electronics

MDPI

Article

Supervised Single Channel Speech Enhancement Method Using UNET

Md. Nahid Hossain [1], Samiul Basir [1], Md. Shakhawat Hosen [1], A.O.M. Asaduzzaman [1], Md. Mojahidul Islam [1], Mohammad Alamgir Hossain [1] and Md Shohidul Islam [1,2,*]

[1] Department of Computer Science and Engineering, Islamic University, Kushtia 7003, Bangladesh
[2] Hong Kong Centre for Cerebro-Cardiovascular Health Engineering (COCHE), The City University of Hong Kong, Kowloon, Hong Kong
* Correspondence: shohid7@cse.iu.ac.bd; Tel.: +88-01-749529735

Abstract: This paper proposes an innovative single-channel supervised speech enhancement (SE) method based on UNET, a convolutional neural network (CNN) architecture that expands on a few changes in the basic CNN architecture. In the training phase, short-time Fourier transform (STFT) is exploited on the noisy time domain signal to build a noisy time-frequency domain signal which is called a complex noisy matrix. We take the real and imaginary parts of the complex noisy matrix and concatenate both of them to form the noisy concatenated matrix. We apply UNET to the noisy concatenated matrix for extracting speech components and train the CNN model. In the testing phase, the same procedure is applied to the noisy time-domain signal as in the training phase in order to construct another noisy concatenated matrix that can be tested using a pre-trained or saved model in order to construct an enhanced concatenated matrix. Finally, from the enhanced concatenated matrix, we separate both the imaginary and real parts to form an enhanced complex matrix. Magnitude and phase are then extracted from the newly created enhanced complex matrix. By using that magnitude and phase, the inverse STFT (ISTFT) can generate the enhanced speech signal. Utilizing the IEEE databases and various types of noise, including stationary and non-stationary noise, the proposed method is evaluated. Comparing the exploratory results of the proposed algorithm to the other five methods of STFT, sparse non-negative matrix factorization (SNMF), dual-tree complex wavelet transform (DTCWT)-SNMF, DTCWT-STFT-SNMF, STFT-convolutional denoising auto encoder (CDAE) and casual multi-head attention mechanism (CMAM) for speech enhancement, we determine that the proposed algorithm generally improves speech quality and intelligibility at all considered signal-to-noise ratios (SNRs). The suggested approach performs better than the other five competing algorithms in every evaluation metric.

Keywords: speech enhancement (SE); short-time Fourier transforms (STFT); convolutional denoising auto encoder (CDAE); dual-tree complex wavelet transform (DTCWT); sparse non-negative matrix factorization (SNMF); casual multi-head attention mechanism (CMAM); UNET

Citation: Hossain, M.N.; Basir, S.; Hosen, M.S.; Asaduzzaman, A.O.M.; Islam, M.M.; Hossain, M.A.; Islam, M.S. Supervised Single Channel Speech Enhancement Method Using UNET. *Electronics* **2023**, *12*, 3052. https://doi.org/10.3390/electronics12143052

Academic Editors: Maysam Abbod and Gwanggil Jeon

Received: 23 February 2023
Revised: 18 April 2023
Accepted: 20 April 2023
Published: 12 July 2023

1. Introduction

The presence of ambient noise and other interfering signals can often distort speech in communication systems making it difficult to understand. By removing this deficiency, the area of speech enhancement (SE) [1] employs signal processing methods to increase the quality of speech signals. When the signal-to-noise ratio (SNR) is high, several conventional techniques, including the Wiener filter [2], spectral subtraction [3], vocal activity detection and others perform well.

A recent study introduces a new architecture called Fast FullSubNet [4] for real-time single-channel speech enhancement. This architecture builds upon the success of the original FullSubNet which achieved excellent performance on the Deep Noise Suppression

Challenge dataset. The Fast FullSubNet improves upon the original by optimizing its processing speed, making it more suitable for real-time applications.

In another study, a new method for single-channel speech enhancement using an improved progressive deep neural network and masking-based harmonic regeneration was developed [5]. The method consists of two stages: a progressive deep neural network for speech denoising, followed by masking-based harmonic regeneration to enhance the speech signal.

For several years, researchers have utilized nonnegative matrix factorization (NMF) to improve single-channel speech signals. NMF was first introduced by Paatero and Tapper [6], and later, Lee and Seung [7] proposed its application in speech enhancement (SE). NMF belongs to a set of multivariate analytical techniques that decompose a matrix into two nonnegative matrices based on its component parts and weights.

Uhlich et al. [8] proposed an alternative approach that employs only Fully Connected Layers (FCL) and accepts input in the form of multiple frames of the magnitude spectrogram of the mixture. This technique models timbre characteristics over different time intervals. Although these methods have been successful, they do not fully exploit the time-frequency properties of a particular region. Instead, they depend on long-term global characteristics that extend across the entire frequency range.

These findings imply that a variety of conventional signal processing issues may benefit greatly from the use of machine learning methods. A Deep Neural Network (DNN)-based system was trained to lower the root mean square of loud speech in a logarithmic magnitude spectrum in one publication [9]. Without adding "musical noise", this system demonstrated highly promising noise reduction capabilities for non-stationary noise. The paper aimed to enhance a prior study employing DNN-based speech enhancement [10] to build a more complete system. Model training had to be accelerated by this program utilizing a GPU. Rather than simply repeating previous findings, this paper attempted to analyze the system and the effect of different types on performance. Several potential alternate settings were investigated in an attempt to interpret some of the learned parameters using an understanding of the input and output signals.

The Generalized Sidelobe Canceller (GSC) [11] is a commonly used method for noise reduction that combines signals from multiple microphones to create a reference signal, which is then used to cancel out noise from the target signal. However, the GSC can sometimes introduce distortion and phase errors into the target signal. The Phase Error Filter (PEF) is used to control the phase of the reference signal and minimize phase errors. The PEF is designed to estimate the phase of the target signal and adjust the phase of the reference signal accordingly. The combination of the GSC and PEF has the potential to improve the performance of systems used for signal detection [12], speech recognition, hearing aids and other applications that require accurate signal processing.

The paper by Mohammadiha et al. [13] proposes a method for enhancing speech signals by using nonnegative matrix factorization (NMF) [14] in both supervised and unsupervised settings. In the supervised setting, a dictionary of spectral bases is learned from clean speech signals and used to enhance noisy speech signals [15]. In the unsupervised setting, NMF is used to separate speech signals from background noise by factorizing the spectrogram [16] of a noisy speech signal into a speech basis matrix and a noise basis matrix.

Another recent work proposes a new model for real-time single-channel speech enhancement using a causal attention mechanism [17]. The authors introduce the importance of speech enhancement and review previous methods before presenting their model architecture which includes an encoder, decoder and causal attention mechanism. The authors also used a weighted loss function in both time and frequency domains to guide the optimization direction of the training.

This study aimed to identify strategies for enhancing speech recognition in noisy surroundings using a single microphone. We created a complicated matrix from speech recordings in noise using the STFT to mimic real-world situations. Next, we divided this matrix into real and imaginary components before putting them into the proposed model

UNET. We aimed to increase the accuracy of our findings by including information from both the imaginary and real sections of the signal. With DNN serving as the learning machine, this study aimed to employ supervised learning to discriminate between the board voice and background noise or reverberation. The design of neural networks and acoustic characteristics are only two examples of the several facets of supervised speech production that are examined.

