Sound & Vibration 2026, 60(4), 4065.
doi: https://doi.org/10.59400/sv4065

Article

Robust speech denoising using a parallel target-field non-causal WaveNet under stationary and non-stationary noise

Jella Sandhya [1,*], Shafiq UI Rehman [2,*], Shaik Mazhar Hussain[3]

[1] Electronics and Communication Engineering, Jawaharlal Nehru Technological University (JNTU), Hyderabad 500085, India
[2] College of Information Technology, Kingdom University, Riffa P.O. Box 40434, Bahrain
[3] Electronics and Communication Engineering, Malla Reddy (MR) Deemed to be University, Hyderabad 500100, India
*** Corresponding author:** Jella Sandhya, sandhyaphd801@gmail.com; Shafiq Ur Rehman, s.rehaman@ku.edu.bh

CITATION

Sandhya J, UI Rehman S, Hussain SM. Robust speech denoising using a parallel target-field non-causal WaveNet under stationary and non-stationary noise. Sound & Vibration. 60(4): 4065. https://doi.org/10.59400/sv4065

ARTICLE INFO

Received: 24 February 2026
Revised: 8 April 2026
Accepted: 14 April 2026
Available online: 21 July 2026

COPYRIGHT

Copyright © 2026 Author(s). *Sound & Vibration* is published by Academic Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: The presence of stationary or non-stationary noise substantially impairs speech intelligibility and its perceptual quality in present-day communication systems. Many existing spectral-domain methods like Wiener filtering and spectral subtraction rely on strong statistical assumptions about complex-valued data, while state-of-the-art neural models such as Conv-TasNet, DCCRN, and MetricGAN+ achieve remarkable performance gains for speech enhancement, but tend to suffer from heavy computational complexity and inference latency due to their complex architecture. In this paper, we present a new supervised non-causal WaveNet with parallel target-field prediction for efficient end-to-end speech denoising. The proposed model captures both past and future temporal context by enabling symmetric receptive fields and removing autoregressive dependencies. Traditional autoregressive WaveNet models generate samples one at a time, which require considerable redundant convolution operations due to the repetitive nature of their structures and tasks, while the presented method is capable of predicting a target field of samples in a single forward pass, allowing for parallel inference. We validate the proposed model on the NSDTSEA dataset for input SNR levels of 2.5 dB to 17.5 dB and under stationary as well as non-stationary noise types. We showcase experimental results that prove that the proposed framework consistently outperforms strong classical baselines, with over 1.22 dB gain in output SNR and 46% reduction in Mel Cepstral Distortion (MCD) under non-stationary noise. Moreover, when compared to state-of-the-art deep learning models such as Conv-TasNet, DCCRN, and MetricGAN+, the proposed method achieves a viable trade-off between performance, robustness, and computational efficiency.

Keywords: parallel target–field; non-causal WaveNet; stationary; non-stationary noise

1. Introduction

Speech enhancement is the process of recovering a clean speech signal from noisy observations and is an important technology for telecommunication, hearing aids, ASR, and other voice-controlled systems. Over the years, a host of speech processing approaches have been developed that can hamper intelligibility and perceptual quality, especially when speech signals are corrupted by stationary and non-stationary noise in real-world acoustic scenes. Traditional speech denoising methods like spectral subtraction and Wiener filtering were popular because of their computational efficiency and analytical tractability. Nevertheless, these methods depend on strict statistical assumptions about noise stationarity and often suffer from musical noise

artifacts or speech distortions in highly non-stationary conditions [1, 2]. Over the past decade, many deep learning-based speech enhancement methods have emerged to overcome such limitations. Previous neural network methods are based on spectral mask estimation or spectral mapping strategies, where time-frequency representations are used in the input [3]. For instance, SEGAN presented adversarial learning frameworks for end-to-end speech enhancement in the time domain [4], and Conv-TasNet showed that convolutional time-domain audio separation could replace spectrogram representation [5]. More recently, recurrent and convolutional-recurrent architectures like DCCRN have shown better performance by jointly learning temporal and spectral dependencies [6]. Transformer-based and attention-based architectures which still afford better robustness against complex acoustic environments [7]. Among the time-domain generative models, WaveNet has been particularly successful in the backbone with its autoregressive modeling of raw audio waveforms employing dilated causal convolutions [8]. WaveNet was originally designed for speech synthesis but it has been proven effective for modeling long-range temporal dependencies. Subsequent derivatives of WaveNet were developed for the task of speech enhancement, including Bayesian and discriminative approaches aimed at increasing robustness to noise [9, 10]. While powerful, traditional autoregressive WaveNet models are constrained by sequential inference and the degree of future context they use is severely limited, making them impractical for the application of denoising. In response to these issues, we present an end-to-end speech denoising solution that leverages a supervised non-causal WaveNet architecture with parallel target-field prediction. Eliminating the autoregressive restriction and incorporating symmetric receptive fields enables the model to take both past and future context into account for better reconstruction fidelity. In addition, an energy-consistent regression loss is used to maximize the noise suppression without distorting the speech structure. We further examine robustness across diverse acoustic environments by evaluating the proposed approach under both stationary and non-stationary noise conditions. To overcome these limitations in the existing approaches, we propose an end-to-end supervised speech denoising framework based on a non-causal non-autoregressive parallel target-field predicting WaveNet architecture. The proposed model fully removes autoregressive constraints and has a symmetric receptive field to model temporal context in both past and future directions, thereby allowing for higher data fidelity. Furthermore, to stimulate more noise elimination under the constraints of speech structure, we propose an energy-conserving regression loss. We evaluate the proposed approach for robustness with respect to stationary and non-stationary noise.

Recent speech enhancement methods either based on autoregressive inference, spectral-domain transformations, or computationally intensive generative frameworks, we first propose a non-causal WaveNet architecture with parallel target-field predictions. The main contribution is to remove the restriction that in a conventional WaveNet model the output samples cannot be computed in parallel, while establishing a symmetric context model by decoding using not just past but in fact future samples. This enables one-shot denoising with substantially lower computation redundancy. In addition, reconstruction quality is improved thanks to the energy-conserving loss

function explicitly preserving speech structure and reducing noise. All these design decisions together grant a reasonable trade-off between performance, robustness, and computational complexity, especially with respect to difficult non-stationary noise conditions.

The main contributions of this work are summarized as follows:

1. We propose a new non-causal WaveNet architecture that can symmetrically model context using past and future temporal information to mitigate the need for conventional causal designs.
2. A parallel target-field prediction method that can predict multiple samples with a single forward pass, which reduces the redundancy of computation in inference and increases the inference efficiency compared with autoregressive models.
3. A loss function that conserves energy, optimizing speech reconstruction and noise suppression together, which enhances the perceptual quality and maintains the fidelity of the signal.
4. A thorough evaluation framework that examines model robustness under both stationary and non-stationary noise conditions, highlighting the strength, effectiveness, and generalizability of the proposed approach.

The rest of this paper is organized as follows. The review of related works on classical and deep learning–based speech enhancement approaches is outlined in Section 2. Section 3 gives the proposed non-causal WaveNet architecture and its modifications for speech denoising. The dataset, training strategy, and experimental setup for model development are discussed in Section 4. In Section 5, we provide a comprehensive experimental evaluation of the proposed method while comparing our results to classical and modern data-driven, deep learning-based methods under both stationary and non-stationary noise scenarios. Section 6 concludes the paper and discusses future directions for work.

2. Related works

There have been efforts in recent years to enhance speech, and most have focused on developing deep learning-based end-to-end approaches [11], and this represents a significant shift from the traditional statistical signal processing methods [12]. They can generally be grouped into time-domain models, spectral-domain methods, adversarial learning frameworks, attention-based architectures, and diffusion-based models.

