Non-Stationary Speech Denoising via a Wavenet Architecture with Dilated Convolutions and Gated Units

Pradeep Kumar Sriperamboodhuru
Jawaharlal Nehru Technological University Hyderabad, India
sams.pradeep@gmail.com (corresponding author)

Anitha Sheela Kancharla
Jawaharlal Nehru Technological University Hyderabad, India
kanithasheela@jntuh.ac.in

Received: 6 February 2026 | Revised: 13 March 2026 | Accepted: 24 March 2026

 DOI: https://doi.org/10.48084/etasr.18019

ABSTRACT

Since the intelligibility of speech signals in speech communication systems can be affected by ambient noise, researchers have developed a number of methods to improve intelligibility. WaveNet is a promising deep learning model to overcome this constraint. WaveNet is a generative model that employs autoregression to produce the probability distribution of the subsequent sample based on fragments of the preceding sample. A supervised version of WaveNet, in which the model learns by minimizing regression loss, can be used to address speech denoising. The proposed model introduces noncausality, discrimination, target field prediction, and conditioning, improving computational efficiency by making the model highly parallelizable. Evaluations show that the proposed method performs better than classical methods, such as a commonly used method based on processing magnitude spectrograms. The proposed method yields higher SNR gains up to 19.23 dB and lower MCD values as low as 9.41, achieving promising results for speech denoising in non-stationary environments.

Keywords-deep learning; WaveNet; speech denoising; end- to-end processing

I. INTRODUCTION

Speech denoising is the process of removing undesired sounds from speech while maintaining its quality, improving the speech signal by removing or minimizing noise while preserving essential speech features [1, 2]. Effectively separating the desired speech signal from unwanted noise components is difficult, particularly when they overlap in the time and frequency domains. Speech denoising techniques aim to reduce noise while minimizing distortion and preserving speech integrity [3].

Speech denoising is used to accurately estimate the speech signal from the mixed signal. Speech enhancement research has focused primarily on dealing with additive background noise, which is generally easier to handle than convolutive nonlinear noise [4]. Existing approaches in this field often incorporate statistical constraints on speech and noise inputs. However, accurately predicting the characteristics of speech or disturbance noise presents significant challenges. Consequently, there is a trade-off between noise suppression and preserving the fidelity of the speech signal, which may lead to distortion in the processed speech.

Classical methods, such as spectral gating and Wiener filtering, have been widely used for speech denoising [5]. They typically operate in the spectral domain, aiming to estimate the clean speech spectrum by exploiting the statistical characteristics of the noise [6, 7]. Deep neural networks in an encoder-decoder framework can be used to implement speech enhancement systems trained using the Short Time Fourier Transform (STFT) of the signals, which captures significant contextual information [8].

Classical methods are computationally efficient and can be effective in situations where the noise characteristics are well understood and stationary; however, they may struggle in handling non-stationary noise when the noise characteristics vary significantly. WaveNet uses dilated convolutions to capture long-range dependencies in the signal [9], trained end-to-end on large amounts of speech data, allowing it to learn complex patterns and generate high-quality denoised speech [10]. WaveNet operates directly in the time domain, capturing temporal variations and fine-grained details in the speech signal. It can handle non-stationary noise and adapt to varying noise characteristics without relying on explicit assumptions. Nevertheless, it typically requires more computational resources and training data than classical methods.

II. METHODS OF SPEECH DENOISING

A. Spectral Subtraction

The spectral subtraction technique is considered a basic algorithm for noise reduction. Using FFT, the noisy speech signal is first converted into the frequency domain to produce $X(\omega)$ [11]. Then, the magnitude spectrum is calculated for processing. In order to estimate and iteratively update the noise spectrum, especially for stationary noise conditions, the noise estimation/update block uses a voice activity detector to detect speech-absent intervals. The estimated noise spectrum is subtracted from the noisy spectrum to obtain a clean speech. Lastly, the enhanced speech signal is reconstructed using the Inverse FFT (IFFT) [12] and the phase of the noisy speech, as shown in Figure 1.

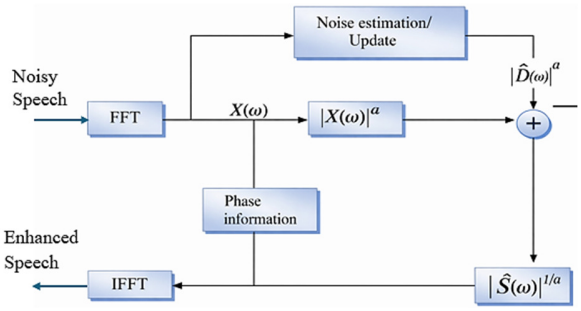

Fig. 1. Block diagram of spectral subtraction.

