symmetry

MDPI

Article

Symmetric Combined Convolution with Convolutional Long Short-Term Memory for Monaural Speech Enhancement

Yang Xian [1,*], Yujin Fu [2,*], Peixu Xing [2,*], Hongwei Tao [1] and Yang Sun [3]

[1] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[2] College of Mathematics and Information Science, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[3] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK
* Correspondence: xianyang@zzuli.edu.cn (Y.X.); 332310030837@zzuli.edu.cn (Y.F.); peixuxing@163.com (P.X.)

Abstract

Deep neural network-based approaches have obtained remarkable progress in monaural speech enhancement. Nevertheless, current cutting-edge approaches remain vulnerable to complex acoustic scenarios. We propose a Symmetric Combined Convolution Network with ConvLSTM (SCCN) for monaural speech enhancement. Specifically, the Combined Convolution Block utilizes parallel convolution branches, including standard convolution and two different depthwise separable convolutions, to reinforce feature extraction in depthwise and channelwise. Similarly, Combined Deconvolution Blocks are stacked to construct the convolutional decoder. Moreover, we introduce the exponentially increasing dilation between convolutional kernel elements in the encoder and decoder, which expands receptive fields. Meanwhile, the grouped ConvLSTM layers are exploited to extract the interdependency of spatial and temporal information. The experimental results demonstrate that the proposed SCCN method obtains on average 86.00% in STOI and 2.43 in PESQ, which outperforms the state-of-the-art baseline methods, confirming the effectiveness in enhancing speech quality.

Keywords: monaural speech enhancement; Symmetric Combined Convolution; separable convolution; ConvLSTM

Academic Editor: Xiaogang Qi

Received: 8 August 2025
Revised: 8 September 2025
Accepted: 11 September 2025
Published: 20 October 2025

Citation: Xian, Y.; Fu, Y.; Xing, P.; Tao, H.; Sun, Y. Symmetric Combined Convolution with Convolutional Long Short-Term Memory for Monaural Speech Enhancement. *Symmetry* **2025**, *17*, 1768. https://doi.org/10.3390/sym17101768

1. Introduction

Speech enhancement aims at recovering clear and intelligible target speech from noisy speech signals, which is realized by improving the quality of the speech signal and removing background noise and interfering sounds [1]. Therefore, speech information can be accurately transmitted and understood in noisy environments [2]. Monaural speech enhancement deals with the challenging case in which the noisy signal is captured using a single microphone and both the clean speech and interference sources are unknown. Monaural speech enhancement technology has wide applications in various fields, especially in telecommunication, automatic speech recognition (ASR) [3], and hearing aid (HA) [4,5]. Recent studies have shown increasing interest in combining speech enhancement and ASR, where speech enhancement acts as a crucial front-end to improve ASR performance [6].

Conventional speech enhancement methods, including Wiener filtering [7,8], Minimum Mean Square Error (MMSE) estimation [9], and Spectral Subtraction [10], mainly

rely on mathematical distributions of the speech signal. Although they offer comparable performance in stable and predictable conditions, they struggle with complex conditions.

In recent years, using deep neural networks (DNNs) to improve speech quality has become a common approach, shifting away from traditional math-based techniques. Unlike conventional mathematics-based speech enhancement methods, DNNs learn complicated nonlinear relationships between noisy speech and clear speech automatically, by optimizing parameters such as weights and biases through training on large datasets [11]. The DNN subsequently predicts the target speech. DNN-based approaches can be split into two main types: masking-based and mapping-based. The masking-based approaches focus on estimating a masking matrix, which is used to recover the target speech from noisy mixtures in the time and time-frequency domains [12]. For mapping-based methods, the neural networks are trained to directly translate noisy speech into target speech [13].

While DNN-based speech enhancement methods have shown impressive results, they have limitations in utilizing the temporal information, especially in capturing the interdependency of time frames. To overcome this challenge, the recurrent neural network (RNN) [14] was introduced, which can extract the interdependency of adjacent frames in sequential data. However, RNN shows disadvantages in exploiting long-term dependencies. Long short-term memory networks (LSTMs) [15], as an improvement of RNNs, effectively retain long-term dependency information through gating mechanisms.

Recently, convolutional neural networks (CNNs) have become the mainstream method for speech enhancement. With local receptive fields and shared weights, CNNs can efficiently recover clean speech from time-frequency features. Early studies employed CNNs to learn the mapping between the complex spectrograms of noisy and clean speech [16,17]. The combination of CNNs with RNNs has emerged as a promising strategy, leveraging both spatial and temporal modeling capabilities to deliver competitive results in speech enhancement tasks [18]. To capture broader temporal context and enhance the modelling capacity of long-term dependencies, dilated convolutions were introduced, significantly expanding the receptive field without increasing the computational burden [19]. Moreover, the convolutional encoder–decoder architectures show competitive enhancement performance, which can better estimate target speech signals from noisy inputs [20].

In addition to CNN- and RNN-based approaches, generative adversarial networks (GANs) have also been explored for speech enhancement. A representative example is SEGAN [21], where the generator is trained with a combination of adversarial and time-domain reconstruction losses. Moreover, a conformer-based metric generative adversarial network (CMGAN) is proposed to address the multiple speech enhancement-related tasks. Two-stage conformer blocks are used to leverage the magnitude and complex spectra, respectively [22]. The predicted masked magnitude, real and imaginary components of the complex spectrum are jointly incorporated to reconstruct enhanced speech.