2.1. Time domain models

Recently, time-domain modeling methods have attracted prominent interest due to the fact they process raw waveforms directly and avoid explicit phase reconstruction needed by spectral-domain methods. Full convolutional architectures have been shown to achieve state-of-the-art performance in speech denoising and separation tasks [13]. DeMucs is an example of this, where the model uses an encoder–decoder structure with skip connections to retain temporal information [14]. Analogous to this, TasNet-based models have been shown to learn efficient time-domain representations and significantly outperform conventional STFT-based masking approaches [15].

Conv-TasNet further optimizes this paradigm to a fully convolutional end-to-end framework for speech separation and denoising [5].

2.2. Spectral and complex-valued models

While these spectral-domain approaches are still widely used owing to their solid theoretical basis [12, 13]. Complex-numbered neural networks have been proposed as a way to better represent phase information. Deep Complex U-Net has shown performance improvements by jointly modeling the magnitude and the phase [16]. The Deep Complex Convolutional Recurrent Network (DCCRN) builds on this concept by using a hybrid of convolutional and recurrent architectures to achieve significant robustness against non-stationary noise [17]. Although powerful, these methods are constrained by their dependency on STFT representations and thus are limited by trade-offs in both temporal and frequency resolution.

2.3. Adversarial learning-based methods

Generative adversarial networks (GANs) have been extensively investigated for improving perceptual quality in transformed speech [18]. MetricGAN was a framework that directly optimized perceptual metrics such as PESQ with the help of a discriminator network [11]. MetricGAN+ [19] provides stability during training and improves the perceptual score. While effective, GAN-based methods involve careful training and can be sensitive to hyperparameter tuning.

2.4. Transformer-based models

Transformer-based architectures have drawn the most attention recently for long-range dependency modeling capabilities through self-attention mechanisms. Recent improvements have come from dual-path transformer networks and Conformer-based models [20, 21] which could achieve further improvements in intelligibility, clarity, or robustness over highly non-stationary environments [22,23]. These models, however, depend on massive corpora and computational power.

2.5. Diffusion-based models

Recent advancements in speech enhancement are focused on diffusion-based models. Iterative refinement is employed by models like DiffWave to gradually reconstruct clean speech signals from noise inputs [24, 25]. Although these methods give high-quality results, their multi-step sampling process incurs an increase in inference latency and computational expense [26,27].

2.6. Non-autoregressive and efficient generative models

Many works focused on making inference more efficient to overcome the limitations of autoregressive methods. Parallel WaveNet proposed a distillation-based method for faster waveform generation [28]. However, these approaches still introduce additional training complexity [29,30].

Despite recent advancements, existing approaches suffer from trade-offs in computational efficiency, inference latency, and scalability that hinder their

deployment for real-time applications. Because time-domain models need deep architectures, spectral models have representation limits, and generative approaches bear more computational burden. In contrast, the proposed work presents a non-causal WaveNet architecture with parallel target-field prediction to allow for an efficient end-to-end speech denoising without impairing robustness to both stationary and non-stationary noise conditions [31]. **Table 1** shows the summary of existing speech enhancement models.

Table 1. Summary of existing speech enhancement models.

Method type	Representative models	Strengths	Limitations
Time-domain	Conv-TasNet, Demucs	End-to-end learning, strong performance	High computational cost
Spectral/Complex	DCCRN, Deep Complex U-Net	Good phase modeling	STFT limitations, resolution trade-off
Adversarial (GAN)	MetricGAN, MetricGAN+	High perceptual quality	Training instability, tuning complexity
Transformer-based	Conformer, Dual-Path	Long-range dependency modeling	Data-hungry, high computational demand
Diffusion-based	DiffWave	High-quality reconstruction	High inference latency
Proposed Method	Non-causal WaveNet	Efficient, parallel, robust	Requires computational resources

3. WaveNet architecture for speech denoising

3.1. WaveNet architecture elements

WaveNet is a variation of the Pixel CNN model that is commonly used for image generation. While maintaining the advantages of Pixel CNN, including design for conditioning, output as discrete softmax distributions, causation, and gating between units, it introduces dilated convolutions and non-linear quantization methods. **Figure 1** shows the WaveNet architectural elements.

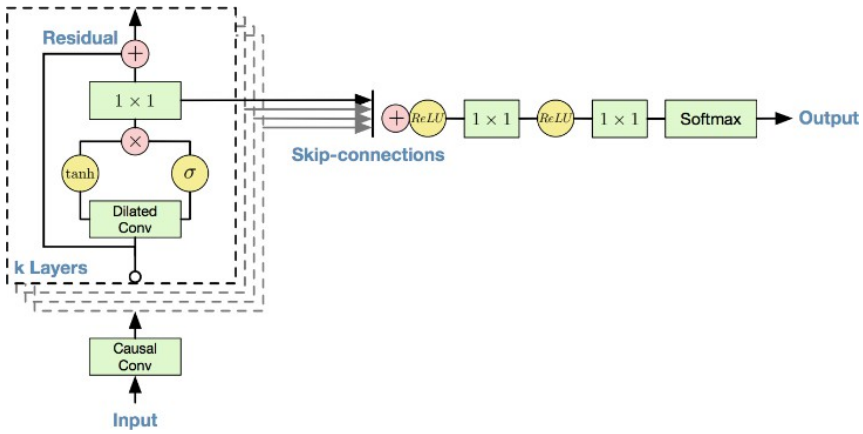


Figure 1. WaveNet architectural elements.

The following are some important characteristics of WaveNet.

3.1.1. Gated-activation units

Sigmoidal gates, like in LSTM networks, govern the activation in each layer of the WaveNet architecture. The gated activation unit is defined as [8],

$$O_t' = tanh\,(W_f * x_t)\ \sigma\,(W_g * x_t).$$

Where * represents convolution and Θ represents element-wise multiplication, *f* represents Filter, *t* represents input time, *t'* represents output time, and *g* represents gate

indices. 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.

3.1.2. Causal dilated convolutions

Due to the strong temporal dependence of audio waveforms, a large receptive field is needed. We could just increase the depth of a simple convolution stack to increase the receptive field, but this necessitates the use of an extremely deep network, which is not ideal. Therefore, we use dilated convolution in turn. Padding on the left and right unevenly according to the dilation so that causality is preserved and activations do not flow back in time. **Figure 2** illustrates the causal convolution, standard convolution, and causal dilated convolution.

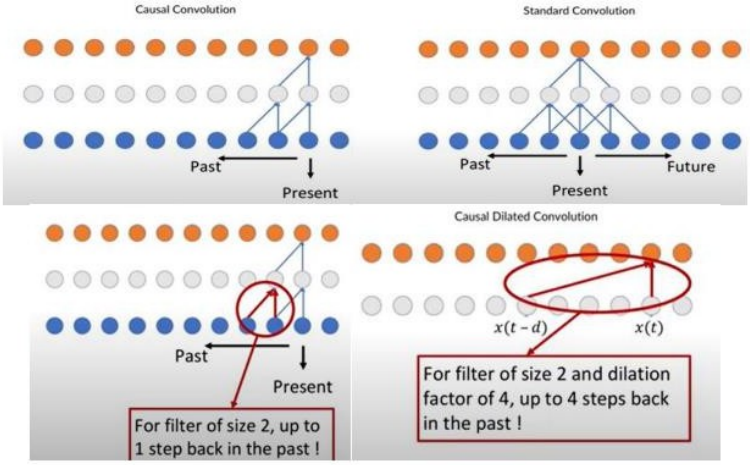


Figure 2. Casual convolution, standard convolution, and casual dilated convolution.