The noisy signal $x(n)$ comprises the clean speech signal $s(n)$ and the undesired additive noise signal $d(n)$:

$$x(n) = s(n) + d(n) \quad (1)$$

In this scenario, the noise is assumed to be a stationary or slowly varying:

$$X(\omega) = s(\omega) + D(\omega)$$
$$X(\omega) = |X(\omega)|e^{j\theta_{x(\omega)}} \quad (2)$$

where $|X(\omega)|$ represents the magnitude spectrum and $\theta_{x(\omega)}$ is the phase of the noisy signal. The polar form of the noise spectrum is represented as follows [13]:

$$D(\omega) = |D(\omega)|e^{j\theta_d(\omega)} \quad (3)$$

Here, $|D(\omega)|$ denotes the magnitude of the noise spectrum. The phase of the noisy spectrum can be considered to be $\Theta x(\omega)$. A prediction of the clean signal spectrum can be obtained as:

$$S'(\omega) = (|X(\omega)| - |D'(\omega)|)e^{j\Theta x(\omega)} \quad (4)$$

where $|D'(\omega)|$ represents the measured noise spectrum. IFFT can be used to generate the improved voice signal of $S'(\omega)$.

B. Wiener Filtering

The Wiener filter is an LTI filter that approximates a random signal by linearly filtering an obtained noisy signal. It minimizes the Mean Squared Error (MSE) between the predicted and required signals through known signal and noise spectra. Given the measurements of a WSS random process $x(n)$, the goal is to find the unit sample response of an LTI system [14]. This allows the filter's output $y'(n)$ to be the MMSE estimate of a joint WSS target process $y(n)$. The error $e(n)$ can be written as: $e[n] = y'[n] - y[n]$. The aim is to minimize it as follows: $\min \varepsilon = E(\varepsilon^2[n])$. Figure 2 depicts the Wiener filter, where noisy speech statistics include the sum of speech and noise components in the time and frequency domains. The filters $h(t)$ and $H(\omega)$ yield the MMSE-optimal estimate of the clean signal.

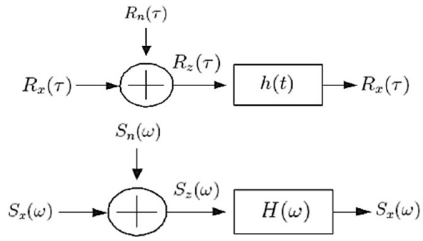

Fig. 2. Block diagram of the filter: time and frequency domain.

C. Spectral Gating

Spectral gating is a method to reduce noise from speech signals by selectively attenuating or deleting noisy spectrum components. This technique is based on the fact that speech and noise frequently occupy distinct frequency areas of the spectrum. Spectral gating methods analyze the spectral properties of the input signal to identify and discriminate the speech and noise components. The method begins by evaluating the Power Spectral Density (PSD) of noisy voice signals [15]. This can be accomplished using the short-time Fourier Transform or other such techniques. The PSD is then compared with a predetermined noise threshold to identify which frequency bins are noise-dominated. After the noisy spectral components are identified, the spectral gating technique uses a gain function to reduce or eliminate the noise. Typically, the gain function is designed to lower the amplitudes of the noise components while maintaining the amplitudes of the speech components. Many methods can be utilized to determine the gain, including simple binary masking, in which noise-dominated parts are entirely reduced, and speech-dominated regions are kept intact.

D. Deep Learning Methods

Deep learning-based models are becoming increasingly popular due to their ability to learn and accomplish a task without the headache of feature engineering [16]. Deep learning algorithms for audio denoising are classified as mask-based or mapping-based. Based on noisy speech input, masking-based models construct masks in the time or frequency domain to reduce disturbances in the input [17]. In contrast, given a large amount of both noisy and cleaned speech, algorithms based on mapping attempt to obtain clean speech directly from noisy audio.

Recently, significant research in speech enhancement has been based on deep learning models that correlate noisy speech to clean spectral representations. In [18], a fully convolutional R-CED network worked as well as or better than traditional FNN and RNN models while using much fewer parameters. The main reasons for this efficiency are that it does not use

pooling and that it uses optimized 1D convolution operations. This CNN-based architecture is great for applications that need to run in real time and with a lack of resources.

Generative Adversarial Networks (GANs), first introduced in [19], are a powerful way to model data and learn complex distributions. A GAN consists of a generator and a discriminator trained through a min-max game. However, GANs face challenges such as training instability and mode collapse. Better designs, such as WGAN and CGAN, are more stable and easier to control. Models based on GANs, such as SEGAN, MetricGAN, and UNetGAN [20], have made great progress in enhancing speech. However, many studies show that WaveNet and its parallel versions still give better perceptual quality, better harmonic reconstruction, and more natural speech. This means that WaveNet is the best design for applications to improve speech quality.