Furthermore, boosting the parameter efficiency of CNN is a promising research direction, various structures have been proposed to maintain comparable performance with fewer parameters. Lightweight architectures such as ShuffleNet [23] optimized convolutional operations by incorporating grouped convolutions and channel shuffle mechanisms, improving information flow across channels and reducing computational complexity. Depthwise separable convolutions [24] further improved efficiency by decomposing standard convolutions into depthwise and pointwise operations, which significantly decreased model complexity and computational demands. The convolutional fusion network (CFN) integrated vanilla and separable convolutions to achieve superior performance compared to traditional CNN-based approaches [25].

Despite significant advancements in speech enhancement, several limitations persist in practical applications. The main research gaps can be summarized as follows:

- Existing CNN-based methods often rely on larger kernels or deeper architectures to boost the model capacity, which increases parameters, reduces efficiency, and makes training more difficult.
- The fixed kernels are better at capturing global or local patterns, which are unable to extract global and local information simultaneously, resulting in degradation in model performance.
- CNNs and LSTMs have limitations in long-term temporal and spatial features modeling. CNNs mainly capture local patterns, while LSTMs struggle with high-dimensional information flow, leading to information loss.
- Many state-of-the-art methods require large parameter sizes and heavy computation, making them unsuitable for deployment on resource-constrained devices.

Motivated by the above research gaps, the objectives of this study can be summarized as follows:

- To design a parameter-efficient speech enhancement model that achieves competitive performance while reducing computational costs, ensuring suitability for resource-constrained devices.
- To develop an architecture capable of modeling both local features and long-term temporal/spatial dependencies without significant information loss.
- To comprehensively evaluate the proposed method against state-of-the-art baselines using PESQ and STOI metrics under noisy conditions.

To achieve the above objectives, we propose a novel network architecture named Symmetric Combined Convolution Network with ConvLSTM (SCCN), featuring the following innovations.

Firstly, we propose the Combined Convolution Block (CCB), which integrates three parallel branches, including a standard convolution and two depthwise separable convolutions. One of the depthwise separable convolutions employs a larger kernel size with dilation to significantly expand the receptive field within a single layer, while the other utilizes a smaller kernel with dilation combined with an increased depth multiplier to enhance network width, thereby reinforcing the model capacity. The CCBs are strategically stacked to construct the encoder, effectively balancing receptive field expansion and hierarchical depth enhancement for robust representation learning.

Secondly, we propose a novel decoder module to upsample the encoder output, which is constructed by stacking multiple Combined Deconvolution Blocks (CDBs). The CDB comprises a standard deconvolution and two depthwise separable convolutions. Similarly, the CDB incorporates one separable convolution that employs a large kernel for contextual information extraction, and another with an increased depth multiplier to capture diverse features per input channel, which improves the breadth of network representation. Moreover, the upsampling operations are followed by two separable convolutions to realize the dimension increase.

Thirdly, the SCCN encoder and decoder employ dilated CCBs and CDBs, which use kernels with exponentially increasing dilation rates. As a result, the receptive field is further extended, enabling the model to capture long-range temporal dependencies in sequential data.

Fourthly, we introduce grouped ConvLSTM to process the high-dimensional information flow, which can effectively utilize spatial information. The input of ConvLSTM is divided and rearranged into two distinct groups, which decreases model complex-

ity and at the same time, effectively retains the ability to model both temporal and spatial dependencies.

The rest of the paper is organized as follows. Section 2 outlines the problem statement. Section 3 details the proposed approach. Experimental setup and results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Problem Statement

The monaural speech enhancement problem aims to separate clean speech s from the noisy speech mixture y. The noisy speech mixture can be modeled as the sum of clean speech and noise n, and this relationship can be expressed as follows:

$$y(m) = s(m) + n(m) \tag{1}$$

where m is the time index. By applying the Short-Time Fourier Transform (STFT), the time domain noisy speech mixture is transformed into the time-frequency domain as

$$Y_{t,f} = S_{t,f} + N_{t,f} \tag{2}$$

where $S_{t,f}$ and $N_{t,f}$ denote the time-frequency domain clean speech signal and noise, respectively. And t and f stand for the time frame and the frequency bin.

Methodology

The neural network model is trained to approximate a mapping function G_θ, parameterized by θ, which transforms the magnitude spectrum of the noisy speech mixture $|Y_{t,f}|$ into that of the clean speech signal $|S_{t,f}|$. The model is evaluated using the mean absolute error (MAE) loss function, calculating the mean of absolute differences across all time-frequency bins between the magnitude of the estimated target speech $|\hat{S}_{t,f}|$ and that of the clean speech $|S_{t,f}|$,

$$\begin{aligned} Loss &= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\left|G_\theta(|Y_{t,f}|) - |S_{t,f}|\right| \\ &= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\left||\hat{S}_{t,f}| - |S_{t,f}|\right| \end{aligned} \tag{3}$$

The flowchart of speech enhancement using the proposed SCCN model is illustrated in Figure 1. In the training stage, training speech and training noise are mixed to generate the training noisy mixture. Spectral features are extracted from both the training speech and the training noisy mixture. These features of the training noisy mixture are used as input to train the SCCN model. The process enables the model to learn how to recover clean speech from noisy mixtures by analyzing the relationship between the noisy mixture and clean speech, ultimately resulting in a trained SCCN model. In the testing stage, testing speech is mixed with testing noise to form the testing noisy mixture. Spectral features are extracted from testing noisy mixture. These features are fed into the trained SCCN model for inference, which is applied to generate the enhanced magnitude spectrum of speech. The phase information is extracted from the spectrum of the testing noisy mixture. The enhanced magnitude of speech is incorporated with the phase of the noisy mixture to generate the enhanced speech signal. Finally, the generated enhanced speech is compared with the original testing speech and noisy speech mixture for performance evaluation, providing a quantitative measure of the model's effectiveness.