Notations: x&X (small & capital), **x** (small bold), **X** (capital bold), *X* (capital italic) and ***X*** (bold capital italic) denote variable, vector, matrix, function and method, respectively.

2. Literature Review

The following discussion covers the most modern deep learning techniques such as DNN, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) that try to address the problem of voice improvement.

2.1. Research and Development of DNNs

The Feed Forward Neural Network is another name for the Deep Neural Network, often referred to as a multilayer perceptron. Every node in this network is linked to the layers above it, making it completely connected. According to a technique proposed by Karjol et al. [18], the average PESQ for observed and unseen noise on the TIMIT corpus is 2.65 and 2.19, respectively.

A recent study by Zhentao Lin et al. [19] provided an overview of existing works on speech enhancement and federated learning (FL) and their limitations. It introduced a proposed approach, the self-adaptive noise distribution network for speech enhancement (SASE) which utilizes a complex-valued time-frequency gate attention mechanism and a self-adaptive Gaussian unitary ensemble attention block to address data heterogeneity in FL. The section also discusses the CommonVoice dataset with noise and proposes loss-based and PESQ-based optimization weighting strategies to update the server model with a large-scale heterogeneous dataset intelligently.

In another recent study by Shubo Lv et al. [20], the authors proposed a novel multi-channel denoising neural network called Spatial-DCCRN for speech enhancement in the STFT domain. The proposed model extends S-DCCRN and benefits from both local and global channel information processing. An angle feature extraction module is adopted to extract frame-level angle features to assist the network in perceiving spatial information more accurately. Additionally, a masking and mapping filtering method is employed to replace the traditional filter-and-sum operation. The proposed model has outperformed several competitive models on the L3 DAS22 Challenge dataset and achieved superior performance compared to the state-of-the-art MIMO-UNet model in multiple evaluation metrics on the multi-channel ConferencingSpeech2021 Challenge dataset.

In a recent study by Sivaramakrishna and Yechuri [21], the authors discussed various conventional and deep learning-based methods for speech enhancement and highlighted the limitations of existing methods in fully exploiting contextual information from multiple scales. They proposed a new model, ECAD3 MUNet, which leverages multi-scale feature extraction blocks with D3 Net and efficient channel attention to improve information flow while addressing aliasing problems associated with dilated convolution in DenseNets models. The proposed model aims to improve the performance of speech enhancement systems by addressing the limitations of existing methods.

In addition to deep learning and GAN-based approaches, some recent studies have explored the use of non-negative matrix factorization (NMF) for speech enhancement. For instance, a study by Bao et al. [22] proposed an NMF-based method for speech enhancement that uses a weighted constraint on the magnitude spectrogram to preserve speech harmonics. The experimental results showed that the proposed method outperformed several baseline methods in terms of objective measures and subjective listening tests.

2.2. RNN-LSTM in Speech Recognition

When dealing with sequence-based data such as voice signals, RNN-LSTM can manage the context information. Information from both the previously hidden layer and the current stage is used by this network. Mass et al. [23] used RNN to denoise the corrupted features such as MFCC. RNN-based SE was determined to be more successful than DNN with different SNR levels.

A framework employing LSTM was presented by Gao et al. [24] to improve the performance of DNN-based speech in low SNR to solve the problems of noisy multi-channel speech [25], reverberation [26] and very non-stationary additive noise. RNN-LSTM significantly improved speech denoising but its learning parameters are challenging and require more effort to master.

2.3. CNNs in Speech Enhancement

Researchers studying speech signal processing are interested in convolutional neural networks [27]. In comparison to RNN and the industry-standard FFNN [28], CNN may be seen as being more effective. Park and Lee [29] showed CNN that performed with a network that was 12 times smaller than RNNs. It works well to isolate the speech and noise components from the noisy signals. CNN demonstrates its efficacy in speech denoising in both spectral and waveform domains [30].

To demonstrate spectrum mapping, Park and Lee employed a redundant convolutional encoder–decoder. Here, the input is a spectrum that may be thought of as an illustration of two-dimensional representations in accordance with one channel. The repeating of convolutional layers is what this encoder and decoder refer to. Skip connections are used to preserve information while encoding and enhance performance during decoding. Thus, CNN may succeed in its goal of discovering an effective denoising method. Compared to DNN and RNN-LSTM, CNN outperformed them in terms of PESQ and STOI outcomes. By focusing on the timing sequence stage which has the greatest impact, the proposed model can enhance speech separation and partially address the temporal model's limited memory, potentially leading to improved performance.

3. Problem Formulation

In speech enhancement, we have to account for the noise that is added to a clean speech signal and makes it a noisy one. The expression of noisy speech can be characterized as

$$\mathbf{x}(t) = \mathbf{s}(t) + \mathbf{n}(t), \tag{1}$$

where $\mathbf{x}(t)$, $\mathbf{s}(t)$, and $\mathbf{n}(t)$ represent the noisy speech, clean speech, and clean noise, respectively, in discrete time. Then, we apply ***STFT*** [31] of the resulting noisy signal, which can be characterized as

$$\boldsymbol{STFT}\{\mathbf{x}(t)\} = \boldsymbol{STFT}\{\mathbf{s}(t)\} + \boldsymbol{STFT}\{\mathbf{n}(t)\}. \tag{2}$$

The ***STFT*** is used to represent the noisy signal in the complex domain. The real and imaginary parts of the complex domain are then combined to form a new concatenated matrix. This newly concatenated matrix is then sent to the UNET for training and feature extraction. Finally, the clean speech signal $\mathbf{s}(t)$ is achieved by ***ISTFT***. In the following equation,

$$\mathbf{s}(t) = \boldsymbol{ISTFT}\big(\mathbf{E_m}, \mathbf{E_p}\big), \tag{3}$$

$\mathbf{E_m}$ and $\mathbf{E_p}$ denote the enhanced magnitude and enhanced phase, respectively.

4. Proposed Method

This section describes the newly proposed speech enhancement method which is comprised of the training stage, the testing stage shown in Figure 1 and the UNET architecture.