III. WAVENET ARCHITECTURE FOR SPEECH DENOISING

WaveNet is a probabilistic and autoregressive deep neural network that predicts each sample based on previous ones and generates raw audio waveforms [10]. It can be trained efficiently on datasets with high sample rates. WaveNet uses dilated convolution to effectively capture long-range temporal dependencies. Causal dilated convolution, which is appropriate for real-time audio generation, guarantees that each output in the original WaveNet formulation depends only on the current and previous samples [10]. On the other hand, the entire noisy signal is accessible during processing for speech denoising applications. To improve noise suppression, the proposed model uses non-causal dilated convolution, which enables the network to use contextual information [21].

The proposed model (Figure 3) first receives the noisy speech waveform as input, which is then processed through a 1-D convolution layer with a 3×1 kernel size to extract the signal's initial temporal features. Subsequently, several residual blocks are applied to the extracted features. To capture long-range dependencies in speech signals, the network uses non-causal dilated convolution layers with increasing dilation factors. These layers increase the network's receptive field without appreciably adding more parameters. A gated activation mechanism, made up of tanh and sigmoid functions, is incorporated into each convolution block; the tanh function generates candidate feature representations, while the sigmoid function manages the information flow. The sigmoidal gates govern the activation in each layer as follows:

$$O_t' = \tanh(Wf * x_t) \odot \sigma(W_g * x_t) \quad (5)$$

where $*$ represents convolution and $\odot$ represents elementwise multiplication, t denotes the input time, t' is the output time, g represents the gate indices, and W_f and W_g are convolutional filters. Here, the tanh function acts as a learned filter, and the sigmoid function acts as a learned gate.

Further, a conditioning block containing bias terms is added. Skip connections are used to aggregate the intermediate representations produced by various convolution layers, preserving crucial information and enhancing training stability.

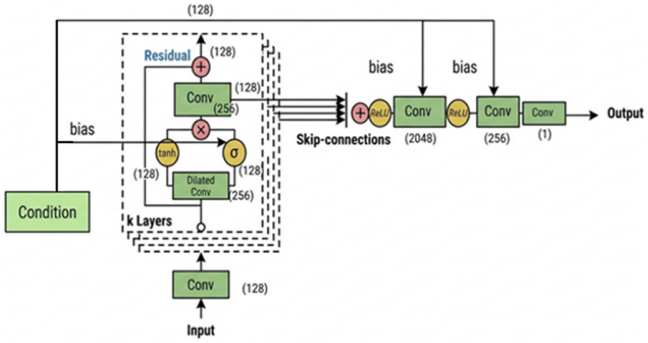

Fig. 3. Complete implementation architecture of the proposed wavenet-based speech denoising model.

The combined features are used by the output convolution layer to reconstruct the improved speech signal. The detailed steps are illustrated in Algorithm 1.

Algorithm 1: Proposed WaveNet for speech denoising
Input: Noisy speech fragment $x(n)$
Output: Denoised speech $y(n)$
Step 1: The noisy speech signal $x(n)$ is first passed through a 1-D convolution layer using a 3×1 for three times to extract initial features.
Step 2: These features are then processed through residual blocks composed of non-causal dilated convolutions, where progressively increasing dilation factors help capture long-range temporal dependencies.
Step 3: A gated activation unit (tanh-sigmoid) with conditioning bias is applied and skip connections are used to combine intermediate feature representations.
Step4: Finally, the aggregated features are fed into an output convolution layer to produce the enhanced speech signal $y(n)$.

IV. TRAINING AND TESTING

A. Dataset and Preprocessing

This study employed the Noisy Speech Database for Training Speech Enhancement Algorithms and TTS models (NSDTSEA) [22], which is a collection of noisy speech samples and their corresponding clean speech counterparts [9]. The dataset encompasses a total of 11,527 training samples and 8,824 testing samples. These recordings feature the voices of 28 distinct speakers recorded in 40 diverse noisy environments, such as parks, buses, and cafés. The testing samples involve the voices of two speakers exposed to various noise conditions. The audio samples are sampled at a rate of 16 kHz. Each sample has an average duration of approximately 3 s, with a standard deviation of 1 s. No audio preprocessing, such as μ-law companding is applied. The pipeline is designed to be an

end-to-end process. To ensure that the model is robust across different noise levels, clean speech signals were mixed with noise at random SNR levels between 2.5 and 17.5 dB. Noisy and clean waveform pairs were used for supervised training. The noisy speech signals were used as input, and the corresponding clean speech signals were used as references to denoise noisy samples for training the model.