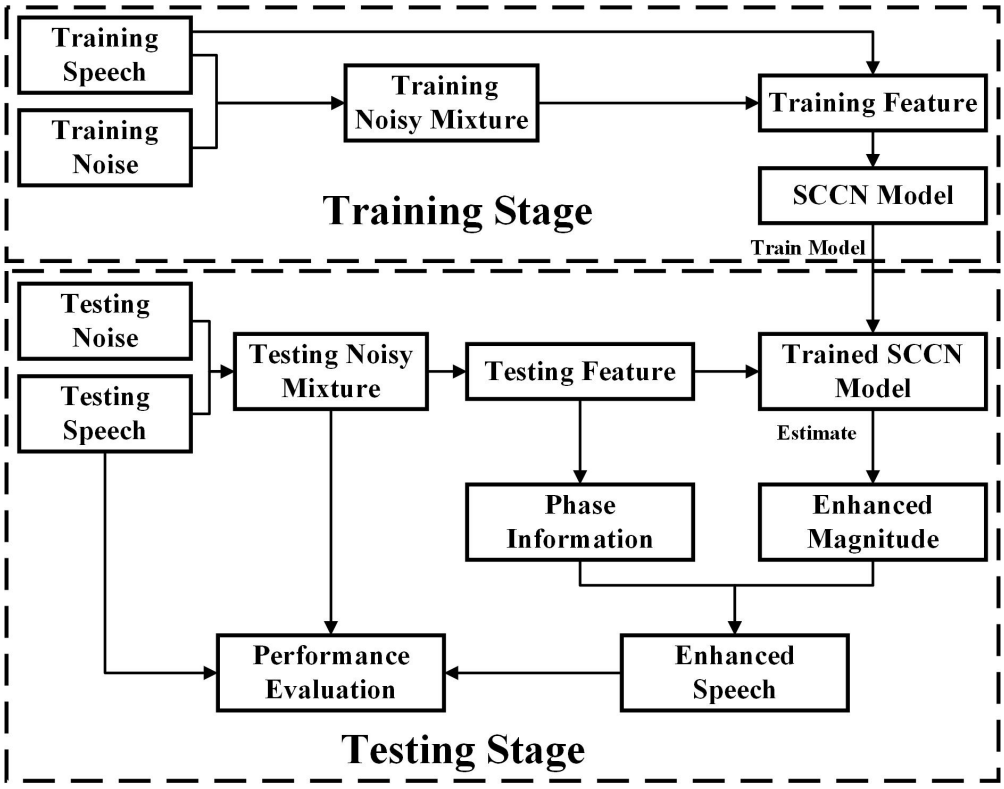


Figure 1. Block diagram of speech enhancement process based on the proposed SCCN model, showing the training and testing stages.

3. Proposed Method

3.1. Proposed Network Architecture

The proposed SCCN is a convolutional encoder–decoder architecture with multiple skip connections, designed to realize monaural speech enhancement, as illustrated in Figure 2. It takes the magnitude spectrum of the noisy mixture as input and generates the estimated magnitude spectrum of the target speech. Afterwards, the phase information of the noisy signal is combined with this estimate for final waveform reconstruction. The architecture consists of three main components: a convolutional encoder, grouped ConvLSTM layers, and a convolutional decoder. The encoder includes a convolutional layer and 12 stacked CCBs, with the number of output channels increasing progressively from 16 to 128. The encoded high-level features are then processed by two grouped ConvLSTM layers to model temporal dependencies. The decoder mirrors the encoder structure, consisting of 12 CDBs and a decovolutional layer, with output channels decreasing from 128 to 16. Varied types of skip connections are employed to promote feature reuse and enhance representation learning. Furthermore, the dilated CCBs and CDBs are applied in SCCN encoder and decoder, respectively.