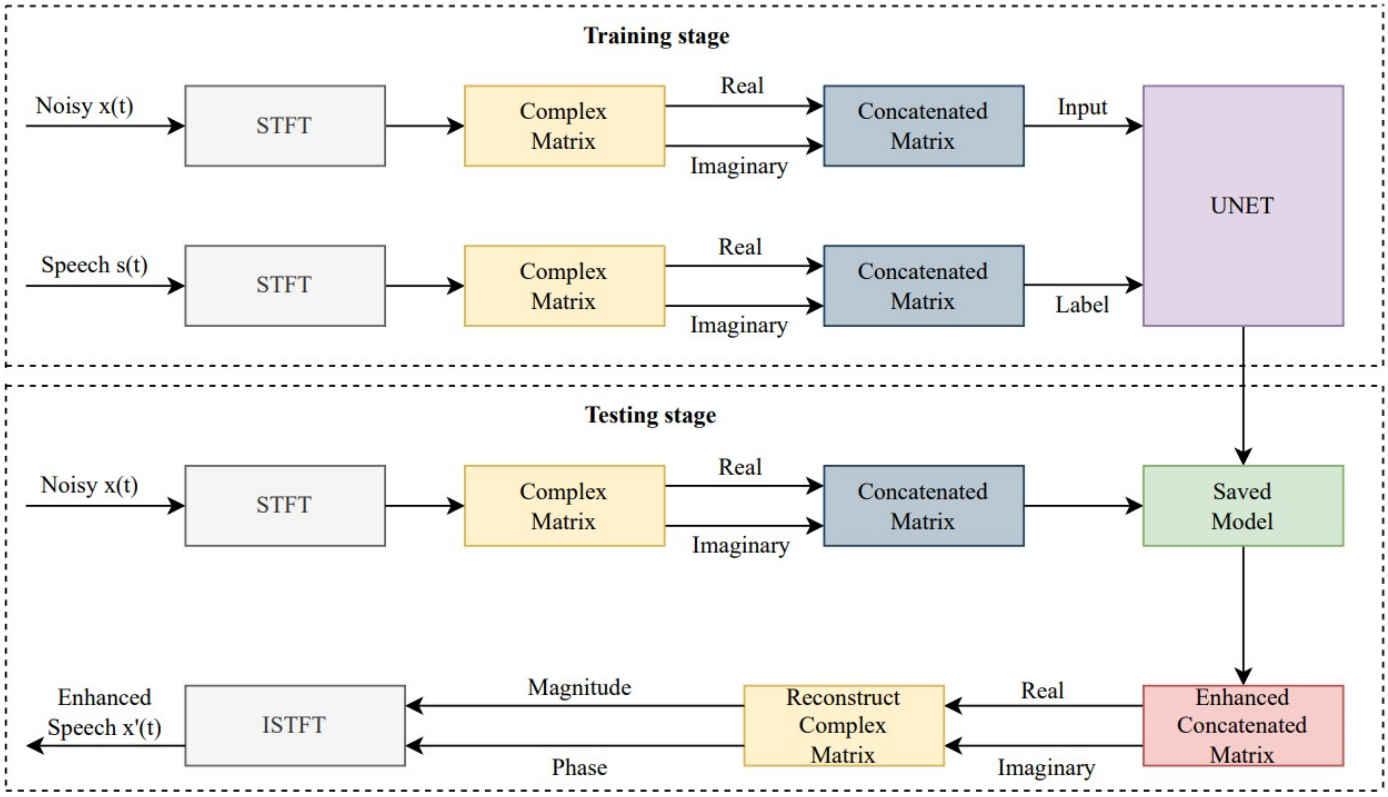


Figure 1. Block diagram of the suggested speech enhancement system.

4.1. Training Stage

The ***STFT*** produces two complex metrics $\mathbf{X}(\mathbf{f})$ and $\mathbf{S}(\mathbf{f})$, for both noisy $\mathbf{x}(\mathrm{t})$ and clean speech $\mathbf{s}(\mathrm{t})$. By concatenating the real and imaginary portions of each matrix, we create two further complex metrics. The UNET model then uses the first concatenated matrix as input and the second one as a label to extract speech components.

$$\boldsymbol{STFT}\{\mathbf{x}(\mathrm{t})\} \rightarrow \mathbf{X}(\mathbf{f}), \tag{4}$$

$$\mathbf{X}(\mathbf{f}) = \mathbf{XR}(\mathbf{f}) + \mathbf{iXI}(\mathbf{f}), \tag{5}$$

$$\boldsymbol{STFT}\{\mathbf{s}(\mathrm{t})\} \rightarrow \mathbf{S}(\mathbf{f}), \tag{6}$$

$$\mathbf{S}(\mathbf{f}) = \mathbf{SR}(\mathbf{f}) + \mathbf{iSI}(\mathbf{f}). \tag{7}$$

The combined form of the real and imaginary parts of $\mathbf{XRI}^{\mathbf{TRAIN}}$ for a noisy combined matrix is sent to the UNET as input and $\mathbf{SRI}^{\mathbf{TRAIN}}$ for speech combined matrix as label data. The system model then decomposes $\mathbf{XRI}^{\mathbf{TRAIN}}$ and $\mathbf{SRI}^{\mathbf{TRAIN}}$ into the bias and weight metrics as follows:

$$\mathbf{XRI}^{\mathbf{TRAIN}} \approx (\mathbf{XRI_W} + \mathbf{b}), \tag{8}$$

$$\mathbf{SRI}^{\mathbf{TRAIN}} \approx (\mathbf{SRI_W} + \mathbf{b}) \tag{9}$$

where $\mathbf{XRI_W} + \mathbf{b}$ express the weight metrics and bias for the noise, $\mathbf{SRI_W} + \mathbf{b}$ specifies the weight metrics and bias for speech signal, g represents the non-linear activation function. Initially, the bias and weight metrics are assigned to zero and random values, respectively. The weight and bias metrics $\mathbf{(XRI_W + b)}$ can be generated by minimizing the cost between $\mathbf{XRI}^{\mathbf{TRAIN}}$ and $g(\mathbf{XRI_W} + \mathbf{b})$ using Equation (10) with the help of Equations (11) and (12)

where α is called the learning rate. During the training, the best model is saved and then bias and weights are updated.

$$\mathbf{XRI_{Error}} = \mathbf{XRI_{label\ output}} - \mathbf{XRI_{predicted\ output}}, \tag{10}$$

$$\mathbf{W_{XRI(New)}} = \mathbf{W_{XRI(Old)}} - \alpha \frac{\partial \mathbf{ZRI_{Error}}}{\partial \mathbf{W_{XRI(Old)}}}, \tag{11}$$

$$\mathbf{b_{XRI(New)}} = \mathbf{b_{XRI(Old)}} - \alpha \frac{\partial \mathbf{ZRI_{Error}}}{\partial \mathbf{b_{XRI(Old)}}}. \tag{12}$$

4.2. Testing Stage

The noisy signal $\mathbf{x}(t)$ is sent to the ***STFT*** algorithm during the testing phase in order to produce a complex spectrogram.

$$\boldsymbol{STFT}\{\mathbf{x}(t)\} \rightarrow \mathbf{X}(\mathbf{f}), \tag{13}$$

$$\{\mathbf{X}(\mathbf{f})\} = \{\mathbf{XR}(\mathbf{f}) + \mathbf{iXI}(\mathbf{f})\} \tag{14}$$

From the complex spectrogram, the imaginary and real parts are concatenated to construct $\mathbf{XRI^{Test}}$ which is passed through the UNET saved model. The model generates an enhanced concatenated spectrogram $\left\{\mathbf{XRI^{Test}}\right\}_{\mathbf{enhanced}}$. The imaginary $\{\mathbf{XI}(\mathbf{f})\}$ and real $\{\mathbf{XR}(\mathbf{f})\}$ parts are then extracted from the enhanced concatenated spectrogram to reconstruct a complex matrix.

$$\{\mathbf{XR}(\mathbf{f}) + \mathbf{iXI}(\mathbf{f})\} = \left\{\mathbf{XRI^{Test}}\right\}_{\mathbf{enhanced}}. \tag{15}$$

The magnitude $\mathbf{E_m}$ and phase $\mathbf{E_p}$ are extracted from the complex spectrogram. Inverse Short-time Fourier transform is applied to the newly generated magnitude and phase to return to the enhanced speech signal as per the following equation:

$$\mathbf{x}'(t) = \boldsymbol{ISTFT}(\mathbf{E_m}, \mathbf{E_p}) \tag{16}$$

4.3. UNET Architecture

We employed the UNET architecture [32] to train the noisy signal which is illustrated in Figure 2. The model architecture includes a pathway that expands on the left and shrinks on the right. The shrinking path follows the typical design of a convolutional network where a leaky rectified linear unit (ReLU) and a 2×2 max pooling operation with a stride of two are used for downsampling after every two 3×3 convolutions applied in succession. As we downsample at each step, we increase the number of feature channels by a factor of four.