B. Model Flow and Hyperparameters Used

The model comprises 30 residual layers with exponentially increasing dilation factors ranging from 1 to 512. This pattern is repeated three times. A 3×1 convolution is applied before the filters in each residual layer, the first dilated convolution to match the input's 1-channel to 128-channel. A ×1 convolution with 128 filters is used in skip connections [15]. The skip connections are summed, followed by ReLU activation. The final two convolutional layers are 3×1 in size, with 2048 and 256 filters, respectively, and not dilated. These layers are separated by ReLU activation. The output layer uses a 1×1 filter, and the feature map is linearly mapped to a single-channel temporal signal. This configuration achieves a receptive field of 7739 samples. Additionally, the model incorporates a binary-encoded scalar representing the speaker's identity as a bias term in each convolutional operation. The model is trained using the following parameters:

- Batch_size: 10 (number of samples per batch)
- Early_stopping_patience: 16 (number of epochs to wait without improvement in accuracy before stopping training)
- Num_epochs: 50 (maximum number of training epochs)
- Num_train_samples: 1000 (the number of training samples presented to the model in one epoch)
- Num_test_samples: 100 (number of validation samples presented to the model in one epoch)

The supervised regression framework utilizes the MSE loss function to train the model [23]. This function minimizes the squared difference between the predicted clean waveform and the ground-truth clean waveform samples. The error is calculated as the average of the errors from all epochs. This technique is particularly useful for handling large datasets. In this setup, the binary-encoded condition input captures the speaker's features through convolutions during the training process. These features are then incorporated as biases in each convolution operation.

V. EXPERIMENTAL EVALUATION

The metrics used for evaluating the performance of the model are the Signal to Noise Ratio (SNR), which estimates the quality of the output, and Mel Cepstral Distortion (MCD), which estimates the intelligibility of the output. SNR is a popular assessment metric in most audio and signal processing domains for determining the quality or fidelity of a signal in the presence of noise, measuring the desirable signal-to-undesired noise power ratio, and providing an objective measure of signal clarity and noise distortion. SNR is calculated using:

$$SNR = 10\log_{10}\left\{\frac{P_{\text{signal}}}{P_{\text{Noise}}}\right\} \quad (6)$$

MCD quantifies the dissimilarity between two sets of mel ceps. It acts as a metric for evaluating the fidelity of parametric speech synthesis systems [12], indicating that a smaller MCD indicates a closer match between synthetic and natural mel cepstral sequences. MCD relies on Mel-Frequency Cepstral Coefficients (MFCCs) to capture the spectral attributes of speech signals. Both the clean reference and denoised signals are first transformed into MFCCs before being used to compute the MCD. The Euclidean distance is calculated between the MFCC vectors of the two signals. MCD is commonly expressed as the mel cepstral distance per frame or per second.

$$MCD = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_i (c_{ti} - \bar{c}_{ti})^2} \quad (7)$$

A. Model

- Dilations: 9 (maximum dilation factor as an exponent of 2)
- Num_stacks: 3 (number of context stacks in the architecture)
- Target_field_length: 1603 (desired length of the output)

B. Optimizer

- Decay: 0.0 (rate of reducing the learning rate over time)
- Epsilon: 1e-08 (a small positive value to ensure stability and prevent division by zero)
- Learning rate: 0.001
- Momentum: 0.9 (moving average of gradients to be maintained)
- Type: 'adam' (type of optimizer used)

C. Results Obtained through Classical Methods

Figure 4 shows the waveforms of the clean speech signal and its corresponding noisy versions under stationary and non-stationary noise conditions. Figure 4 also depicts the distortion added to the speech signal's temporal structure, which is the input for the denoising techniques. Figure 5 shows an amplitude histogram for the high-energy noise components of the signal.

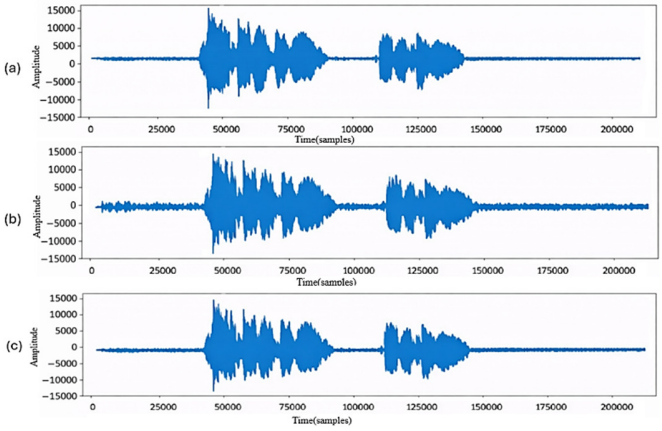

Fig. 4. (a) Clean speech; Noisy speech input to all the methods under (c) non-stationary and (c) stationary.

Figures 6 and 7 show the frequency-domain components used to estimate the noise profile, while Figure 8 shows the cleaned magnitude response after spectral subtraction. Figure 9 shows the output waveform of the Wiener filter (Ŷ). This is the clean speech that was estimated by the MMSE after noise suppression.

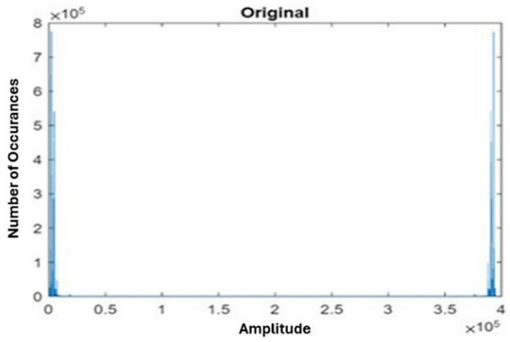

Fig. 5. Histogram of the original signal.