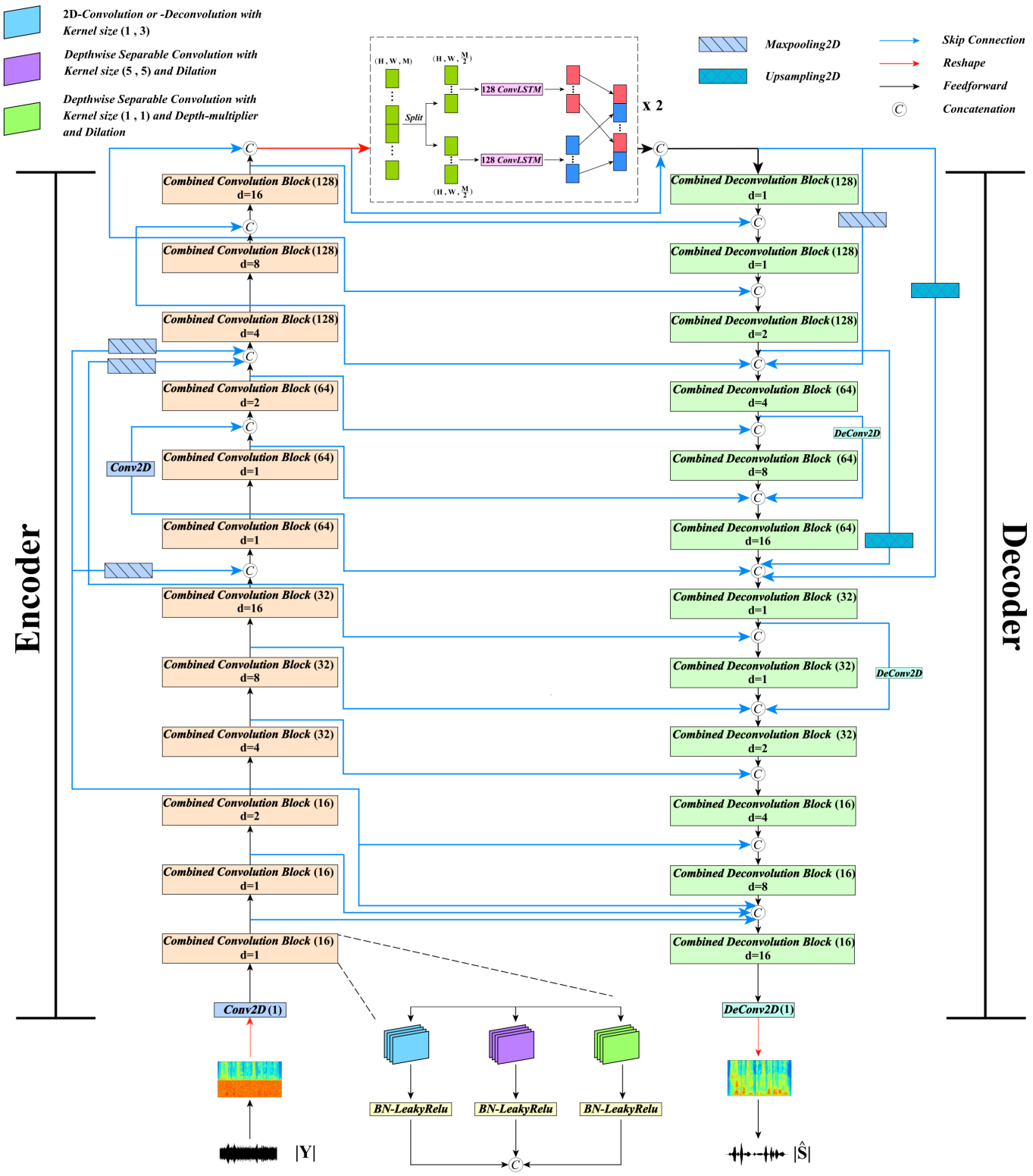


Figure 2. The architecture diagram of the proposed SCCN, with the components and their functions listed at the top of the figure. 'Combined Convolution Block (128) d = 8' refers to a block with three convolutional branches, each with 128 output channels. The dilation rates of two depthwise separable convolution branches are 8. '128 ConvLSTM' denotes ConvLSTM with 128 output channels. The encoder is positioned on the left, while the decoder is on the right side of the figure. The structure diagram of the CCB(16) is shown at the lower center of the figure, while the detailed structure diagrams of CCB can be found in Figure 3.

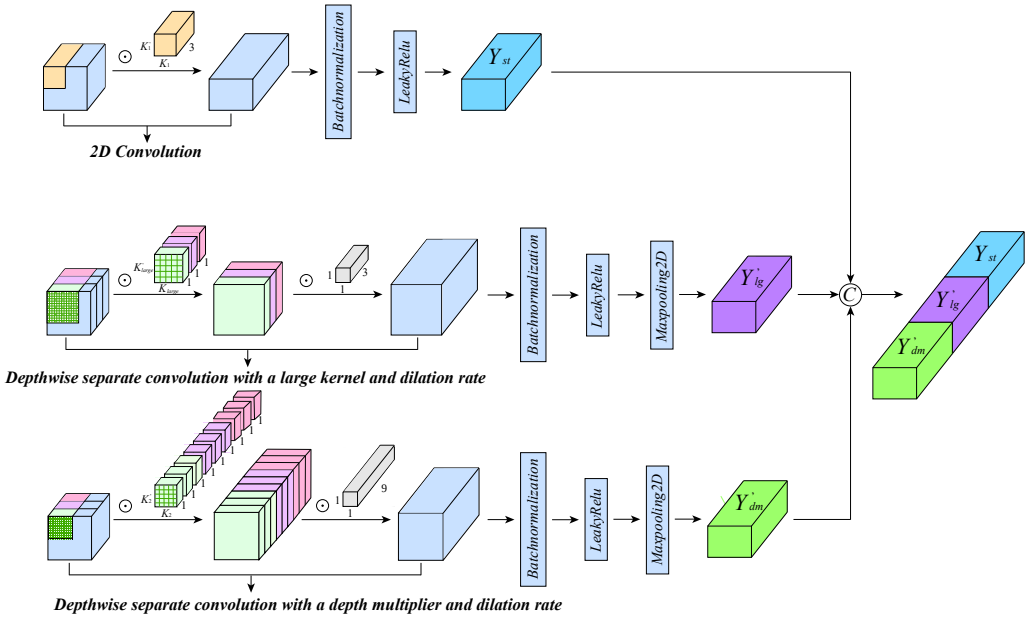


Figure 3. Architecture of the proposed CCB. The module consists of three main operations: 2D convolution, DCLK, and DCLM. Symbols ⊙ and © represent the convolution operation and concatenation, respectively. For Simplify, the figure only displays that each convolutional branch takes three-channel input and generates one-channel output.

