage. The reconstruction loss function is defined as:

$$L = \frac{1}{K} \sum_{k=1}^{K} (|Yk1 - Yk| + |Yk2 - Yk|) \tag{2}$$

Consider a case where clean spectrogram is represented by Y, and the total number of Short-Time Fourier Transform (STFT) bins is K. The Adam optimizer [20] with $\beta_1 = 0.5$, $\beta_2 = 0.9$ and a starting learning rate of 1×10 decaying by a factor of 10 for every 100,000 steps was used during training. It is worth noting that normalization strategies were not used since batch normalization and weight normalization did not produce any improvements in our experiments.

4.1 U-Net Subnetworks

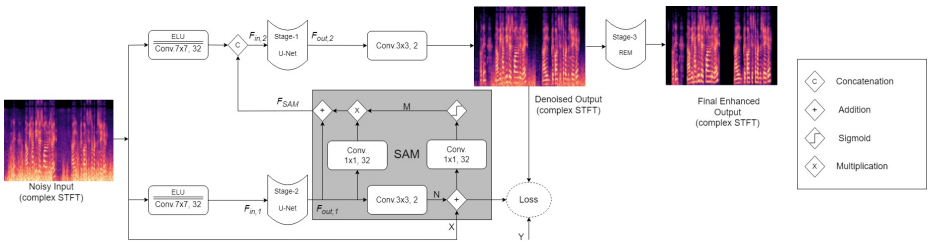

Fig. 3. Three stage denoising model with the SAM module [11].

In computer vision tasks and audio processing, the most common use of the U-Net architecture is documented [24], [25]. We employed U-Net subnetworks that are designed following a symmetrical encoder-decoder design, featuring four downsamplers and upsamplers as shown in Fig. 4 (a) for this study.Each scale contains an intermediate block referred to as either J-Block 4(b) or K-Block 4 (d) as shown in figure 4, which consists of densenet two layer block with residual connection. Concatenating skip connections are used to connect the outputs of the encoder J-Blocks to their respective decoder J-Blocks. Additionally, an I-Block is placed after the fourth downsampler. K-Block is similar to J-Block, however the number of DenseNet block with residual connection increases to 4.The downsampling layers employ strided convolutions with a stride of 2x2 and a kernel size of 4x4 as well as the same number of filters as the next I-Block which consists of three DenseNet blocks, as illustrated in Figure 2(c). In the decoder, the upsampling layers make use of transposed convolutions having

identical hyperparameters to their corresponding downsampling layers. Our experiments showed that while checkerboard artifacts are a common side effect of transposed convolutions, they started to disappear as training progressed.

4.2 Resemble Enhancement Model

The Resemble Enhancement Model (REM)[1] is the ultimate in sound enhancement technology. REM is based on sophisticated algorithms and machine learning, which allows it to do more than just analyse sound. It enhances speech intelligibility, and maintains the original tonality of music and voice. Being a seasoned sound designer, it naturally modifies his techniques to fit various settings and sound kinds. REM works excellent for enhancing speaker volume in a conference context. Its true allure is in its capacity to isolate and accentuate specific audio components while preserving a clean, distortion-free sound quality efficient fine-tuning algorithms that guarantee minimum latency and maximum effect are responsible for this quick performance. The enhancer is a latent conditional flow matching (CFM) model. It consists of an Implicit Rank-Minimizing Autoencoder (IRMAE) and a CFM model that predicts the latents. This first stage involves an autoencoder that compresses the clean mel spectrogram Mclean into a compact latent representation Zclean , which is then decoded and vocoded back into a waveform. The model consists of an encoder, decoder and vocoder.

As we have used pre-trained model, it works as follows. After completing the training of the first stage, In second Stage the latent CFM model is trained. The CFM model is conditioned on a blended Mel Mblend = αM denoised + *(1 -* α*)*Mnoisy , derived from the noisy STFT-spectrogram Mnoisy and a denoised STFT-spectrogram Mdenoised Here, α is the parameter that adjusts the strength of the denoiser. During training, the α is set to follow a uniform distribution $U(0,1)$. During inference, the value of α can be controlled by the user. To predict the latent representation of the clean speech here we have used the loaded pre-trained enhancer model and which is already trained jointly with the latent CFM model. REM does not discriminate between sources or audio formats. REM effortlessly transitions between clear speech recordings, energetic musical performances, and tranquil surround sound to provide an improved listening experience on all media platforms and playback devices.