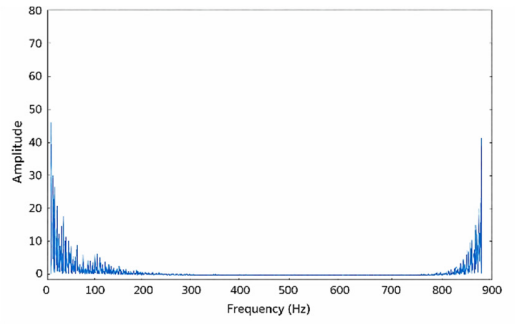

Fig. 6. Frequency spectrum of the noisy speech signal.

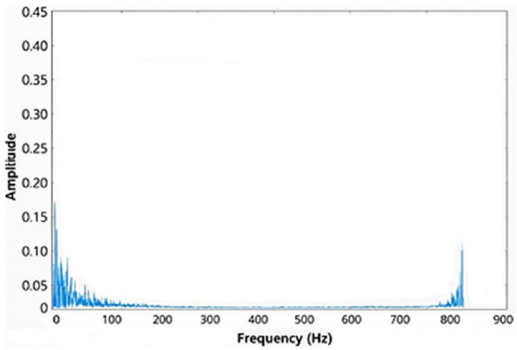

Fig. 7. Frequency spectrum of the noise signal.

Figure 10 shows the filter's frequency response, which shows how the gain shaping was used to reduce the power of the added noise across all frequencies.

The following spectrograms show the spectral-gating process. Figure 11(a) shows the noisy speech with an excessive amount of broadband noise, Figure 11(b) shows the spectral mask, demonstrating the time-frequency regions where speech is dominant, and Figure 11(c) shows the enhanced spectrogram made by using the derived mask to eliminate the noise-dominated segments.

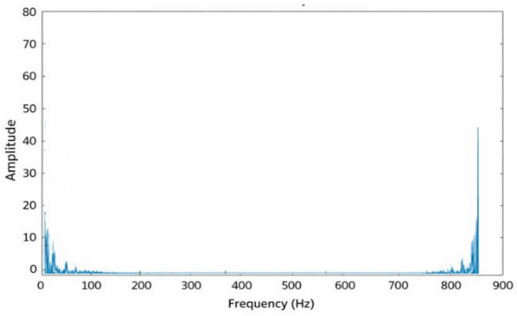

Fig. 8. Noise-subtracted speech spectrum.

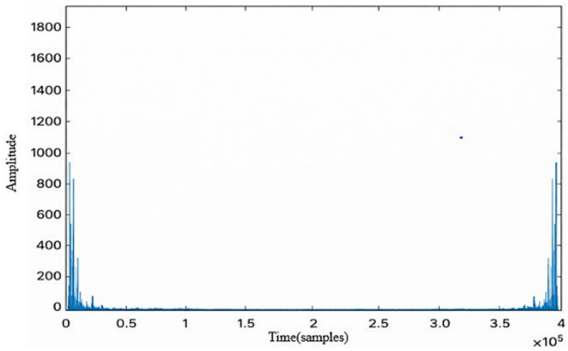

Fig. 9. Wiener filter output (Ŷ).

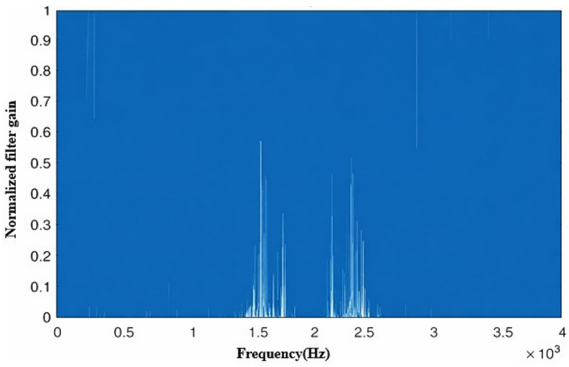

Fig. 10. Wiener filter response.

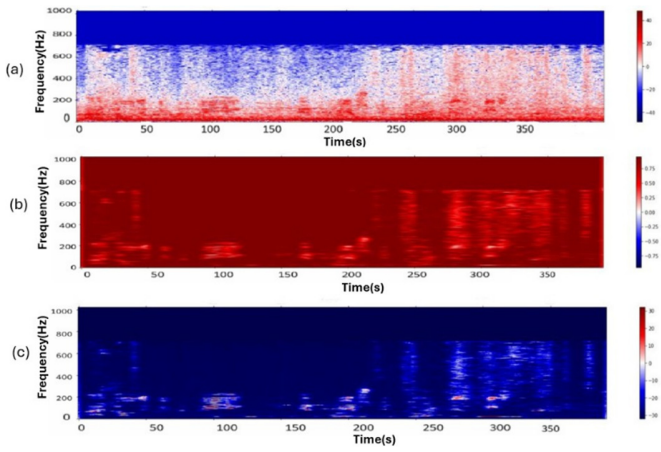

Fig. 11. Spectrograms of the speech signal using the spectral gating method: (a) Original noisy, (b) spectral mask, (c) enhanced spectrogram.