The proposed method involves a two-step process. First, we reduce the number of feature channels using a 2×2 convolution, followed by upsampling the feature map. The resulting feature map is then concatenated with a proportionately reduced feature map from the contracting route. Next, we apply two 3×3 convolutions with leaky ReLU activation at each step in the expanding path.

The UNET model employs a cropping step after each convolution, as boundary pixels are lost. The final layer uses a 1×1 convolution to map the 16 feature vector components to the desired number of classes. The network comprises a total of 24 convolutional layers. To ensure a seamless output segmentation map, the input tile size must be selected such that all 2×2 max-pooling operations can be performed on layers with even x and y dimensions.

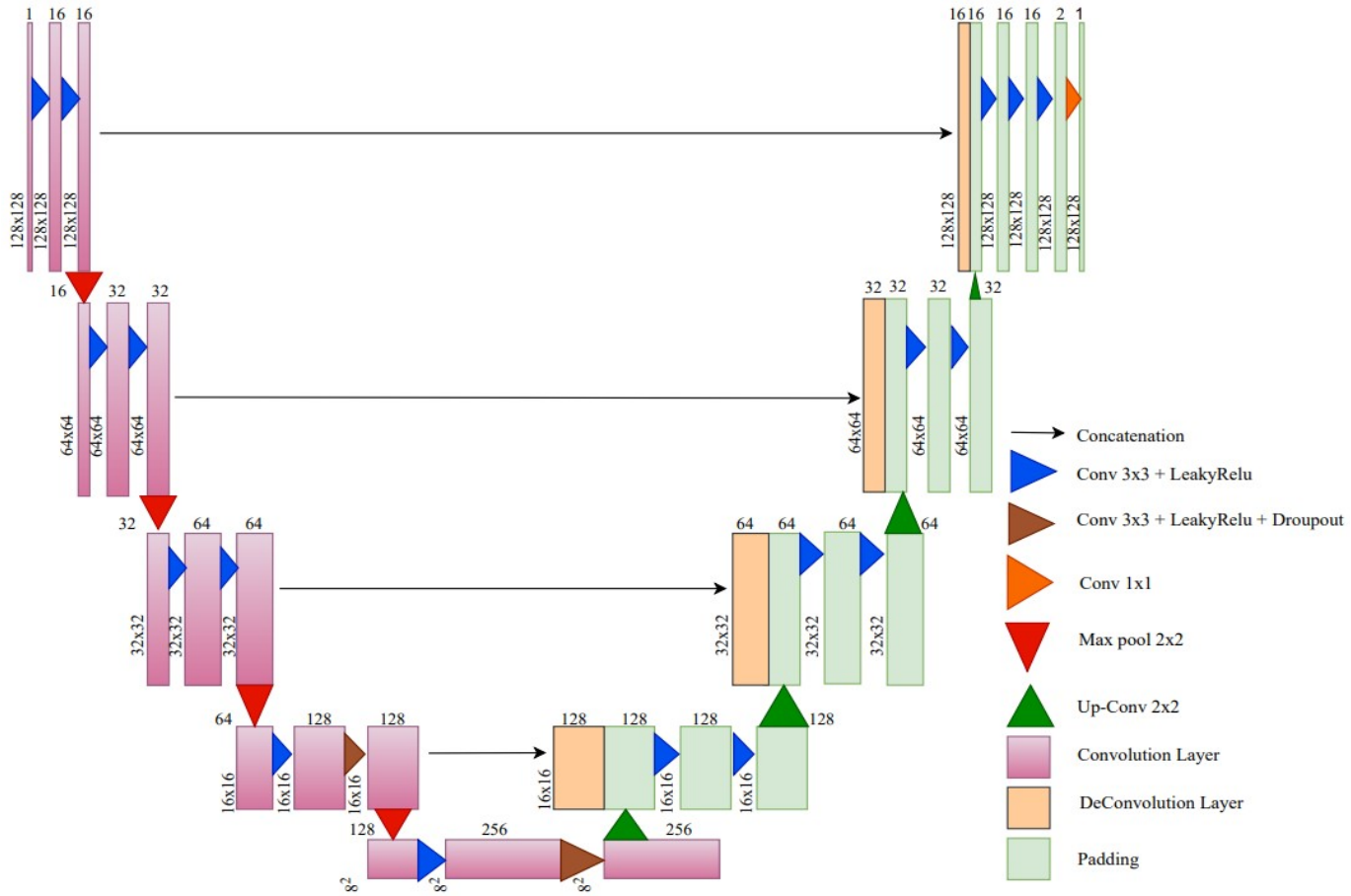


Figure 2. UNET architecture.

The proposed model can be described as a deeply supervised encoder–decoder network that employs nested, dense skip pathways to connect the encoder and decoder sub-networks. These skip pathways were redesigned to effectively bridge the semantic gap between the feature maps of the two sub-networks. By reducing this gap, we believe that the optimizer will have a simpler learning task, as the feature maps of the encoder and decoder networks will exhibit greater semantic similarity.

LadderNet [24], Attention UNET [25], Recurrent and Residual Convolutional UNET (R2-UNet) [26], and UNET with Residual Blocks or Dense Link Blocks are a few of the most well-liked variations.

5. Experimental Results and Discussions

A single male voice speaks 720 utterances in the IEEE corpus [33] at a sample rate of 25 kHz. We chose 200 utterances at random from 1 to 400 to use as our training set, 50 utterances from 400 to 500 as our development set and another 50 utterances from 501 to 720 as our test set. We down-sampled all of the utterances to 8 kHz; the training signal lasts for about 10 min and each test signal lasts for about 6 s. Ten various varieties of noise, including babbles, birds, cafes, cars, casinos, factories, keyboards, machine guns, PC fans, speech-shaped noise (SSN) and streets, may be utilized for training and testing. These sounds originate from many datasets including the NOISEX−92 dataset and the Aurora-2 database. In order to produce noisy test signals, five distinct SNRs [34] are artificially added to clean test signals. These SNRs range from 10 dB to −10 dB with a 5 dB gap. The first 20 min of random noise are used to create the training noise signals while the last 6 s of random noise are used to create the testing noise signals. When computing the applied

magnitude spectrograms for ***STFT***, a hamming window of 8064 samples with 50% overlap is used.