5 Experimental Setup

5.1 Dataset Used

The CHiME1 (Computational Hearing in Multisource Environments) dataset is the first installment in a series designed to facilitate the development and evaluation of robust speech recognition systems in challenging acoustic environments. Recorded in a typical domestic setting, the dataset features binaural recordings of utterances from the GRID corpus, a collection of simple, fixed grammar sentences, spoken by multiple speakers. The CHiME1 dataset includes both clean

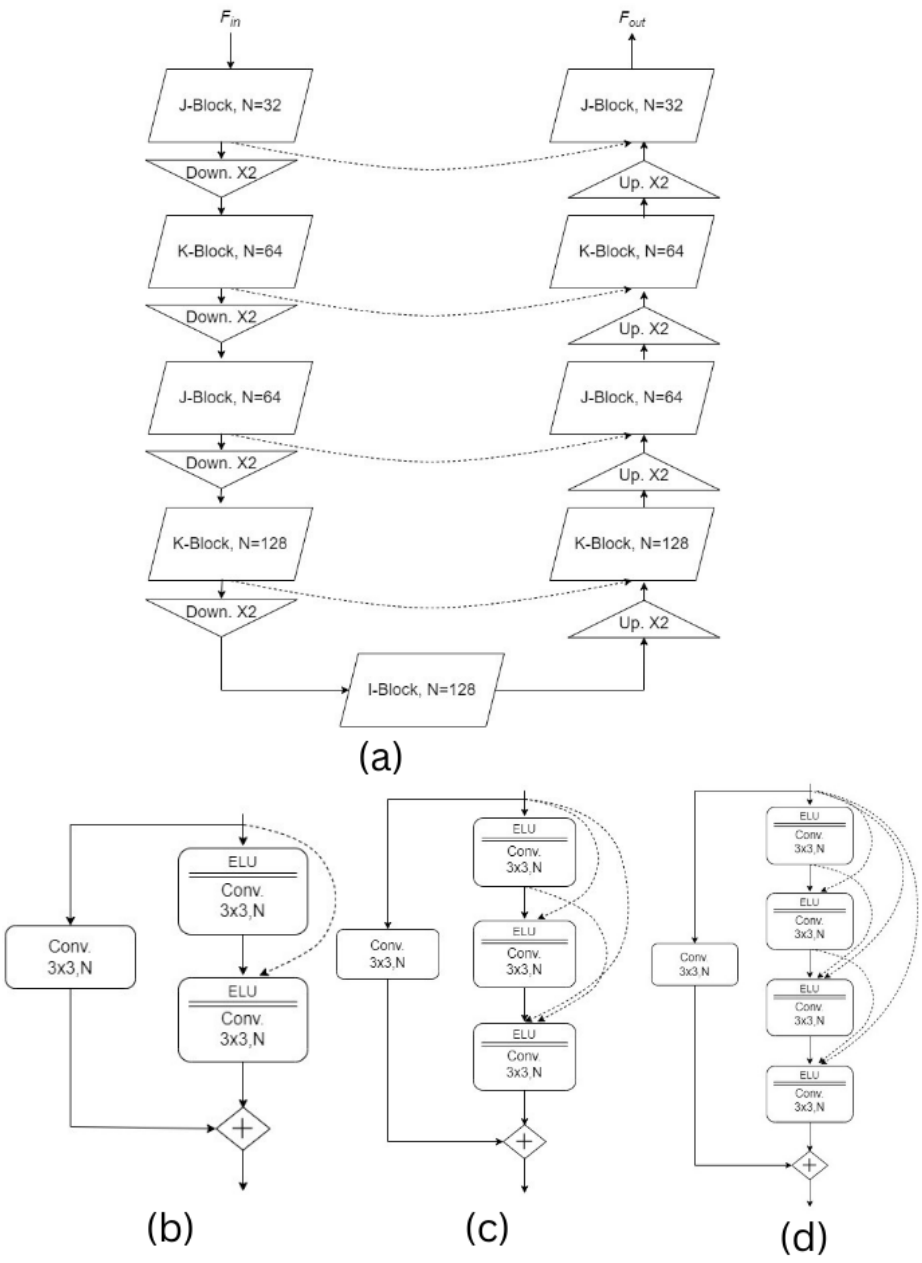

Fig. 4. SAM module embeded with UNet architecture, and blocks.