3.1.3. μ-law quantization

The original audio signal is transformed using a μ-law companding to prevent a large softmax output space (e.g., 65,536 possible for 16-bit audio). It performs a compressed dynamic range transformation on the signal and reduces quantization levels to 256. Such non-linear quantization is better than linear quantization of comparable complexity, providing greater precision for lower amplitude signals which are more perceptually significant.

3.1.4. Skip connections

These have two benefits. For example, they simplify the training of deep models. Second, they enable information to be passed directly between each layer and the final levels. It allows the network to use information obtained at multiple levels in arriving at its final output.

3.1.5. Context stacks

In order to add to the depth of the network without significantly swelling the receptive field, WaveNet employs context stacks. In these stacks, the necessary depth of the network is obtained by stacking layers with different dilation factors.

3.2. Modifications to WaveNet for denoising

The following modifications can be applied to WaveNet for Speech Denoising as shown in **Figure 3**:

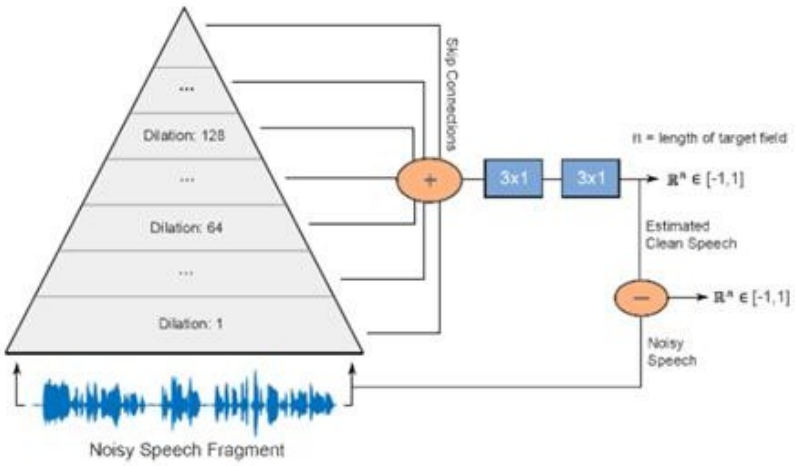


Figure 3. Modifications to WaveNet for denoising.

3.2.1. Non-causality

In speech enhancement, future samples are generally available and can therefore be used for better predictions. Wavenet's asymmetrical dilated convolution pattern can be enhanced by raising the filter length to 3 and applying symmetrical padding around each dilation layer. This displacement of the target sample at the center of the receptive field increases by twice its context and avoids considerations about causality.

3.2.2. Real-valued predictions

Instead of discrete softmax output, WaveNet could exploit real-valued predictions for speech denoising. This prevents creating artifacts in the denoised signal that result from modeling multi-modal distributions.

3.2.3. Energy-conserving loss

An energy-preserving loss function is used, which consists of two terms: the difference between original and denoised speech signals and the difference between original and denoised noise signals. This loss formulation ensures the model can perform the denoising task that we want.

3.2.4. Discriminative model

The altered model is no longer strictly autoregressive and does not even explicitly simulate a probability distribution. Instead, it denoises the cleaned signal directly. The model is supervised trained on minimizing a regression loss.

3.2.5. Conditioning

The model may be conditioned on a binary-valued scalar signal corresponding to the speaker identity. This value of the condition acts as a bias for each conv layer. different speaker identifiers in the speaker identities of the training set (but unknown speakers use an auxiliary code from mode-zero).

3.2.6. Noise-only data augmentation

We can also add a data augmentation technique with a parameter such as if we were to use a fraction of the training set with only background noise.

3.2.7. Denoising step

While denoising, the model processes a noisy signal in consecutive segments and successively concatenates each segment. If the available RAM is large enough, all of this sample can be cleaned in one pass.

Such modifications allow the adapted WaveNet model to efficiently cope with the Speech Denoising problem in a parallelizable and end-to-end fashion.

3.3. Problem formulation

Noisy speech signal is defined as,

$$x(t) = s(t) + n(t).$$

Where, $x(t)$ denotes observed noisy speech, $s(t)$ denotes clean speech signal, $n(t)$ denotes additive noise.

The purpose of speech denoising is to learn a mapping function:

$$\widehat{s}(t) = f\theta(x(t)). \tag{1}$$

Where, $f\theta(.)$ represents a learnable neural network with parameters θ, $\widehat{s}(t)$ is the estimated clean speech.

The proposed method, in contrast to spectral-domain methods, directly learns an end-to-end time-domain regression model with no explicit need for time-frequency transformation.

3.4. Non-causal WaveNet architecture

The conventional WaveNet is autoregressive and causal [8]:

$$p(x) = \prod_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1}). \tag{2}$$

But for denoising in speech, future samples are available during inference. Thus, the autoregressive constraint is relaxed, and symmetric padding is employed.

The receptive field is given by the below [8]:

$$R = 1 + 2\sum_{l=0}^{L-1} (k-1)d_l. \tag{3}$$

Where k = filter size, d_l = dilation factor in layer l, i.e., L = number of layers.

This supports modeling both past and future context,

$$\widehat{s}(t) = f_\theta(x\left(t - \frac{R}{2}\right), \ldots, x(t + \frac{R}{2}). \tag{4}$$

3.5. Gated residual blocks

Each residual block computes [8]:

$$Z = tanh(W_f * x) \odot \sigma(W_g * x). \tag{5}$$

Where, W_f and W_g are convolution filters, $*$ denotes convolution, $\odot$ element-wise multiplication.

Residual and skip connections are applied to improve gradient flow [8]:

$$x_{l+1} = x_l + W_{res} * z. \tag{6}$$

3.6. Target field prediction

The network predicts a target field of length L_T instead of predicting a single sample at each forward pass.

$$\widehat{S_{t:t+L_T}} = f_\theta(X_{t:t+R}). \tag{7}$$

It eliminates the need for repetitive convolution processes per image while allowing inferences to be done in parallel, making it more computationally efficient.

3.7. Energy-conserving loss function

Traditional regression utilizes Mean Squared Error (MSE) [28]:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{t=1}^{N}(s(t) - \widehat{s(t)})^2. \tag{8}$$

We therefore define an energy-conserving loss that better reflects denoising performance.

$$\mathcal{L} = \alpha||s - \widehat{s}||_2^2 + \beta||n - (x - \widehat{s})||_2^2. \tag{9}$$

Where, α, β are the weighting coefficients. The first term is structure-preserving, meaning speech structure is preserved. The second term is for correctly cancelling the noise.

One of these objectives leads to balanced reconstructions of speech and noise removal.

4. Training and testing

4.1. Data set and preprocessing

The dataset employed in this work is NSDTSEA (Noisy Speech Database for Training Speech Enhancement Algorithms). It consists of a collection of noisy speech samples and their corresponding clean speech counterparts. The dataset encompasses a total of 11,527 training samples and 8,824 testing samples. These recordings feature the voices of 28 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 in the dataset 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. Hence, the pipeline is designed to be an end-to-end process. During the training phase, the model is trained using the noisy speech signals as input and the corresponding clean speech signals as references to denoise the noisy samples.