5.1. Experimental Setup

The Adam optimizer was used to train the proposed model across 200 epochs, with a learning rate of 0.0001, β1 of 0.9, and β2 of 0.999. All utterances used for training and testing were captured at 8 kHz. A total of 32 batches were used to train the model. Accordingly, the encoder has 16, 32, 64, 128, and 256 channels per layer, whereas the decoder (shown in Figure 2) has 128, 64, 32, 16 and 1 channel per layer. Speech was preprocessed before being fed into the model which is a similar technique to that of [35]. We set the overlap to 0.5 s (i.e., 50% of the input) and the length of the single-channel speech segments to 1 s (12,000 samples). We also incorporated random shifts between 0 and 1 s for input speech since previous studies showed that data augmentation of the input speech may benefit a model's ability to denoise. Since there are 8064 samples in the random shift, the input speech's length after the shift was around 0.5 s (i.e., 12,000 samples). Before sending the input data to the model, it was standardized using the standard deviation. The standard deviation was subtracted from the output of the model to decrease error. Furthermore, by resampling the input and output of the model during training, interpolated sampling aided in enhancing the model's performance in denoising. The measurements acquired using the random Gaussian matrix had a better likelihood of reproducing the original signal, which aided in enhancing the model's accuracy.

5.2. Evaluation Method

The hearing aid speech quality index (HASQI) [36], the hearing aid speech perception index (HASPI) [37], the perceptual evaluation of speech quality (PESQ) [38] and the short-time objective intelligibility (STOI) [39] were used to assess the speech quality and intelligibility. The HASQI and HASPI were created to evaluate sound quality and perception in people with hearing impairments as well as those with normal hearing [40]. Higher scores corresponded to increased sound quality and intelligibility, respectively. Both of these values ranged from 0 to 1. The PESQ is a frequently used objective quality test for evaluating speech communications. Higher scores indicate greater speech quality, and the results ranged from −0.05 to 4.5. The STOI yields scores that range from 0 to 1 and assesses the correlation coefficient between the temporal envelopes of clean speech and improved speech in short-time areas. Better intelligibility is indicated by a higher STOI rating. In other words, depending on the artifacts and interference present, it determines the overall speech quality of the improved sources. The performance improves as the STOI score rises. It is crucial to utilize this parameter and average it across 20 test signals and 10 different forms of noise when assessing the overall effectiveness of enhancement techniques such as ***STFT-NMF***, ***DTCWT-SNMF***, ***DTCWT-STFT-SNMF***, ***STFT-CDAE*** and ***CMAM***. Using the training set for the training stage and the development set for the test stage, we handled tests 4.1 to 4.2. The total performance of test 4.2 was then tested using the test set in the final step, utilizing the best techniques or backgrounds. This method allowed us to precisely determine the optimal result for our analysis.

5.3. Competing Methods

To evaluate the effectiveness of the proposed model, we trained it with the IEEE corpus dataset and compared it to similar baseline models. The reference methods we used were as follows:

STFT-SNMF [13]; ***STFT*** and ***SNMF*** (KL Cost Function)-based SE Method followed by ***STFT-NMF***, ***DTCWT-SNMF*** [14]; ***DTCWT*** and ***SNMF*** (KL Cost Function)-based SE Method followed by ***DTCWT-NMF***, ***DTCWT-STFT-SNMF*** [15]; ***DTCWT***, ***STFT*** and ***SNMF*** (KL Cost Function)-based SE Method, ***STFT-CDAE*** [16]; Single-Channel Audio Source Separation Using Convolutional Denoising Auto Encoders, ***CMAM*** [17]; Casual Multi-Head Attention

Mechanism, ***STFT-UNET*** [ours]; Supervised Single-Channel Speech Enhancement Method Using UNET.

5.4. Effect of the Proposed Method over Competing Methods

To compare the suggested procedure to other approaches, performance analysis is essential. In this research, we used HASQI, HASPI, PESQ, and STOI metric scores to evaluate the deep learning technique employing the UNET strategy with the ***STFT-SNMF***, ***DTCWT-SNMF***, ***DTCWT-STFT-SNMF***, ***STFT-CDAE*** and ***CMAM*** strategies. The findings in Tables 1 and 2 demonstrate that the suggested method outperforms the alternatives in all SNR scenarios. We determined that the suggested method progressively raises the HASQI score from high to low SNR and that low SNR instances saw an improvement in the HASPI score before high SNR. We know that our suggested method works well in all SNR scenarios, depending on the HASPI score. The PESQ and STOI values are shown in Tables 3 and 4. The proposed deep learning (UNET) model successfully addresses the problem of speech signal distortion following SE processing to some degree since both PESQ and STOI of speech seem to be raised at all SNR levels.

Table 1. Comparison of HASQI values for six different methods under five SNR conditions.

Method	−10	−5	0	5	10
STFT-SNMF [13]	0.145	0.268	0.423	0.583	0.715
DTCWT-SNMF [14]	0.159	0.281	0.437	0.581	0.701
DTCWT-STFT-SNMF [15]	0.174	0.301	0.456	0.607	0.732
STFT-CDAE [16]	0.181	0.361	0.516	0.628	0.745
CMAM [17]	0.183	0.388	0.528	0.654	0.759
STFT-UNET [ours]	**0.215**	**0.399**	**0.566**	**0.698**	**0.795**

Table 2. Comparison of HASPI values for six different methods under five SNR conditions.

Method	−10	−5	0	5	10
STFT-SNMF [13]	0.522	0.605	0.729	0.818	0.901
DTCWT-SNMF [14]	0.543	0.692	0.769	0.854	0.922
DTCWT-STFT-SNMF [15]	0.690	0.742	0.834	0.869	0.949
STFT-CDAE [16]	0.760	0.840	0.867	0.879	0.969
CMAM [17]	0.912	0.932	0.898	0.932	0.981
STFT-UNET [ours]	**0.936**	**0.947**	**0.959**	**0.954**	**0.991**

Table 3. Comparison of PESQ values for six different methods under five SNR conditions.

Method	−10	−5	0	5	10
STFT-SNMF [13]	1.529	1.776	2.148	2.483	2.782
DTCWT-SNMF [14]	1.526	1.918	2.268	2.519	2.748
DTCWT-STFT-SNMF [15]	1.598	2.039	2.414	2.692	2.900
STFT-CDAE [16]	1.675	2.128	2.503	2.601	3.012
CMAM [17]	1.835	2.377	2.702	2.922	3.139
STFT-UNET [ours]	**2.035**	**2.583**	**2.906**	**3.022**	**3.341**

Table 4. Comparison of STOI values for six different methods under five SNR conditions.

Method	−10	−5	0	5	10
STFT-SNMF [13]	0.538	0.649	0.759	0.845	0.906
DTCWT-SNMF [14]	0.555	0.677	0.780	0.849	0.903
DTCWT-STFT-SNMF [15]	0.587	0.706	0.803	0.873	0.920
STFT-CDAE [16]	0.598	0.745	0.832	0.882	0.932
CMAM [17]	0.601	0.775	0.852	0.904	0.944
STFT-UNET [ours]	**0.688**	**0.794**	**0.889**	**0.978**	**0.995**

Depending on the HASPI score, we ensured that the proposed technique performs well in all SNR conditions. Tables 3 and 4 show the PESQ and STOI values, respectively. Since the PESQ and STOI of speech seem to be increased at all SNR levels, our deep learning (UNET) model partially solves the issue of speech signal distortion after SE processing.