and noisy versions of the recordings, allowing us to do controlled experimentation in noise robustness. Additionally, it provides detailed transcriptions of the speech content and separate recordings of the background noises, enabling researchers to test and improve noise suppression and speech enhancement algorithms. This dataset has been widely used for benchmarking and advancing techniques in the field of speech recognition, particularly in scenarios involving multiple, dynamic noise sources. Six types of different wind noise was collected from freesound[2].

5.2 Performance Metrics

Mean Absolute Error (MAE) MAE measures the average magnitude of errors between predicted and actual values without considering their direction. It is calculated as the average of the absolute differences between predictions and actual observations, providing a straightforward interpretation of the prediction accuracy.

Coherence In the context of topic modeling, coherence measures the degree of semantic similarity between high-scoring words in a topic. A higher coherence score indicates that the words within a topic are more related to each other, suggesting better quality of the topic model.

Δ Mean Squared Deviation (Δ MSD) Δ MSD is a variation of Mean Squared Deviation (MSD), often used to assess the variability of a set of values. It measures the squared differences between predicted and actual values, providing a sense of how predictions deviate from actual outcomes. Delta MSD specifically emphasizes the changes or differences between these deviations over time or across different conditions.

Short-Time Objective Intelligibility (STOI) STOI is an objective metric for evaluating speech intelligibility. It compares short-time segments of the clean and degraded speech signals to estimate how intelligible the speech is to human listeners. STOI is widely used in speech enhancement and hearing aid algorithms to assess and improve the clarity of speech signals.

6 Experimental Results

We trained model for 200 epoch, and 2000 steps per epoch. At the end of training, we were able to obtain training loss of 3 % and validation loss of around 6 %.

[2] freesound.org "https://freesound.org/"

6.1 Comparison With Existing Works

This sub-Section compared obtained results with few existing works results. However, the authors of this study were not able to compare the work with more existing works, as not all authors released their trained model. Due to lack of time, and insufficiency of resources, authors of this study were also not able to retrain existing models, and decided to compare them with open sources released models available. Table 1 denotes the obtained results on various existing studies, compared with proposed methodology. Observations based on Table1 denotes that, after completing process of multi stage U-Net on a noisy audio, the words are clear enough for a subject to be interpreted and have high coherence then other existing models as observed. Coherence denotes that how much words within topic are related to each other. On the other hand, low coherence on existing models is due to less denoising of model, or removal of important speech properties while denoising. Alternatively, Deepfilternet, being an advance speech denoising / enhancement model, outperforms proposed method in various aspects. At the point controversy arises that is the proposed model an optimal

Table 1. Comparison op proposed approach with existing works.