4.2. Model configurations

The model comprises 30 residual layers with exponentially growing dilation factors ranging from 1 to 512. This pattern is repeated three times. A 3 × 1 convolution is applied before the first dilated convolution to match the input's 1-channel to 128 channels, the number of filters in each residual layer. Skip connections use 1x1 convolutions with 128 filters. The skip connections are summed, followed by a ReLU activation. The final two convolutional layers are 3 × 1 in size with 2,048 and 256 filters, respectively, and they are not dilated. These layers are separated by a ReLU activation. The output layer, utilizing a 1 × 1 filter, linearly projects the feature map to a single-channel temporal signal. This configuration achieves a receptive field of 7,739 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 parameter settings:

1. Training

a. Batch_size: 10 (number of samples per batch);
b. Early_stopping_patience: 16 (number of epochs to wait without improvement in accuracy before stopping training);
c. Num_epochs: 50 (maximum number of training epochs);
d. Num_train_samples: 1,000 (number of training samples presented to the model in one epoch);
e. Num_test_samples: 100 (number of validation samples presented to the model in one epoch).

2. Model

a. Dilations: 9 (maximum dilation factor as an exponent of 2);
b. Num_stacks: 3 (number of context stacks in the architecture);
c. Target_field_length: 1,603 (desired length of the output).

3. Optimizer

a. Decay: 0.0 (rate of reducing the learning rate over time);
b. Epsilon: 1×10^{-8} (a small positive value to ensure stability and prevent division by zero);
c. Learning rate: 0.001 (determines the step size for updating weights at each iteration);
d. Momentum: 0.9 (moving average of gradients to be maintained);
e. Type: 'adam' (type of optimizer used).

The dataset is presented to the model using mini-batch gradient descent, where 1,000 randomly selected samples are used per epoch. The error is calculated as the average of 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.

5. Experimental evaluation

5.1. Evaluation metrics used

The metrics to be used for the evaluation of the performance of the model are Signal to Noise Ratio (SNR) and Mel Cepstral Distortion (MCD). Higher SNR indicates better noise suppression, and lower MCD indicates better speech reconstruction fidelity.

5.1.1. Signal to noise ratio (SNR)

The Signal-to-Noise Ratio (SNR) is a popular assessment metric in a variety of domains, including audio and signal processing, for determining the quality or fidelity of a signal in the presence of noise. It measures the ratio of desirable signal power to undesired noise power, providing an objective measure of signal clarity and noise distortion.

SNR is calculated using the formula [26],

$$SNR = 10\ \log_{10} \frac{P_{signal}}{P_{noise}}. \tag{10}$$

5.1.2. Mel cepstral distortion (MCD)

Mel cepstral distortion (MCD) quantifies the dissimilarity between two sets of mel cepstra. It serves as a metric for evaluating the fidelity of parametric speech synthesis systems, 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 MCD. The Euclidean distance between the two signals' MFCC vectors is then calculated. MCD is commonly expressed as mel cepstral distance per frame or per second [27].

$$MCD = \frac{10\sqrt{2}}{ln10} \frac{1}{T} \sum_{i=1}^{T} \sqrt{\sum_{i} \left(C_{ti} - \widehat{C_{ti}}\right)^2}. \tag{11}$$

5.2. Performance evaluation using classical methods

A noisy audio sample was given as input to different classical speech denoising methods, and the outputs were obtained as follows

Transcript: "Where there is a will, there is a way"

Figure 4 shows the comparison of clean and noisy speech signals. The top panel shows the clean speech waveform. The middle panel illustrates the speech signal corrupted by non-stationary noise, characterized by irregular fluctuations in both amplitude and temporal occurrence. The bottom panel presents the speech signal corrupted by stationary noise, where the noise characteristics remain relatively constant over time. This comparison highlights the temporal variability of non-stationary noise in contrast to the consistent nature of stationary noise, demonstrating the different noise conditions used in the experimental evaluation.

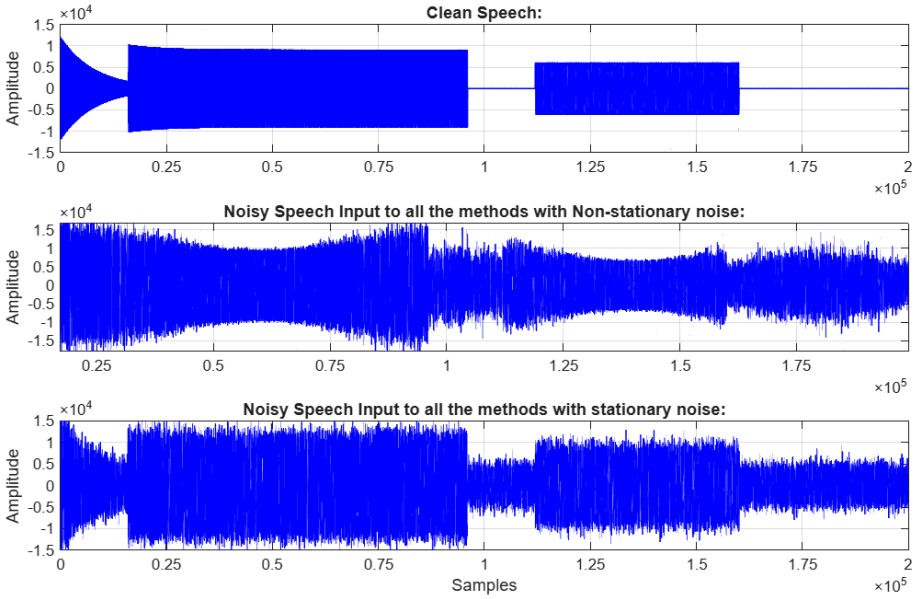


Figure 4. Comparison of clean and noisy speech signals.

5.2.1. Spectral subtraction

Figure 5 shows the spectral subtraction. The three plots show how a signal and its noise behave in the frequency domain using a Fast Fourier Transform (FFT). The Noisy Signal FFT displays the spectrum of the original signal combined with noise, where the true frequency components are partially hidden by a broad spread of noise across frequencies. The Noise Signal FFT shows the spectrum of the noise alone, which is mostly low in magnitude but spread across many frequencies, giving it that "background" appearance. The Noise-subtracted Spectrum represents the result after removing or reducing the noise component from the noisy signal; here, the important frequency peaks (the actual signal content) stand out much more clearly, especially around the higher-frequency region, demonstrating how noise subtraction improves signal clarity and makes dominant frequencies easier to identify.