The proposed methodology for SNR cases works well based on the HASQI and HASPI ratings in Figures 3 and 4. Our figures show that the technique we propose is effective in increasing HASQI scores from high to low SNR cases and it does so with more success in low SNR cases. This suggests that the proposed methodology is effective despite the SNR conditions.

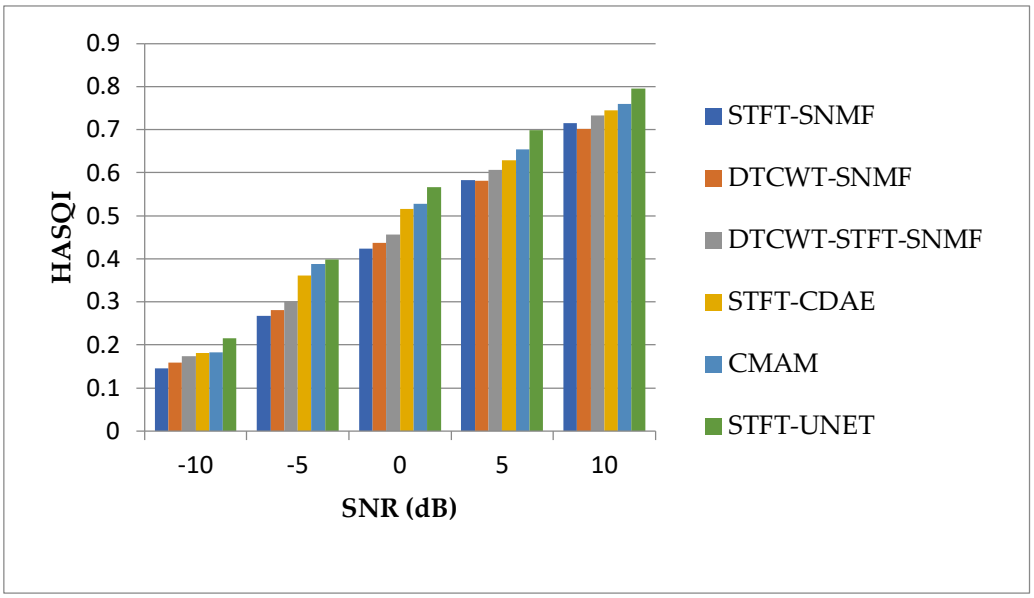


Figure 3. Comparison of the six methods HASQI values at five SNR conditions.

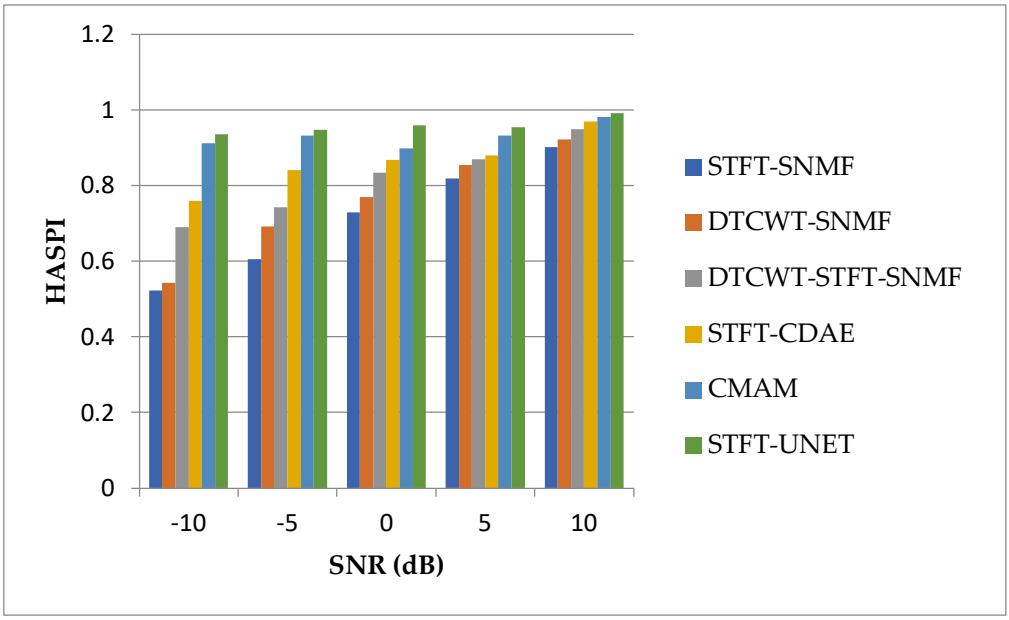


Figure 4. Comparison of the six methods HASPI values at five SNR conditions.

The PESQ findings are outlined in Figure 5, which demonstrates the proposed strategy performance across five SNR conditions. As seen in Figure 5, the proposed strategy outperforms the other five approaches in all SNR situations. In terms of STOI results in

Figure 6, the proposed system's deep learning approach using UNET has preferentially increased the STOI scores over high SNR in low SNR situations.

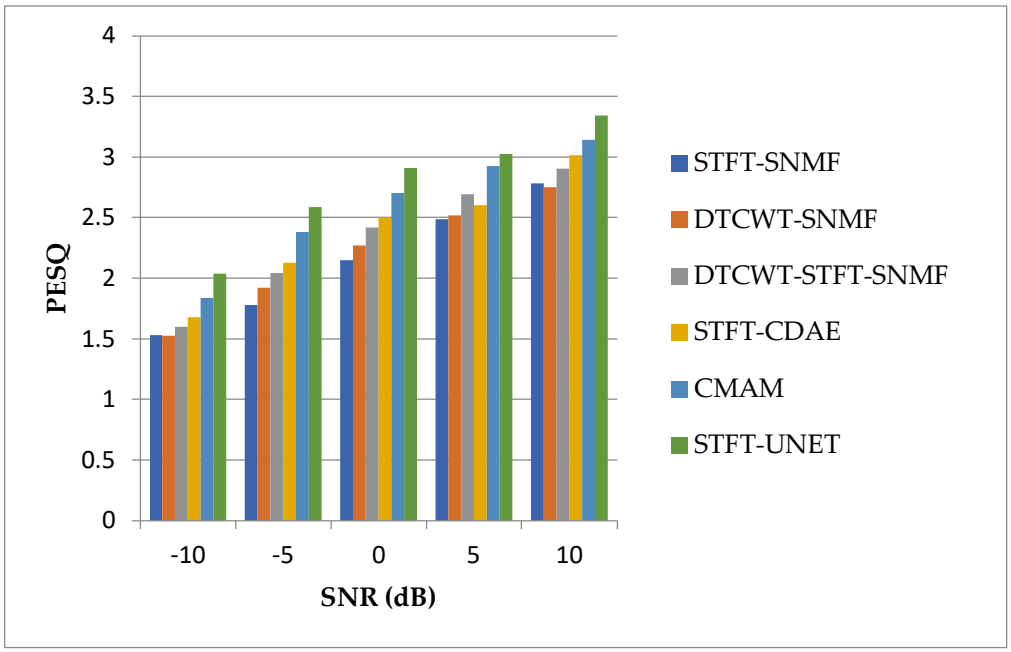


Figure 5. Comparison of the six methods PESQ values at five SNR conditions.