Noise	Model	Coherence	Δ MSD	STOI	MAE
-5dB	Spectral-Gatting [26]	0.88227	5.3702	0.77388	1559.27
	DeepfilterNET [27]	0.8538	1.0882	0.9128	728.71
	Nsnet2-denoiser [28]	0.343	1.6175	0.1715	56299977.16
	Two-Stage Unet [11]	0.8972	4.9898	0.82	3854.66
	Multiblock-Unet (Without REM)	0.9091	4.4575	0.82166	3578.19
	Multistage-MultiBlock-Unet (Proposed)	**0.9642**	2.4395	0.9189	1037.99
0dB	Spectral-Gatting [26]	0.89179	5.5726	0.8473	1559.21
	DeepfilterNET [27]	0.97794	0.6378	0.9416	454.64
	Nsnet2-denoiser [28]	0.36411	1.2707	0.1328	41908510.01
	Two-Stage Unet [11]	0.95397	2.5139	0.8776	2108.84
	Multiblock-Unet (Without REM)	0.95741	2.2291	0.8799	1946.9
	Multistage-MultiBlock-Unet (Proposed)	**0.983**	2.2743	0.9516	726.57
5dB	Spectral-Gatting [26]	0.89889	5.5219	0.8658	1559.27
	DeepfilterNET [27]	0.9826	0.5256	0.9571	382.57
	Nsnet2-denoiser [28]	0.36342	0.8017	0.1627	35356174.96
	Two-Stage Unet [11]	0.97894	0.8662	0.9071	1217.19
	Multiblock-Unet (Without REM)	0.97863	0.78482	0.9094	1103.75
	Multistage-MultiBlock-Unet (Proposed)	**0.98713**	2.3324	0.9611	665.61
10dB	Spectral-Gatting [26]	0.90091	5.3941	0.914	1559.27
	DeepfilterNET [27]	0.98662	0.248	0.977	237.07
	Nsnet2-denoiser [28]	0.3661	0.5275	0.1313	26015898.19
	Two-Stage Unet [11]	0.98689	0.3694	0.9614	613.36
	Multiblock-Unet (Without REM)	0.98722	0.3004	0.9633	572.68
	Multistage-MultiBlock-Unet (Proposed)	**0.9903**	2.6847	0.9817	641.1

model? Resulting into an positive aspect, we also calculate Mean Opinion Score (MOS) for Deepfilternet, resulting into proving superiority of proposed methodology. MOS being an evaluation metrics to validate model on real life situation,

by asking an subject to rate the denoised speech between 1 (Bad) to 5 (Good). Table 2 denotes the ratings of 108 users after hearing speech from various speech models. In Table 1, proposed model did not performed well on other metrics as compared to DeepFilternet, as DeepFilternet is denoising more as compared to proposed method, however, it also degrades the speech quality while denoising and user is not able to get better experince due to degraded quality of sound, which can be observed and concluded in table 2.

Table 2. MOS obtained on 108 participants (78 male + 30 Female).

Model	MOS
Spectral Gatting [26]	3.67
Nsnet2-denoiser [28]	2.33
Two-Stage Unet [11]	3.89
DeepFilternet [27]	4.21
Proposed	**4.52**

7 Summary and Conclusions

In this work, we presented significance of multistage UNet formed from various types of blocks to denoise hearing aids speech with additive white noise. We enquired end-to-end methodology for identifying the pattern of wind noise in the speech signal. This study investigated three stage UNet, formed by three different types of blocks, namely, I-Block, J-Block and K-Block. The primary goal of this study is to restore and enhenance the quality of speech, which has been degraded due to presence of various types of wind noise at various SNR levels. The features extracted (STFT Spectrograms) by signal processing concepts were then fed into the UNet model for the process of maping an noisy spectrogram to clean spectrogram. For this study, we made a variety of observations based on models capability, such as, evaluations based on coherence, Δ MSD, STOI, MAE, and MOS. In comparison to existing widely used and open-source denoising models, we achieved significantly better results for the task selected. We also discussed the difference and improvement between proposed and existing methodology. The proposed system have been explored for only six type of noise, which we aim to extend the work to various different types of noise, to analyze effect of different types of noise on model, as a future task. Future works also involve more detailed mathematics on proposed methodology, and exploring variety of blocks for proposed approach.

Acknowledgements

This study has been funded by Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, under the sponsorship of project 'BHASHINI', (Grant ID: 11(1)2022-HCC.(TDIL)). Authors thank authorities of DA-IICT Gandhinagar for their kind support and cooperation to carry out this research work.

References

[1] S. K. Swain, "Hearing loss and its impact in the community," *Matrix Science Medica*, vol. 8, no. 1, pp. 1–5, 2024, {Last Accessed Date : $2^{nd} July, 2024$}.

[2] W. H. Organization *et al.*, *World report on hearing*. World Health Organization, 2021, {Last Accessed Date : $5^{th} June, 2024$}.

[3] C. Schremmer, T. Haenselmann, and F. Bomers, "A wavelet based audio denoiser," in *Proc. IEEE International Conference on Multimedia and Expo*, Citeseer, 2001, 145–148, Tokyo, Japan.