3.2. Combined Convolution Block

The CCB consists of three main operations: 2D convolution, depthwise separable convolution with a large kernel (DCLK), and depthwise separable convolution with a larger depth multiplier (DCLM). The output of these operations is aggregated along the channel dimension to form a unified representation of the characteristics. Figure 3 illustrates the module architecture.

The input feature map X is processed by the 2D convolution, producing an output feature map Y. The 2D convolution operation is mathematically represented as:

$$Y_{\text{st}}[i,j,c] = \sum_{m=1}^{K_1} \sum_{n=1}^{K_1'} \sum_{k=1}^{C_{in}} X[i+m-1, j+n-1, k] \cdot W_{\text{st}}[m,n,k,c] \tag{4}$$

where W_{st} denotes the convolution kernel, with K_1 and K_1' representing the spatial dimensions (height and width) of the kernel, respectively. $Y_{\text{st}}[i,j,c]$ is an element of the output feature map Y, indicating the convolution result at position (i,j) and output channel c. The index c ranges from 1 to C_{out}, where C_{out} is the number of output channels. The element $X[i+m-1, j+n-1, k]$ corresponds to the value at position $(i+m-1, j+n-1)$ in the spatial domain and input channel k of the input feature map X, where k ranges from 1 to C_{in}, with C_{in} representing the number of input channels.

For the DCLK, the depthwise convolution uses a large kernel for each input channel. This large kernel enables the model to capture a broader context. Dilation rate d increases the spacing between kernel elements, allowing the model to capture larger receptive fields without extra computational cost. This depthwise convolution is followed by a pointwise convolution (1×1 kernel). The operations are mathematically represented as below.

1. **Depthwise Convolution with Large Kernels**:

$$Y_{\text{lg}}[i,j,c] = \sum_{m=1}^{K_{\text{larger}}} \sum_{n=1}^{K'_{\text{larger}}} X[i+(m-1)\cdot d, j+(n-1)\cdot d, c] \cdot W_{\text{lg}}[m,n,c] \tag{5}$$

where W_{lg} is the convolution kernel of size $K_{\text{larger}} \times K'_{\text{larger}}$, and Y_{lg} is the output of the depthwise convolution applied to the input feature map X using this large kernel along with a dilation rate d. and c represents the index of the output channel of the depthwise convolution ,with $c = 1, 2, \ldots, C_{\text{in}}$.

2. **Pointwise Convolution with 1 × 1 Kernel**:

$$Y'_{\text{lg}}[i, j, c'] = \sum_{c=1}^{C_{\text{in}}} Y_{\text{lg}}[i, j, c] \cdot W'_{\text{lg}}[c, c'] \tag{6}$$

where W'_{lg} is the 1×1 convolution kernel, and Y'_{lg} is the output of the pointwise convolution after the depthwise convolution. The kernel size is 1×1, meaning that it combines information from different channels while maintaining the spatial resolution. c denotes the input channel index and the output channel index associated with the pointwise convolution, where $c' = 1, 2, \ldots, C_{\text{out}}$.

For the DCLM, each input channel is convolved with multiple kernels, generating multiple output channels. The depth multiplier parameter determines how many filters are applied to each input channel, thus increasing the number of output channels. Such a configuration facilitates the extraction of more complex representations, including cross-channel information. By applying multiple filters to each input channel, the DCLM improves the model's capacity for extracting richer, more abstract features. In our model, we further incorporate a dilated encoder and decoder, where dilated convolutions are applied to DCLK and DCLM. The operations are mathematically represented as follows:

1. **Depthwise Convolution with Depth Multiplier**:

$$Y_{\text{dm}}[i, j, c''] = \sum_{m=1}^{K_2} \sum_{n=1}^{K'_2} X[i + (m-1) \cdot d, j + (n-1) \cdot d, c''] \cdot W_{\text{dm}}[m, n, c''] \tag{7}$$

where Y_{dm} is the output of depthwise convoluiton for DCLM, and c'' represents the index of the output channel of the depthwise convolution, with $c'' = 1, 2, \ldots, C_{\text{in}} \cdot \text{depth\_multiplier}$.

2. **Pointwise Convolution with 1 × 1 Kernel**:

$$Y'_{\text{dm}}[i, j, c'''] = \sum_{c''=1}^{C_{\text{in}} \cdot \text{depth\_multiplier}} Y_{\text{dm}}[i, j, c''] \cdot W'_{\text{dm}}[c'', c'''] \tag{8}$$

In the Equation (8), c''' denotes the output channel index of the pointwise convolution, and $c''' = 1, 2, \ldots, C_{\text{out}}$.