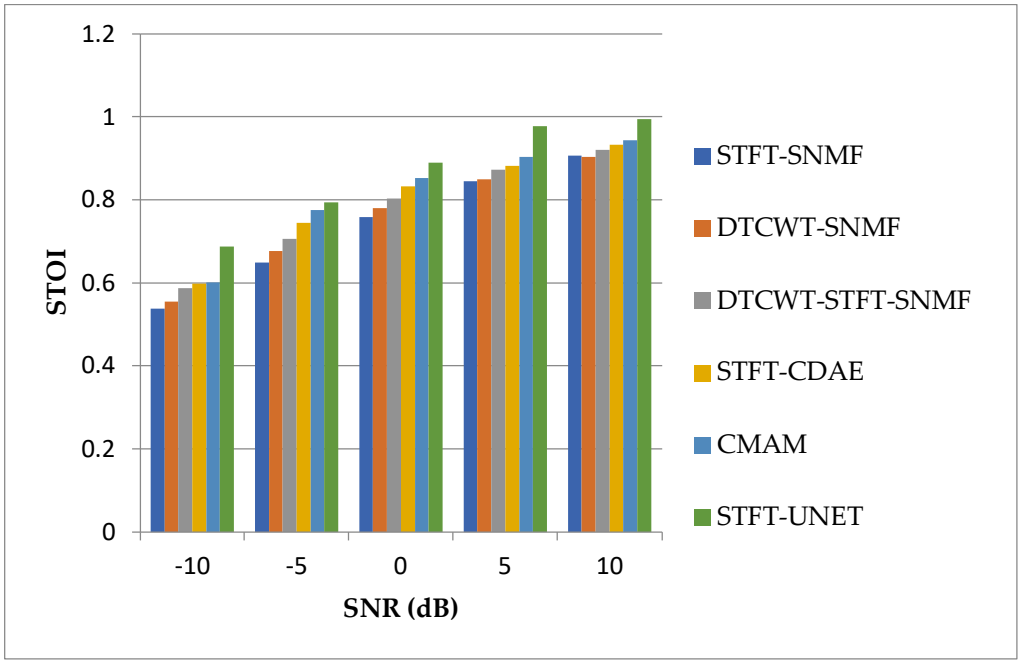
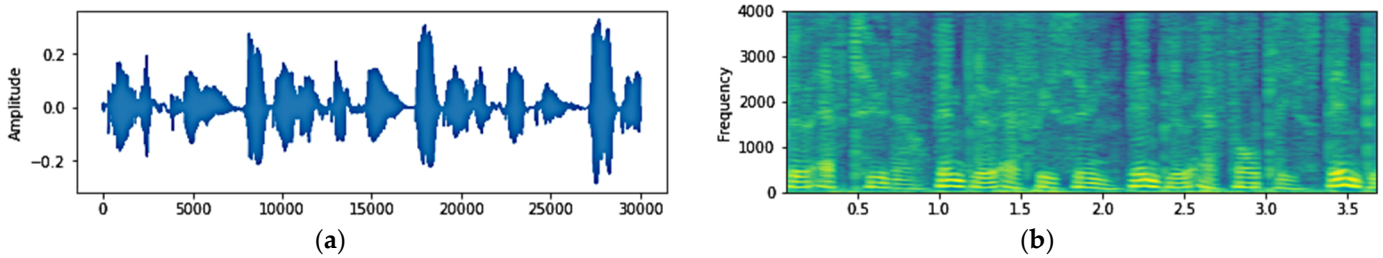


Figure 6. Comparison of the six methods STOI values at five SNR conditions.

5.5. Original and Enhanced Signals Time Domain and Spectrogram Representation

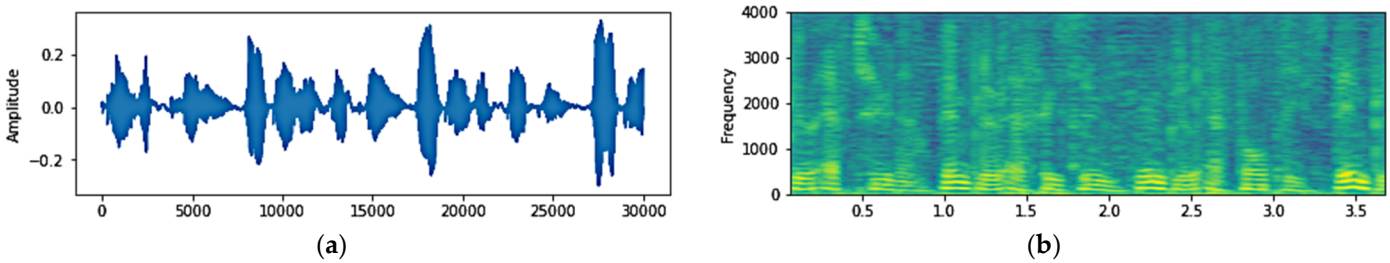
Figure 7 illustrates the time-domain and spectrogram representations of one original speech signal that was randomly selected for 60 s.



(**a**) (**b**)

Figure 7. Original time-domain waveform and spectrogram. (**a**) Original speech waveform, (**b**) original speech spectrogram.

Figure 8 shows that the proposed model approximates the noisy signal to improve it and provide a clear speech signal. Because it accounts for all the many factors that might impact signal quality, this strategy is far more successful than other approaches. As we can see, this technique works quite well at extracting speech from noisy signals.



(**a**) (**b**)

Figure 8. Predicted time-domain waveform and spectrogram. (**a**) Enhanced speech waveform, (**b**) enhanced speech spectrogram.

6. Conclusions

We have developed a convolutional neural network architecture called UNET which improves on the original CNN design. The UNET model architecture is based on two principles. The first is the use of encoder connections which involve max-pooling layers with stride two to reduce data size and the repetition of convolutional layers with more filters in the encoder block. The second principle is the use of decoder blocks and their connections. As we move closer to the decoder, the number of filters in the convolutional layers decreases, followed by continuous up-sampling in the subsequent layers. Skip connections are also used to connect the output of the previous layer to the decoder blocks' layers. By using this network architecture to separate the desired sources, we achieve better performance in all SNR scenarios. The speech signal quality and understandability are enhanced compared to other approaches discussed in this article. The experimental results show that the proposed speech enhancement model outperforms current models in terms of overall performance using various evaluation methodologies. In the future, we plan to explore other training and testing procedures using different deep neural networks.