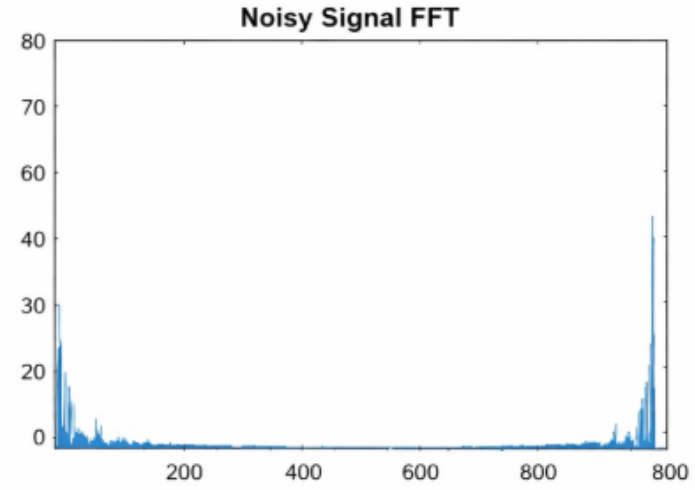


Figure 5. *Cont.*

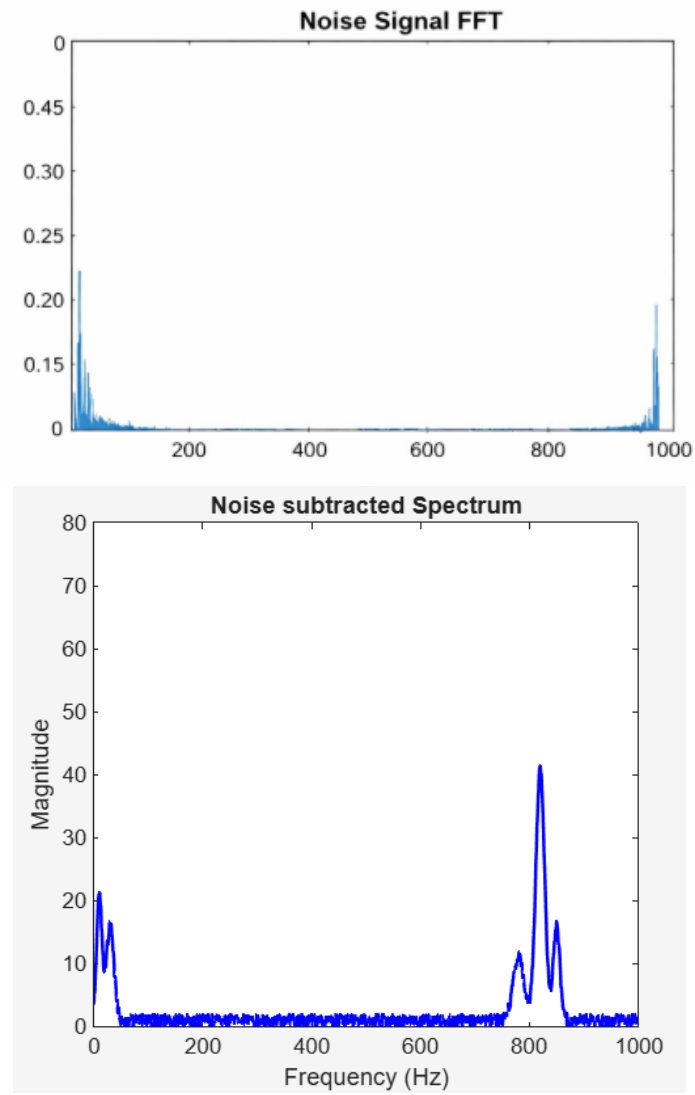


Figure 5. Spectral subtraction.

5.2.2. Weiner filtering

Wiener filtering in **Figure 6** improves speech quality and intelligibility by modeling the signal-to-noise ratio at each frequency and applying a corresponding gain that attenuates noise while preserving speech components.

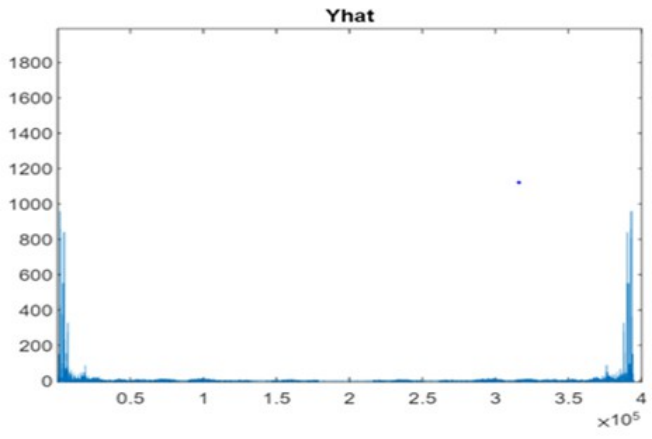


Figure 6. *Cont.*

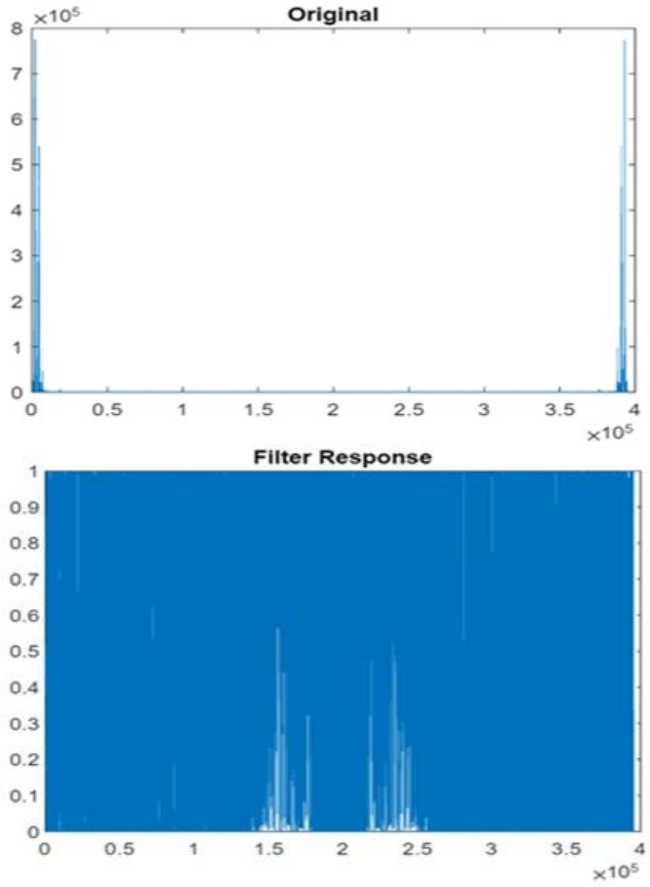


Figure 6. Weiner filtering.

The three plots show the effect of these transformations in the frequency domain. The Original plot displays the FFT of the "raw" input signal, with enormous peaks at both ends dominating the spectrum as expected; although there is a very strong frequency content being measured, it could also be interpreted as an artifact that could lead to a DC offset or mirrored high-frequency content. "Yhat" is the smoothed or filtered signal after some sort of reconstruction; it is much lower in magnitude than X, and some dominant peaks have been lost, so unwanted components (like noise or extreme spikes) have disappeared while assuming a useful structure. We can see in the Filter Response how this filter works. A high absolute value (close to 1) means that the frequency is passed through, while a low one (near 0) indicates that it is attenuated; the scattered lower values tell us that the filter has removed specific bands, and since those bands were being imposed on "Yhat," we are left with something cleaner without big peaks.

5.2.3. Spectral gating

In spectral gating shown in **Figure 7**, a threshold mask based on frequency is applied to the noisy spectrogram, masked areas are chosen to centralize around the regions dominated by speech, and it results in an attenuation of noise components.

These four spectrograms illustrate the corrupting effects of noise on a signal and then a partial restoration via masking/filtering. The noise signal shows the noise alone, smeared over time and frequencies, appearing as a somewhat uniform background. The Noisy Signal shows the original signal mixed with this noise; we can still see structured patterns in it (likely meaningful signal components), but it has been obscured by the

added noise. The Mask applied shows the same filter or mask in the time–frequency domain, where brighter (stronger) regions represent parts of the spectrum being retained, and weaker areas are being suppressed—this mask is formulated to retain necessarily signal-related features while attenuating noise-dominated areas. And finally, the recovered spectrogram shows what happens when we apply the mask to our noisy signal: we see that now the underlying structures are more visible again, but they have not been fully restored; some noise is still present, and some of the details of our signals may be missing, which represents a compromise between suppressing noise and keeping as much information as possible.