Batch Normalization and LeakyRelu follow three convolutions to improve nonlinearity. Moreover, Maxpooling is employed to downsample the feature maps of two separable convolutions when needed; Otherwise, it is omitted. And finally, the processed outputs of the three operations: standard 2D convolution Y_{st}, DCLK Y'_{lg}, and DCLM Y'_{dm} are aggregated across channels to construct a unified feature map Z. This operation is mathematically represented as:

$$Z = \text{concat}(Y_{\text{st}}, Y'_{\text{lg}}, Y'_{\text{dm}}) \tag{9}$$

3.3. Combined Deconvolution Block

For the convolutional encoder–decoder architecture, the decoder takes the low-dimensional information flow produced by the encoder as input and reconstructs it into a higher-dimensional form that aligns with the original input sequence. To recover information flow, the CDBs are stacked to build the convolutional decoder. The CDB employs

three key operations: standard 2D deconvolution, DCLK, and DCLM. These operations aim to expand feature dimensionality while effectively restoring spatial information. The structure is similar to the CCB asshown in Figure 3.

The standard 2D deconvolution utilizes strides to upsample the information flow. Batch Normalization and LeakyReLU are followed to improve training stability and nonlinearity. These operations are also used to process the outputs of two separable convolutions. Moreover, the upsampling is employed to expand the spatial dimensions of two separable convolutions when needed, which enhances the model's ability to capture fine details. Finally, the processed outputs of all three convolutions are concatenated to provide an informative representation of speech signals.

$$Y_{\text{agg}} = \text{Concat}(Y_{d1}, Y_{d2}, Y_{d3}) \tag{10}$$

where Y_{d1}, Y_{d2}, and Y_{d3} denote the processed outputs of Deconv, DCLK, and DCLM, respectively.

3.4. Dilated Encoder and Decoder

In the SCCN encoder and decoder, dilated convolutions are employed within two depthwise separable convolutions to effectively model long-range dependencies in speech signals. Unlike traditional convolutions, whose receptive field is limited by a fixed kernel size, dilated convolutions progressively expand the receptive field by increasing the dilation rate, allowing the network to capture information from a broader temporal or spatial range without significantly increasing the computational cost.

Specifically, the SCCN model uses an exponentially increasing dilation rate, starting at 1 for input and hidden layer 1, then doubling with each successive layer, and the detailed configuration is illustrated in Figure 4. As the dilation rate increases, the receptive field grows exponentially, allowing the network to model both short-term and long-term dependencies and integrate information from larger regions of the input space.

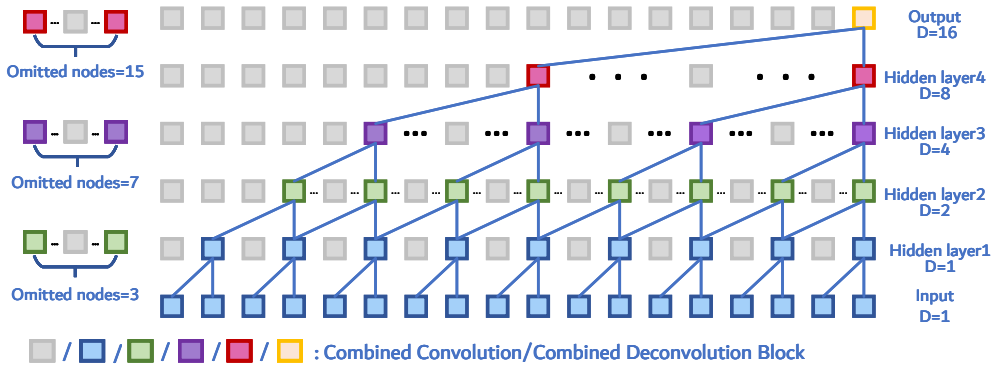


Figure 4. The dilated combined convolution and deconvolution blocks utilized in ecoder/decoder. The different colors denote different dilation rates, the receptive field of the grey block is omitted to facilitate understanding.

3.5. Grouped Convolutional Long Short-Term Memory

We propose a grouped ConvLSTM to capture the spatial information and long-term dependencies simultaneously, which overcomes the limitations of vanilla LSTM structure. In addition, the group structure enhances parameter efficiency. The detailed information about the grouped ConvLSTM is provided in Figure 5a.

The grouped ConvLSTM comprises two layers, and each layer contains two parallel ConvLSTMs. The input of the first layer is divided into two streams. The first stream is fed to the initial ConvLSTM, while the second stream is processed by the subsequent ConvLSTM. Their outputs are concatenated and rearranged, which is passed to the second

grouped ConvLSTM layer. Therefore, grouped ConvLSTM demonstrates higher parameter efficiency by dividing the output channels into distinct groups.

For example, under equivalent architectural settings, a standard ConvLSTM layer typically has more parameters than a grouped ConvLSTM layer that separates the processes of the part of input. Then, a parameter-free representation rearrangement layer facilitates cross-group information exchange by rearranging output of grouped ConvLSTM. Feature rearranging boosts the model's ability to capture inter-group correlations, as illustrated in Figure 5c. Without this rearrangement layer, as depicted in Figure 5b, the model struggles to effectively capture inter-group dependencies, which limits its ability to learn across different feature groups. The introduction of the representation rearrangement layer addresses this limitation by promoting effective information flow across groups while preserving the model's capacity to capture both long-term temporal and spatial dependencies. This approach strikes a balance between reducing complexity and enhancing effective learning.