VI. RESULTS

The results of the comparison show that classical methods, such as spectral subtraction, Wiener filtering, and spectral gating, leave behind artifacts and do not suppress noise, especially when the noise is not stationary. The WaveNet-based method, on the other hand, shows better noise separation and signal reconstruction. Figure 12 demonstrates the proposed WaveNet model's denoising capabilities, where the reconstructed waveform closely resembles the clean reference despite significant input speech signal distortion. Figure 13 shows the model's discrimination behavior under noise-only input, where the noise-output branch precisely captures the underlying noise features while the data-output branch successfully eliminates undesirable speech-like components.

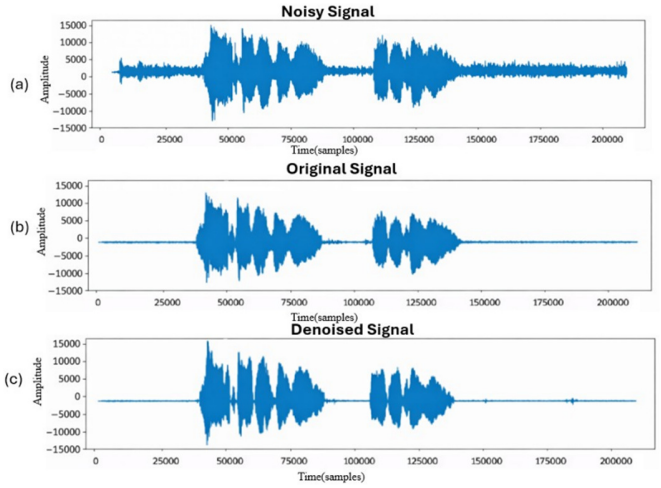

Fig. 12. Output of the WaveNet-based model for noisy speech: (a) Noisy signal, (b) original signal, (c) denoised signal.

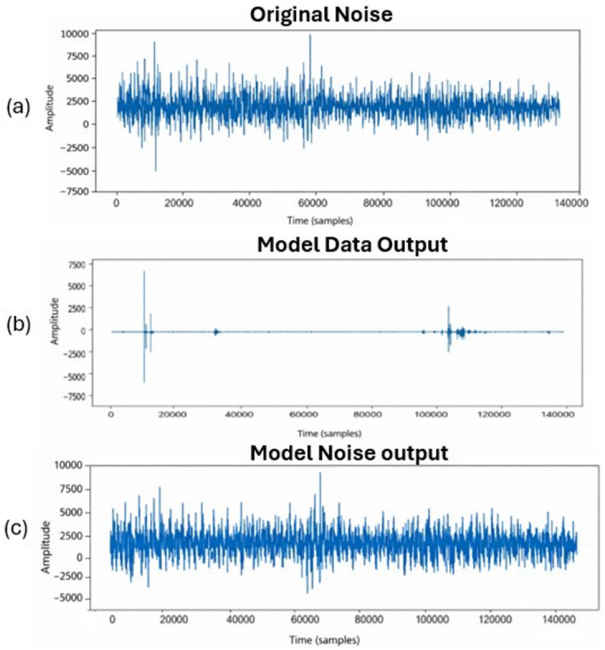

Fig. 13. Wavenet-based model output corresponding to noise-only input: (a) Original noise, (b) model data output, (c) model noise.

Figure 14 shows the training and validation loss curves. These curves illustrate that the model can generalize well and learn consistently, as shown by the smooth and steady convergence throughout the optimization phase. Figure 15 shows that the proposed WaveNet model terminates training early. It halts at epoch 49 and keeps the best model at epoch 33 based on minimum validation loss, thereby preventing overfitting and making the model work optimally.

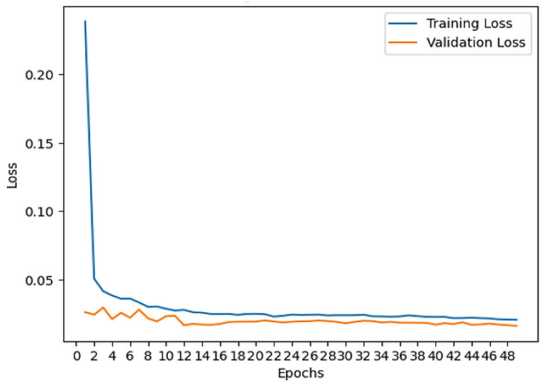

Fig. 14. Training and validation loss vs. epochs.