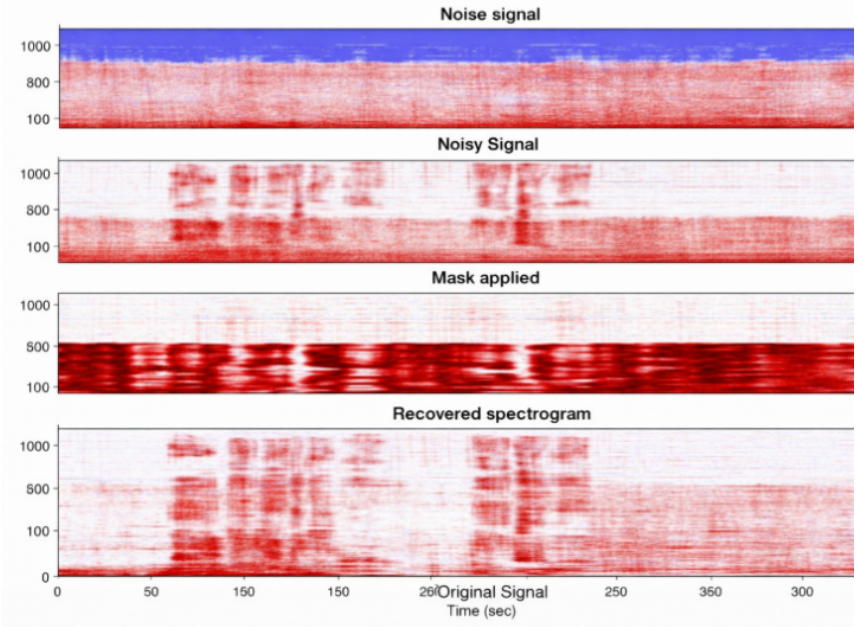


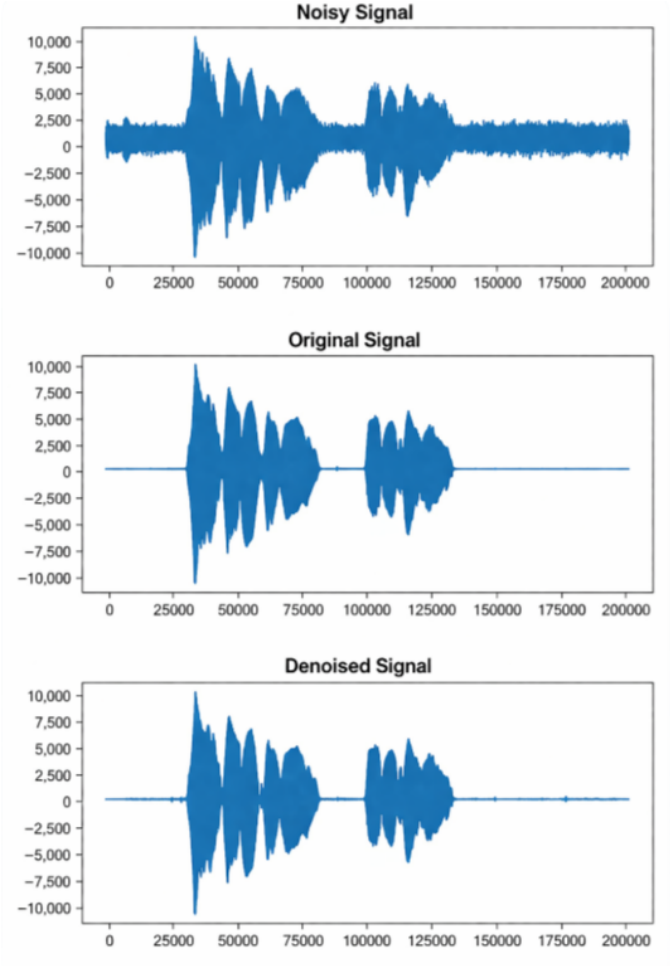
Figure 7. Spectral gating.

5.2.4. Speech denoising WaveNet

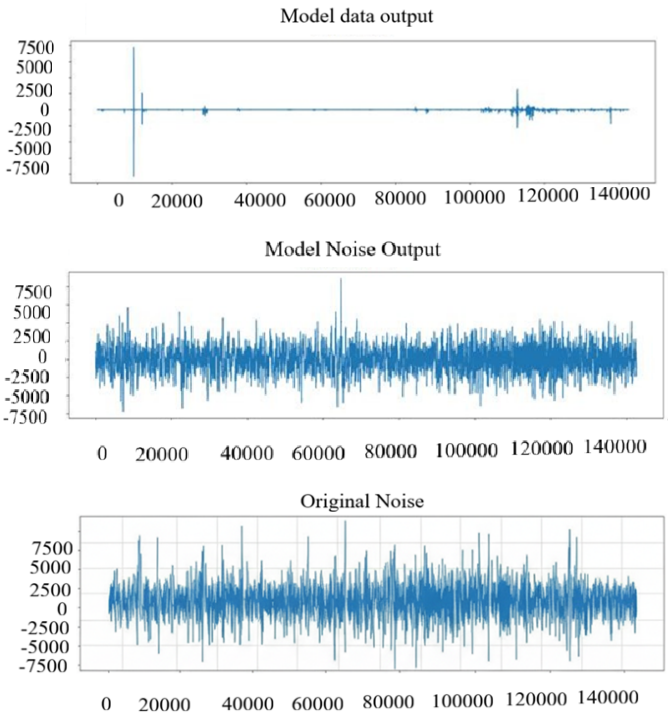
Three stages of a signal-processing pipeline are shown in **Figure 8a**. The Noisy Signal is the direct observation, in which a clean underlying waveform has been polluted by random noise, complicating the identifiability of any meaningful patterns. The Original Signal shows the clean ground-truth waveform without noise, where there is a clear structure —bursts of higher amplitude interspersed by quieter regions. The Denoised Signal is the model's attempt to recover a clean signal from the noisy input. This waveform closely follows the shape of the original signal while filtering out most of the background noise, even though some minor residual noises and slight smoothing may still be present.

Figure 8b has to do with how the model handles noise internally. The original noise represents the random noise added to the clean signal, which has the appearance of rapid, irregular fluctuations. The model data output shows the model's estimate of what represents the actual (true) signal component, which is not zero everywhere but almost so with exceptions where events are present, showing that the model identifies meaningful structure. The model noise output, the estimated noise of the model, is almost identical to the original noise pattern. Combined, these plots show that the model performs well in breaking down the input into signal and noise

components—recovering the relevant features while filtering out noise.



(a) Illustration of noisy signal, original signal, and denoised signal



(b) Illustration of model data output, model noise output, and original noise

Figure 8. Speech denoising WaveNet.

Classical methods, such as spectral gating and Wiener filtering, have been widely used for speech denoising. These methods are based on statistical signal processing techniques and often rely on assumptions about the statistical properties of the speech and noise signals. They typically operate in the spectral domain and aim to estimate the clean speech spectrum by exploiting the statistical characteristics of the noise. Classical methods are computationally efficient and can be effective in scenarios where the noise characteristics are well understood and stationary. However, they may struggle with handling non-stationary noise or in situations where the noise characteristics vary significantly, as shown in **Figure 9**.

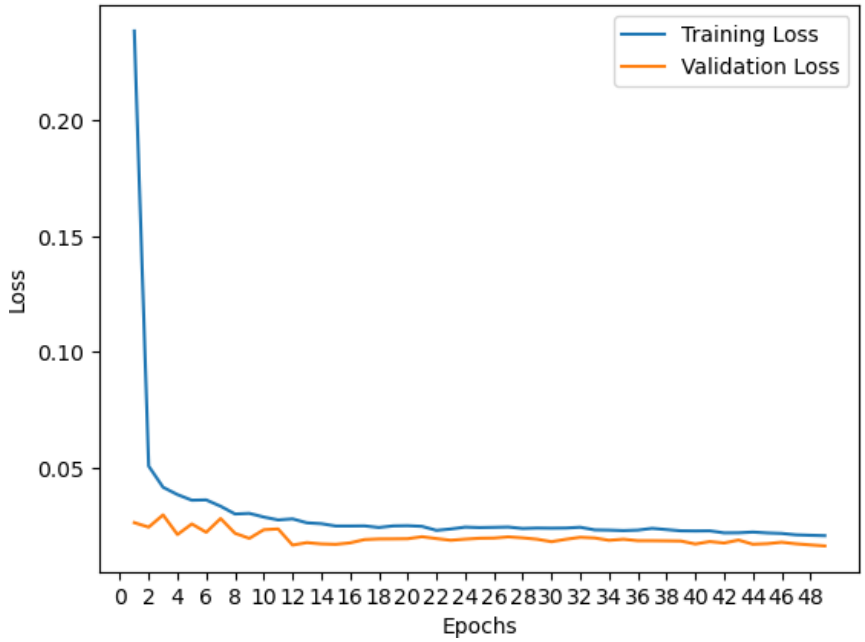


Figure 9. Training and validation loss.