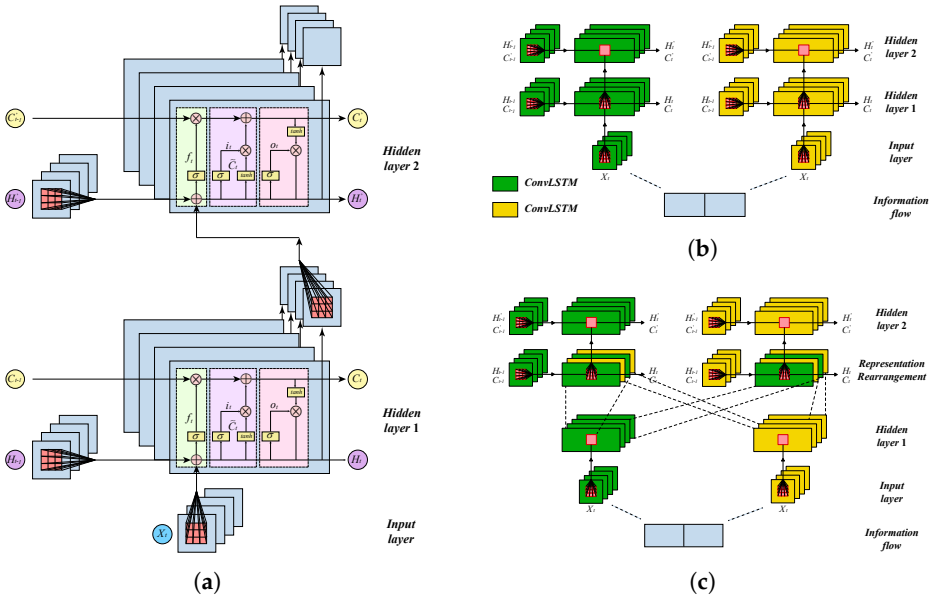


Figure 5. The architecture diagram of the proposed Grouped ConvLSTM. (**a**) A standard ConvLSTM. (**b**) A Grouped ConvLSTM without representation rearrangement. (**c**) A Grouped ConvLSTM with representation rearrangement.

More specifically, ConvLSTM modifies the traditional LSTM architecture by replacing fully connected operations with convolutional ones within its recurrent units, enabling the effective capture of both temporal dynamics and spatial features [26]. A key characteristic of ConvLSTM is its gated structure, implemented through convolutional filters, which reduces the number of parameters through weight sharing while effectively capturing long-term dependencies. This design enhances both computational efficiency and model generalization [27]. The primary computational formulas are as follows:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \tag{11}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \tag{12}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * H_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \tag{13}$$

$$H_t = o_t \circ \tanh(C_t) \tag{14}$$

Notation $*$ indicates the convolution operation, while symbol $\circ$ signifies the Hadamard product. i_t, f_t, and o_t represent the input gate, forget gate, and output gate, respectively.

W represents the different convolution weight matrices, corresponding to the weights from the current input X_t, the previous hidden state H_{t-1}, and the previous cell state C_{t-1} to each gate. b is the bias term for each gate. C_t and H_t represent the cell state and hidden state in the current time step, respectively.

4. Exeperiments

4.1. Data and Setup

We evaluate our system with two experiments using the commonly used database. In the first experiment, the proposed approach is evaluated based on noise signals of the Diverse Environment Multichannel Acoustic Noise Database (DEMAND) [28]. More specifically, we randomly select clean utterances from TIMIT [29] and VCTK [30] corpora and mix them with noise signals from DEMAND to thoroughly assess the model's effectiveness under low SNR scenarios, where the noise signals have higher energy than clean speech signals. The training set includes noisy speech mixtures generated at four SNR levels (−15 dB, −10 dB, −5 dB, 0 dB). In total, there are 108,000 (27,000 × 4 = 108,000) noisy mixtures that are exploited to train the network model. To fully evaluate the proposed model, the speakers in the training sets differ from those in the testing sets. Similarly, the testing noisy mixtures are generated by mixing the clean speech and noise signals at −15 dB, −10 dB, −5 dB, and 0 dB. The testing noise contains OMMEETING, TMETRO, STRAFFIC, DLIVING, Nriver. In summary, about 1.67 h (500 × 4 × 3 ÷ 3600 = 1.67) of noisy speech mixtures are utilized to test the proposed methods. In the second experiment, the proposed method is tested by using the noise signals from the NOISX-92 [31] and Nonspeech [32] databases. A random selection of clean utterances is made from the TIMIT corpus and mixed with noise signals from the Nonspeech and NOISEX-92 databases to generate the training set. This training set comprises 66,000 noisy mixtures that are created using 21 noise signals. A total of 1200 noisy mixtures are exploited to test the baseline and the proposed models. The testing noises include Machine, Traffic and Car, Crowd from Nonspeech database, and Machinegun from NOISX-92. The Machine noise is seen in the training set, the Traffic and Car noise and Machinegun noise are semi-unseen, which means the training set utilizes the first half of the noise signals and the testing employs the second half noise. Moreover, the Crowd noise is completely unseen, meaning it is not used in the training set.