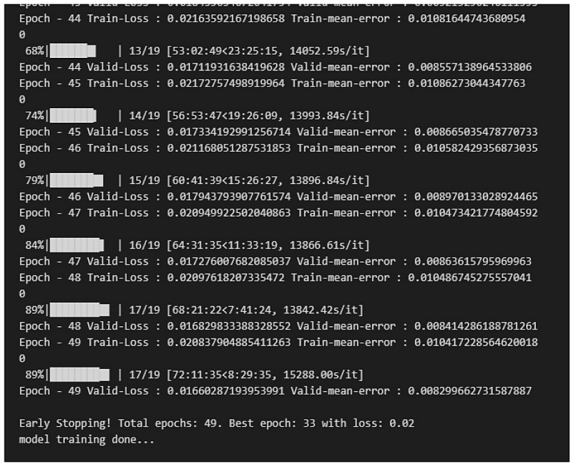

Fig. 15. Training process with early stopping.

A noise integrated audio sample with input SNR values of 2.5, 7.5, 12.5, and 17.5 dB is given to the classical methods, spectral subtraction, Wiener filtering, spectral gating, and the proposed model. Wiener filtering performs very well for stationary noise, even for low SNRs, but its performance degrades significantly for non-stationary signals, even for high SNRs. Spectral subtraction performs well for stationary noises, but not for non-stationary noises. This is because speech noise, which is very unpredictable in non-stationary noise, is not taken into consideration when estimating noise. Spectral gating performs moderately for both stationary and non-stationary noises, but its performance highly depends on the threshold value considered. Tables I and II compare SNR and MCD when the noise is stationary. Classical methods do better in this case.

TABLE I. COMPARISON OF SNR IN STATIONARY NOISE

Noise level (dB)	**2.5**	**7.5**	**12.5**	**17.5**
Proposed model (using WaveNet)	-0.1952	-0.0419	-0.0853	-0.5178
Wiener filtering	2.7570	7.7873	12.2648	18.4330
Spectral subtraction	2.5218	7.7373	12.2889	17.4464
Spectral gating	2.2031	8.4283	12.3437	17.2617

TABLE II. COMPARISON OF MCD IN STATIONARY NOISE

Noise level (dB)	**2.5**	**7.5**	**12.5**	**17.5**
Proposed model (using WaveNet)	26.7007	26.7136	22.5623	24.7155
Wiener filtering	16.9635	13.3563	11.5209	10.5177
Spectral subtraction	17.3250	14.2529	11.7084	9.3824
Spectral gating	29.5452	30.5783	30.7721	27.4745

TABLE III. COMPARISON OF SNR IN NON-STATIONARY NOISE

Noise level (dB)	**2.5**	**7.5**	**12.5**	**17.5**
Proposed model (using WaveNet)	3.1747	9.0269	14.7213	19.2280
Wiener filtering	2.6263	7.2069	12.4578	17.3524
Spectral subtraction	2.8397	8.6263	13.5001	18.2486
Spectral gating	2.5426	7.5711	13.3794	17.5745

TABLE IV. COMPARISON OF MCD IN NON-STATIONARY NOISE

Noise level (dB)	**2.5**	**7.5**	**12.5**	**17.5**
Proposed model (using Wavenet)	21.0143	12.5769	11.1937	9.4091
Wiener filtering	48.2019	42.7465	37.7465	38.3308
Spectral subtraction	24.6643	20.3668	19.4356	17.5753
Spectral gating	39.1009	43.2238	40.2238	38.3517

The proposed model effectively suppresses non-stationary environmental noise while maintaining speech quality, as evidenced by the consistent enhancement in output SNR and the reduction in MCD across varying input SNR levels. The proposed WaveNet architecture achieves significant quantitative improvements, yielding higher SNR gains (up to 19.23 dB) and markedly lower MCD values (as low as 9.41) under non-stationary noise conditions, outperforming all classical baselines across all tested SNR levels. Tables III and IV show how well the model works when there is previously unseen non-stationary noise. To test its robustness, the model was tested with speech data from speakers who were not in the training set, ensuring that the evaluation was not based on a specific speaker. The fact that SNR and MCD continue to improve at different SNR levels evidences that the proposed WaveNet model works well with different speakers and noise levels.

VII. CONCLUSION

WaveNet performs exceptionally well for speech denoising in terms of non-stationary noise, irrespective of its amount. The proposed model predicts target fields as opposed to individual samples, which greatly enhances its effectiveness while lowering its temporal complexity. Moreover, the model is time dimensionally adjustable because of its convolutional nature. Regardless of how it is trained, it is capable of denoising audio of variable length. Consequently, the proposed model supports one-shot denoising. Since training and inference can be performed on various hardware with varied amounts of memory, this flexibility is helpful. The proposed model is highly robust as it is efficiently trained in fewer epochs. However, although it has fewer parameters (approx. 6.3M), it requires a significant amount of computational resources, since it works directly in the time domain. Future research can focus on reducing the use of computational resources.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests that could have influenced the results of this work.