On the other hand, the WaveNet-based method is a deep learning-based approach. It is a model that uses dilated convolutions to capture long-range dependencies in the speech signal. A WaveNet-based model is trained end-to-end on large amounts of speech data, allowing it to learn complex patterns and generate high-quality denoised speech. It operates directly in the time domain, which enables it to capture temporal variations and fine-grained details in the speech signal. A WaveNet-based model can handle non-stationary noise and adapt to varying noise characteristics without relying on explicit assumptions. However, it typically requires more computational resources and training data compared to classical methods.

5.2.5. Stationary noise results

An audio sample mixed with noise with SNRs of 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB is given as input to classical method algorithms spectral subtraction, Wiener filtering, and spectral gating, and also to a WaveNet-based model. The values of the performance evaluation metrics for various methods and different inputs with different SNRs are obtained as follows (**Tables 2** and **3**):

Table 2. Stationary noise (SNR).

Noise level	Using WaveNet	Wiener filtering	Spectral subtraction	Spectral gating
2.5dB	−0.1952	2.7570	2.5218	2.2031
7.5 dB	−0.0419	7.7873	7.7373	8.4283
12.5 dB	−0.0853	12.2648	12.2889	12.3437
17.5 dB	−0.5178	18.4330	17.4464	17.2617

Table 3. Stationary noise (MCD).

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

5.2.6. Non-stationary noise results

For low SNR levels, classical methods are slightly better than the proposed model under stationary noise conditions. This behavior is due to stationary noise having deterministic spectral properties, so this is where the statistical filtering methods shine. On the other hand, input SNR gets higher, performance gaps decrease (**Tables 4** and **5**).

Table 4. Stationary noise (SNR).

Noise level	Using WaveNet	Wiener filtering	Spectral subtraction	Spectral gating
2.5dB	−3.1747	2.6263	2.8397	2.5426
7.5 dB	9.0269	7.2069	8.6263	7.5711
12.5 dB	14.7213	12.4578	13.5001	13.3794
17.5 dB	19.2280	17.3524	18.2486	17.5745

Table 5. Non-stationary noise (MCD).

Noise level	Using WaveNet	Wiener filtering	Spectral subtraction	Spectral gating
2.5dB	21.0143	48.2019	24.6643	39.1009
7.5 dB	12.5769	42.7465	20.3668	43.2238
12.5 dB	11.1937	37.7465	19.4356	40.2238
17.5 dB	9.4091	38.3308	17.5753	38.3517

5.3. Discussion

The experimental results provide some important insights into the behaviour and effectiveness of the proposed non-causal WaveNet model. First, the proposed method shows superior performance compared with classical methods in terms of SNR and MCD under non-stationary noise conditions. This owes to the models' ability to learn complex temporal dependencies directly from raw waveforms, which enables them to quickly adapt to fast-evolving noise configurations. On the other hand, typical approaches like Wiener filtering and spectral subtraction make stationary noise assumptions, which limit performance in dynamic conditions. For stationary noise, classical algorithms are competitive or slightly better at lower values of SNR. However, this behaviour is expected because the statistics are the most appropriate for noise that has stable spectral characteristics. But at a higher SNR, the performance of classical methods and our model is showing a much closer gap, which explains why

a learning-based method can generalise the noise range. One essential reason for the performance of our method is the non-causal architecture with symmetric receptive fields. The model can include both past and future context in the reconstruction, allowing for more accurate reconstructions than causal architectures that only take the past into account. Meanwhile, the target-field prediction mechanism produces different output samples in one single forward path, which saves costly redundant computations of autoregressive inference sample by sample. The proposed approach thus represents a novel trade-off between performance and computational efficiency in relation to modern deep learning models such as Conv-TasNet, DCCRN, and MetricGAN+. Although these models tend to offer superior denoising performance, they generally necessitate deeper architectures, sophisticated training techniques, or iterative processing. On the other hand, the proposed framework produces comparable performance with a simpler architecture and parallel inference in the testing phase. However, the proposed method also has certain limitations. The performance of the proposed model with respect to stationary noise is relatively less compared to the classical method, mainly because the stationary noise statistics are not explicitly modeled during training. Furthermore, although model reuse could reduce inference redundancy in theory, it is still impractical because the model consumes too much computational resource per inference since it needs to operate in the time domain. In summary, the results show that the proposed non-causal WaveNet model performs well in non-stationary noise conditions and achieves a nice trade-off of robustness, efficiency, and scalability for real-world voice enhancement applications.

Results presented in **Figures 5–9** show that minima tracking produces speech spectrograms that clean up beautifully over the noisy spectrograms, thereby demonstrating its utility in speech enhancement. The spectral subtraction method is shown in **Figure 5**, which is used as a baseline for comparison. Spectral subtraction can reduce a portion of background noise, but it also has two main drawbacks: it creates audible musical noise artifacts and it cannot keep spectral details of the speech signal well. This shortcoming drives the demand for more advanced methods.

The comparative analysis in **Figures 6** and **7** shows that the proposed model shows gains against traditional techniques. In particular, the vocal enhancement produces improved speech signals that suppress the noise while preserving the signal. In contrast to these traditional approaches, the proposed approach achieves a better trade-off between noise suppression and speech distortion; this gain is demonstrated in perceptual quality scores.

Figure 8 shows the performance of the WaveNet-based speech denoising model. The comparison of approaches can be seen from the above plot, where the classical techniques are mistakenly running behind, as the deep learning-based approach is majorly outperforming them by capturing those complex temporal patterns of speech signals. The restored waveform exhibits more gradual changes and retains speech more accurately, confirming that our model outperforms in non-stationary noise conditions.

Furthermore, the loss curves for training and validation are shown in **Figure 9**, which gives an indication of how the model learned. Training loss and validation loss both started to drop linearly at the same time and continued to do so, indicating success

in the case of no or very little overfitting. The closeness of the two curves indicates that, for unseen data, the model has good generalization capacity.

Overall, the results from **Figures 5–9** confirm that the proposed framework provides substantial improvements in speech quality and noise reduction over classical approaches. This can help go beyond the limitations of classical methods and enable deep learning-based techniques to make the model robust and useful in practical applications where speech enhancement is needed.

6. Conclusion

This paper proposed a supervised non-causal WaveNet-based approach to speech denoising with parallel target-field prediction and an energy-conserving loss function. The suggested model removed autoregressive limitations and allowed symmetric context modeling by employing temporal information in both the past and the future. Moreover, the target-field prediction mechanism improved inference efficiency by enabling multiple samples to be predicted in one forward pass and reducing unnecessary calculations. The performance of the proposed approach was validated with extensive experiments in terms of stationary and non-stationary noise.

The experimental results demonstrated that the proposed WaveNet-based model achieved impressive performance at all SNR levels under non-stationary noise conditions. On the other hand, classical approaches like Wiener filtering and spectral subtraction showed improved performance in stationary noise environments due to their statistical noise assumptions but poor performance in dynamic environments. Spectral gating yielded moderate performance in both scenarios and was sensitive to threshold selection.