Author Contributions: Conceptualization, M.N.H. and S.B.; Methodology, M.N.H. and M.S.I.; Software, M.N.H.; Writing—original draft preparation, M.N.H. and M.A.H.; Writing—review and editing, M.S.H., A.A. and M.M.I.; Supervision, M.S.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data will be provided upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Loizou, P. Contrained Wiener Filtering. In *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2013.
2. Ephraim, Y.; Malah, D. Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 443–445. [CrossRef]
3. Cohen, I.; Gannot, S. Spectral Enhancement Methods. In *Springer Handbook of Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 873–902.
4. Hao, X.; Li, X. Fast FullSubNet: Accelerate Full-Band and Sub-Band Fusion Model for Single-Channel Speech Enhancement. *arXiv* **2022**, arXiv:2212.09019. [CrossRef]
5. Ping, H.; Yafeng, W. Single-Channel Speech Enhancement Using Improved Progressive Deep Neural Network and Masking-Based Harmonic Regeneration. *Speech Commun.* **2022**, *145*, 36–46. [CrossRef]
6. Paatero, P.; Tapper, U. Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics* **1994**, *5*, 111–126. [CrossRef]
7. Lee, D.D.; Seung, H.S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
8. Uhlich, S.; Giron, F.; Mitsufuji, Y. Deep Neural Network Based Instrument Extraction from Music. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Queensland, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 2135–2139.
9. Boll, S. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
10. Lim, J.S.; Oppenheim, A.V. Enhancement and Bandwidth Compression of Noisy Speech. *Proc. IEEE* **1979**, *67*, 1586–1604. [CrossRef]
11. Kim, S.; Kim, H. Multi-Microphone Target Signal Enhancement Using Generalized Sidelobe Canceller Controlled by Phase Error Filter. *IEEE Sens. J.* **2016**, *16*, 7566–7567. [CrossRef]
12. Hua, X.; Peng, L.; Liu, W.; Cheng, Y.; Wang, H.; Sun, H.; Wang, Z. LDA-MIG Detectors for Maritime Targets in Nonhomogeneous Sea Clutter. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
13. Mohammadiha, N.; Smaragdis, P.; Leijon, A. Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2140–2151. [CrossRef]
14. Farrokhi, D.D. Single Channel Speech Enhancement in Severe Noise Conditions. Ph.D. Thesis, University of Western Australia, Perth, Australia, 2011.
15. Islam, M.S.; Zhu, Y.; Hossain, M.I.; Ullah, R.; Ye, Z. Supervised Single Channel Dual Domains Speech Enhancement Using Sparse Non-Negative Matrix Factorization. *Digit. Signal Process.* **2020**, *100*, 102697. [CrossRef]
16. Grais, E.M.; Plumbley, M.D. Single Channel Audio Source Separation Using Convolutional Denoising Autoencoders. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1265–1269.
17. Fan, J.; Yang, J.; Zhang, X.; Yao, Y. Real-Time Single-Channel Speech Enhancement Based on Causal Attention Mechanism. *Appl. Acoust.* **2022**, *201*, 109084. [CrossRef]
18. Karjol, P.; Kumar, M.A.; Ghosh, P.K. Speech Enhancement Using Multiple Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5049–5052.
19. Lin, Z.; Zeng, B.; Hu, H.; Huang, Y.; Xu, L.; Yao, Z. SASE: Self-Adaptive Noise Distribution Network for Speech Enhancement with Federated Learning Using Heterogeneous Data. *Knowl.-Based Syst.* **2023**, *226*, 110396. [CrossRef]
20. Shubo, L.; Fu, Y.; Yukai, J.; Xie, L.; Zhu, W.; Rao, W.; Wang, Y. Spatial-DCCRN: DCCRN Equipped with Frame-Level Angle Feature and Hybrid Filtering for Multi-Channel Speech Enhancement. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 436–443.
21. Yechuri, S.; Vanambathina, S. A Nested U-Net with Efficient Channel Attention and D3Net for Speech Enhancement. *Circuits Syst. Signal Process.* **2023**. [CrossRef]
22. Xu, L.; Wei, Z.; Zaidi, S.F.A.; Ren, B.; Yang, J. Speech Enhancement Based on Nonnegative Matrix Factorization in Constant-Q Frequency Domain. *Appl. Acoust.* **2021**, *174*, 107732. [CrossRef]
23. Maas, A.; Le, Q.V.; O'neil, T.M.; Vinyals, O.; Nguyen, P.; Ng, A.Y. Recurrent Neural Networks for Noise Reduction in Robust ASR. 2012. Available online: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45168.pdf (accessed on 22 February 2023).
24. Gao, T.; Du, J.; Dai, L.-R.; Lee, C.-H. Densely Connected Progressive Learning for Lstm-Based Speech Enhancement. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5054–5058.
25. Li, X.; Horaud, R. Multichannel Speech Enhancement Based on Time-Frequency Masking Using Subband Long Short-Term Memory. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 298–302.
26. Weninger, F.; Geiger, J.; Wöllmer, M.; Schuller, B.; Rigoll, G. The Munich Feature Enhancement Approach to the 2nd CHiME Challenge Using BLSTM Recurrent Neural Networks. In Proceedings of the 2nd CHiME Workshop on Machine Listening in Multisource Environments, Vancouver, BC, Canada, 1 June 2013; pp. 86–90.

27. Lee, G.W.; Kim, H.K. Multi-Task Learning U-Net for Single-Channel Speech Enhancement and Mask-Based Voice Activity Detection. *Appl. Sci.* **2020**, *10*, 3230. [CrossRef]
28. Fu, S.-W.; Tsao, Y.; Lu, X.; Kawai, H. Raw Waveform-Based Speech Enhancement by Fully Convolutional Networks. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6–12.
29. Park, S.R.; Lee, J. A Fully Convolutional Neural Network for Speech Enhancement. *arXiv* **2016**, arXiv:1609.07132. [CrossRef]
30. Veselinovic, D.; Graupe, D. A Wavelet Transform Approach to Blind Adaptive Filtering of Speech from Unknown Noises. *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.* **2003**, *50*, 150–154. [CrossRef]
31. Paliwal, K.K.; Alsteris, L.D. On the Usefulness of STFT Phase Spectrum in Human Listening Tests. *Speech Commun.* **2005**, *45*, 153–170. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Lecture Notes in Computer Science 9351; Springer: Berlin/Heidelberg, Germany, 2015.
33. Rothauser, E. IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio Electroacoust.* **1969**, *17*, 225–246.
34. Gerkmann, T. Bayesian Estimation of Clean Speech Spectral Coefficients given a Priori Knowledge of the Phase. *IEEE Trans. Signal Process.* **2014**, *62*, 4199–4208. [CrossRef]
35. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. A Regression Approach to Speech Enhancement Based on Deep Neural Networks. *IEEEACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 7–19. [CrossRef]
36. Kates, J.M.; Arehart, K.H. The Hearing-Aid Speech Quality Index (HASQI). *J. Audio Eng. Soc.* **2010**, *58*, 363–381.
37. Kates, J.M.; Arehart, K.H. The Hearing-Aid Speech Perception Index (HASPI). *Speech Commun.* **2014**, *65*, 75–93. [CrossRef]
38. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 2, pp. 749–752.
39. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
40. Healy, E.W.; Yoho, S.E.; Wang, Y.; Wang, D. An Algorithm to Improve Speech Recognition in Noise for Hearing-Impaired Listeners. *J. Acoust. Soc. Am.* **2013**, *134*, 3029–3038. [CrossRef] [PubMed]