ACKNOWLEDGEMENT

Not applicable in this paper.

DATA AVAILABILITY

The dataset used in this study is publicly available in [22], and further details are provided in [24].

REFERENCES

[1] K. U. Shajeesh, K. S. Sachin, D. Pravena, and K. P. Soman, "Speech Enhancement based on Savitzky–Golay Smoothing Filter," *International Journal of Computer Applications*, vol. 57, no. 21, pp. 39–44, Nov. 2012.

[2] S. J. Lee and H. Y. Kwon, "A Preprocessing Strategy for Denoising of Speech Data Based on Speech Segment Detection," *Applied Sciences*, vol. 10, no. 20, Oct. 2020, Art. no. 7385, https://doi.org/10.3390/app10207385.

[3] J. Benesty, J. Chen, Y. (Arden) Huang, and S. Doclo, "Study of the Wiener Filter for Noise Reduction," in *Speech Enhancement*, Springer-Verlag, 2005, pp. 9–41.

[4] M. Tanveer *et al.*, "Ensemble deep learning in speech signal tasks: A review," *Neurocomputing*, vol. 550, Sept. 2023, Art. no. 126436, https://doi.org/10.1016/j.neucom.2023.126436.

[5] S. T. Yousif and B. M. Mahmmod, "Speech Enhancement Algorithms: A Systematic Literature Review," *Algorithms*, vol. 18, no. 5, May 2025, Art. no. 272, https://doi.org/10.3390/a18050272.

[6] S. P. Kumar and K. A. Sheela, "A DNN Based Adaptive Filter for Speech Enhancement," in *2024 Second International Conference on Data Science and Information System (ICDSIS)*, May 2024, pp. 1–5, https://doi.org/10.1109/ICDSIS61070.2024.10594120.

[7] W. Yuan and B. Xia, "A speech enhancement approach based on noise classification," *Applied Acoustics*, vol. 96, pp. 11–19, Sept. 2015, https://doi.org/10.1016/j.apacoust.2015.03.005.

[8] S. Kar and V. Mukherjee, "Acoustic Signal Enhancement Using Deep Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24259–24264, Aug. 2025, https://doi.org/10.48084/etasr.10571.

[9] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, Nov. 2023, Art. no. 101869, https://doi.org/10.1016/j.inffus.2023.101869.

[10] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5069–5073, https://doi.org/10.1109/ICASSP.2018.8462417.

[11] M. Lehekar and V. More, "Implementation of Speech Enhancement Algorithm on Hardware platform," in *2022 International Conference on Industry 4.0 Technology (I4Tech)*, Sept. 2022, pp. 1–4, https://doi.org/10.1109/I4Tech55392.2022.9952668.

[12] P. S. Rao and V. Sreelatha, "Implementation and Evaluation of Spectral Subtraction with Minimum Statistics using WOLA and FFT Modulated Filter Banks," M.S. Thesis, Blekinge Institute of Technology, Sweden, 2014.

[13] M. M. Lynn and C. Su, "Speaker Independent and Text Independent Emotion Recognition System Based on Random Forest Classifier," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 6, no. 12, pp. 9259–9266, Dec. 2018.

[14] E. Lai, "Time-domain representation of discrete-time signals and systems," in *Practical Digital Signal Processing*, Elsevier, 2003.

[15] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, Sept. 1998, https://doi.org/10.1109/89.709670.

[16] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, no. 6, Nov. 2021, Art. no. 420, https://doi.org/10.1007/s42979-021-00815-1.

[17] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Mapping and Masking Targets Comparison using Different Deep Learning based Speech Enhancement Architectures," in *2020 International Joint Conference on Neural Networks (IJCNN)*, July 2020, pp. 1–8, https://doi.org/10.1109/IJCNN48605.2020.9206623.

[18] S. R. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement." arXiv, 2016, https://doi.org/10.48550/ARXIV.1609.07132.

[19] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014, vol. 27.

[20] A. Wali *et al.*, "Generative adversarial networks for speech processing: A review," *Computer Speech & Language*, vol. 72, Mar. 2022, Art. no. 101308, https://doi.org/10.1016/j.csl.2021.101308.

[21] H. R. Guimarães, H. Nagano, and D. W. Silva, "Monaural speech enhancement through deep wave-U-net," *Expert Systems with Applications*, vol. 158, Nov. 2020, Art. no. 113582, https://doi.org/10.1016/j.eswa.2020.113582.

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

[23] K. Zhao and Y. Zhong, "English Speech Distortion Detection and Repair Based on Deep Learning," in *3D Imaging Technologies and Deep Learning*, 2025, pp. 145–156, https://doi.org/10.1007/978-981-96-5249-5_13.

[24] D. Rethage, "drethage/speech-denoising-wavenet." Mar. 26, 2026, [Online]. Available: https://github.com/drethage/speech-denoising-wavenet.