Experimental results show that the proposed framework exceeds classical baselines with more than 1.22 dB improvement in output SNR and a 46% decrease in Mel Cepstral Distortion (MCD) under non-stationary noise conditions.

Our future work will attempt to reduce the computational complexity of this model using model compression techniques, including pruning, quantization, and lightweight architectural modifications. Moreover, hybrid learning mechanisms combining time-domain and frequency-domain representations might be effective under stationary noise. It will also be benefited by extending the training dataset such that it covers a wider spectrum of real-world noise environments and performing evaluation on a benchmark dataset to improve its generalization ability. And last, but not the least, to optimize the model for real-time deployment on embedded and edge devices, to make a practical speech communication system and assistive technologies happen.

Author contributions: Conceptualization: JS, SUR and SMH; methodology: JJS, SUR and SMH; software implementation: JS, SUR and SMH; original draft: JS, SUR and SMH. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Kingdom University, Bahrain, under Grant KU-SRU-2024-13.

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Data availability statement: The data supporting the findings of this study are publicly available through the NSDTSEA dataset. Additional implementation details are available from the corresponding author upon reasonable request.

Acknowledgment: The authors would like to acknowledge their respective organizations for providing all types of support in completing this research work.

Conflict of interest: The authors declare no conflict of interest.

AI use statement: During the preparation of this manuscript, the authors used ChatGPT solely for language refinement. No AI tools were used for data analysis, interpretation, or generation of scientific content. All outputs were critically reviewed and edited by the authors. The authors take full responsibility for the integrity and accuracy of the work.

References

1. Hu Y, Loizou PC. Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Transactions on Audio, Speech, and Language Processing. 2008; 16(1): 229–238. doi: 10.1109/TASL.2007.911054
2. Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979; 27(2): 113–120. doi: 10.1109/TASSP.1979.1163209
3. Wang Y, Wang D. Towards Scaling Up Classification-Based Speech Separation. IEEE Transactions on Audio, Speech, and Language Processing. 2013; 21(7): 1381–1390. doi: 10.1109/TASL.2013.2250961
4. Pascual S, Bonafonte A, Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network. In: Proceedings of the Interspeech 2017; 20–24 August 2017; Stockholm, Sweden. pp. 3642–3646. doi: 10.21437/Interspeech.2017-1428
5. Luo Y, Mesgarani N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019; 27(8): 1256–1266. doi: 10.1109/TASLP.2019.2915167
6. Hu Y, Liu Y, Lv S, et al. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. In: Proceedings of the Interspeech 2020; 25–29 October 2020; Shanghai, China. pp. 2472–2476. doi: 10.21437/Interspeech.2020-2537
7. Pandey A, Wang D. A New Framework for CNN-Based Speech Enhancement in the Time Domain. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019; 27(7): 1179–1188. doi: 10.1109/TASLP.2019.2913512
8. van den Oord A, Dieleman S, Zen H, et al. WaveNet: A Generative Model for Raw Audio. arXiv preprint. 2016. doi: 10.48550/arXiv.1609.03499
9. Qian K, Zhang Y, Chang S, et al. Speech Enhancement Using Bayesian Wavenet. In: Proceedings of the Interspeech 2017; 20–24 August 2017; Stockholm, Sweden. pp. 2013–2017. doi: 10.21437/Interspeech.2017-1672
10. Rethage D, Pons J, Serra X. A Wavenet for Speech Denoising. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 15–20 April 2018; Calgary, AB, Canada. pp. 5069–5073. doi: 10.1109/ICASSP.2018.8462417
11. Fu SW, Liao CF, Tsao Y, et al. MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement. arXiv preprint. 2019. doi: 10.48550/ARXIV.1905.04874
12. Oppenheim AV, Schafer RW. Digital Signal Processing. Prentice-Hall; 1975.
13. Liu C, Wang L, Dang J. Masking based Spectral Feature Enhancement for Robust Automatic Speech Recognition. In: Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA); 27–29 June 2020; Dalian, China. pp. 287–291. doi: 10.1109/ICAICA50127.2020.9181915

14. Défossez A, Usunier N, Bottou L, et al. Music Source Separation in the Waveform Domain. arXiv preprint. 2019. doi: 10.48550/ARXIV.1911.13254
15. Lin J, Van Wijngaarden AJDL, Wang KC, et al. Speech Enhancement Using Multi-Stage Self-Attentive Temporal Convolutional Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021; 29: 3440–3450. doi: 10.1109/TASLP.2021.3125143
16. Choi HS, Kim JH, Huh J, et al. Phase-aware Speech Enhancement with Deep Complex U-Net. arXiv preprint. 2019. doi: 10.48550/ARXIV.1903.03107
17. Lv S, Hu Y, Zhang S, et al. DCCRN+: Channel-wise Subband DCCRN with SNR Estimation for Speech Enhancement. arXiv preprint. 2021. doi: 10.48550/arXiv.2106.08672
18. Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.
19. Fu SW, Yu C, Hsieh TA, et al. MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. In: Proceedings of the Interspeech 2021; 30 August 30–3 September 2021; Brno, Czech Republic. pp. 201–205. doi: 10.21437/Interspeech.2021-599
20. Chen J, Mao Q, Liu D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. In: Proceedings of the Interspeech 2020; 25–29 October 2020; Shanghai, China. pp. 2642–2646. doi: 10.21437/Interspeech.2020-2205
21. Gulati A, Qin J, Chiu CC, et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In: Proceedings of the Interspeech 2020; 25–29 October 2020; Shanghai, China. pp. 5036–5040. doi: 10.21437/Interspeech.2020-3015
22. Luo Y, Chen Z, Yoshioka T. Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. In: Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 4–8 May 2020; Barcelona, Spain. pp. 46–50. doi: 10.1109/ICASSP40776.2020.9054266
23. Nachmani E, Adi Y, Wolf L. Voice Separation with an Unknown Number of Multiple Speakers. arXiv preprint. 2020. doi: 10.48550/arXiv.2003.01531
24. Kong Z, Ping W, Huang J, et al. DiffWave: A Versatile Diffusion Model for Audio Synthesis. arXiv preprint. 2020. doi: 10.48550/ARXIV.2009.09761
25. Richter J, Welker S, Lemercier JM, et al. Speech Enhancement and Dereverberation with Diffusion-Based Generative Models. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023; 31: 2351–2364. doi: 10.1109/TASLP.2023.3285241
26. Welker S, Richter J, Gerkmann T. Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain. In: Proceedings of the Interspeech 2022; 18–22 September 2022; Incheon, South Korea. pp. 2928–2932. doi: 10.21437/Interspeech.2022-10653
27. Emura S. Estimation of Output SI-SDR Solely from Enhanced Speech Signals in Diffusion-Based Generative Speech Enhancement Method. In: Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO); 26-30 August 2024; Lyon, France. pp. 236–240. doi: 10.23919/EUSIPCO63174.2024.10714970
28. van den Oord A, Li Y, Babuschkin I, et al. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. arXiv preprint. 2017. doi: 10.48550/ARXIV.1711.10433
29. Braun S, Tashev I. Data augmentation and loss normalization for deep noise suppression. arXiv preprint. 2020. doi: 10.48550/ARXIV.2008.06412
30. Défossez A. Hybrid Spectrogram and Waveform Source Separation. arXiv preprint. 2022. doi: 10.48550/arXiv.2111.03600
31. Kubichek R. Mel-cepstral distance measure for objective speech quality assessment. In: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing; 19–21 May 1993; Victoria, BC, Canada. pp. 125–128. doi: 10.1109/PACRIM.1993.407206