[4] T. Biswas, C. Pal, S. B. Mandal, and A. Chakrabarti, "Audio de-noising by spectral subtraction technique implemented on reconfigurable hardware," in *2014 Seventh International Conference on Contemporary Computing (IC3)*, IEEE, 2014, 236–241, Noida, India.

[5] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1791–1801, 2007.

[6] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[7] P. Korhonen, "Wind noise management in hearing aids," in *Seminars in Hearing*, Thieme Medical Publishers, Inc., vol. 42, 2021, pp. 248–259.

[8] K. Chung, L. Mongeau, and N. McKibben, "Wind noise in hearing aids with directional and omnidirectional microphones: Polar characteristics of behind-the-ear hearing aids," *The Journal of the Acoustical Society of America (JASA)*, vol. 125, no. 4, pp. 2243–2259, 2009.

[9] Y. Uemura, H. Nakashima, N. Hiruma, and Y.-i. Fujisaka, "Real-time wind noise cancellation based on binaural cues for hearing aids,"

[10] O. Ronneberger, P. Fischer, and T. 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*, Springer, 2015, pp. 234–241.

[11] E. Moliner and V. Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, Singapore, pp. 841–845.

[12] S. Kochkin, "Marketrak viii: Consumer satisfaction with hearing aids is slowly increasing," *The Hearing Journal*, vol. 63, no. 1, pp. 19–20, 2010.

[13] K. T. Walker and M. A. Hedlin, "A review of Wind-Noise Reduction Methodologies," *Infrasound monitoring for atmospheric studies*, vol. 1, pp. 141–182, 2009.

[14] J. A. Zakis, "Wind noise at microphones within and across hearing aids at wind speeds below and above microphone saturation," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 3897–3907, 2011.

[15] M. Strasberg, "Dimensional Analysis of Windscreen Noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 544–548, 1988.

[16] J. M. Kates, *Digital hearing aids*. Plural publishing, 2008, {Last Accessed Date : 10^{th} April, 2024}.

[17] Widex, "Widex supertm power to hear," {Last Accessed Date : 10^{th} April, 2024}.

[18] T. Ricketts, A. Dittberner, and E. Johnson, "High-frequency amplification and sound quality in listeners with normal through moderate hearing loss," *J Speech Lang Hear Res*, vol. 51, no. 1, pp. 160–172, 2008.

[19] K. Chung, N. McKibben, and L. Mongeau, "Wind noise in hearing aids with directional and omnidirectional microphones: Polar characteristics of custom-made hearing aids," *J Acoust Soc Am(JASA)*, vol. 127, no. 4, pp. 2529–2542, 2010.

[20] G. J., "An innovative rie with microphone in the ear lets users hear with their own ears," *GN Hearing AS*, pp. 1–8, 2020.

[21] K. Chung, "Effects of venting on wind noise levels measured at the eardrum," *Ear Hear*, vol. 34, no. 4, pp. 470–481, 2013.

[22] S. W. Zamir, A. Arora, S. Khan, *et al.*, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 821–14 831.

[23] A. Shah, E. Kadam, H. Shah, S. Shinde, and S. Shingade, "Deep residual networks with exponential linear unit," in *Proceedings of the third international symposium on computer vision and the internet*, 2016, pp. 59–65.

[24] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018, {Last Accessed Date : $5^{th} June, 2024$}.

[25] D. Baloch, S. Abdullah, A. Qaiser, S. Ahmed, F. Nasim, and M. Kanwal, "Speech enhancement using fully convolutional unet and gated convolutional neural network," *Int. J. Adv. Comput. Sci. Appl*, vol. 14, pp. 831–836, 2023.

[26] E. Sudheer Kumar, K. Jai Surya, K. Yaswanth Varma, A. Akash, and K. Nithish Reddy, "Noise reduction in audio file using spectral gatting and fft by python modules," in *Recent Developments in Electronics and Communication Systems*, IOS Press, 2023, pp. 510–515.

[27] H. Schröter, T. Rosenkranz, A. Maier, *et al.*, "Deepfilternet: Perceptually motivated real-time speech enhancement," *arXiv preprint arXiv:2305.08227*, 2023, {Last Accessed Date : $5^{th} June, 2024$}.

[28] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*, Springer, 2020, pp. 79–86.