To provide a clearer explanation of Experiment 1 and Experiment 2, we present a detailed summary of both experiments in tabular form, in Table 1.

Table 1. Experimental Summary.

Experiment	Setup	Training / Testing Details
Experiment 1	Clean speech corpus: TIMIT, VCTK Noise corpus: DEMAND SNR levels: −15, −10, −5, 0	Training: 108,000 mixtures Testing: 2000 mixtures Noise types: OMMEETING, TMETRO, STRAFFIC, DLIVING, NRIVER
Experiment 2	Clean speech corpus: TIMIT Noise corpus: NOISEX-92, Nonspeech DataBase SNR levels: −5, 0, 5	Training: 66,000 mixtures Testing: 1200 mixtures Noise types: Machine (seen), Traffic & Car, Machinegun (semi-unseen), Crowd (unseen)

The perceptual evaluation of speech quality (PESQ) [33] and short-time objective intelligibility (STOI) [34] serve as metrics for assessing performance. PESQ scores range from −0.5 to 4.5, representing the quality of speech perception, while STOI scores vary

between 0 and 1, reflecting speech intelligibility. Higher values in both metrics correspond to improved enhancement results.

4.2. Baselines

SCCN (7.2 million) is evaluated against five benchmark approaches, such as the LSTM model (30.7 million), BLSTM model (15.9 million), CFN method (3.5 million) from [25], GRN model (2.5 million) used in [19], AECNN model (6.4 million) applied in [20]. The LSTM model is a five-layer network consisting of four LSTM hidden layers followed by a fully connected output layer. Each LSTM layer has 1024 units and exploits dropout for regularization. Similarly, the BLSTM model is a five-layer network with four BLSTM hidden layers and an output layer. Each BLSTM layer contains 512 units with dropout, and it comprises a forward LSTM sublayer and a backward LSTM sublayer. The outputs of two sublayers are summed and passed to the next layer. The CFN method utilizes the standard convolution and depthwise-separable convolution to construct the convolutional and deconvolutional fusion blocks, which are stacked to form the CFN network. The GRN is a fully connected dilated convolutional network comprising 62 layers with residual connections. The gated convolutions are applied to further enhance information flow control within the GRN network. The AECNN is an 18-layer convolutional encoder–decoder architecture that incorporates skip connections between the encoder and decoder to facilitate feature utilization and fusion.

4.3. Results and Analysis of the First Experiment

Figures 6 and 7 illustrate the performance comparisons among the proposed SCCN method, LSTM, BLSTM, and CFN based on STOI and PESQ for speaker-independent cases involving Tmetro, Straffic, Dliving, Nriver, and Omeeting noises. Overall, all models yield significant improvements over the unprocessed noisy mixtures based on STOI and PESQ scores, which indicate all models are capable of reducing noise to varying degrees.

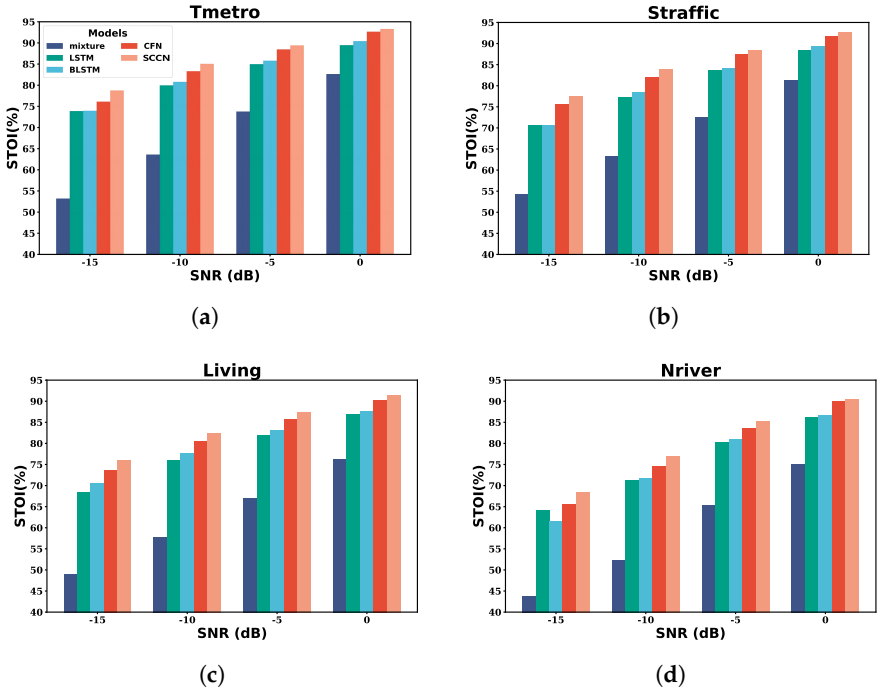
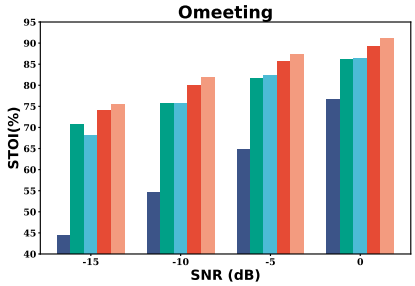
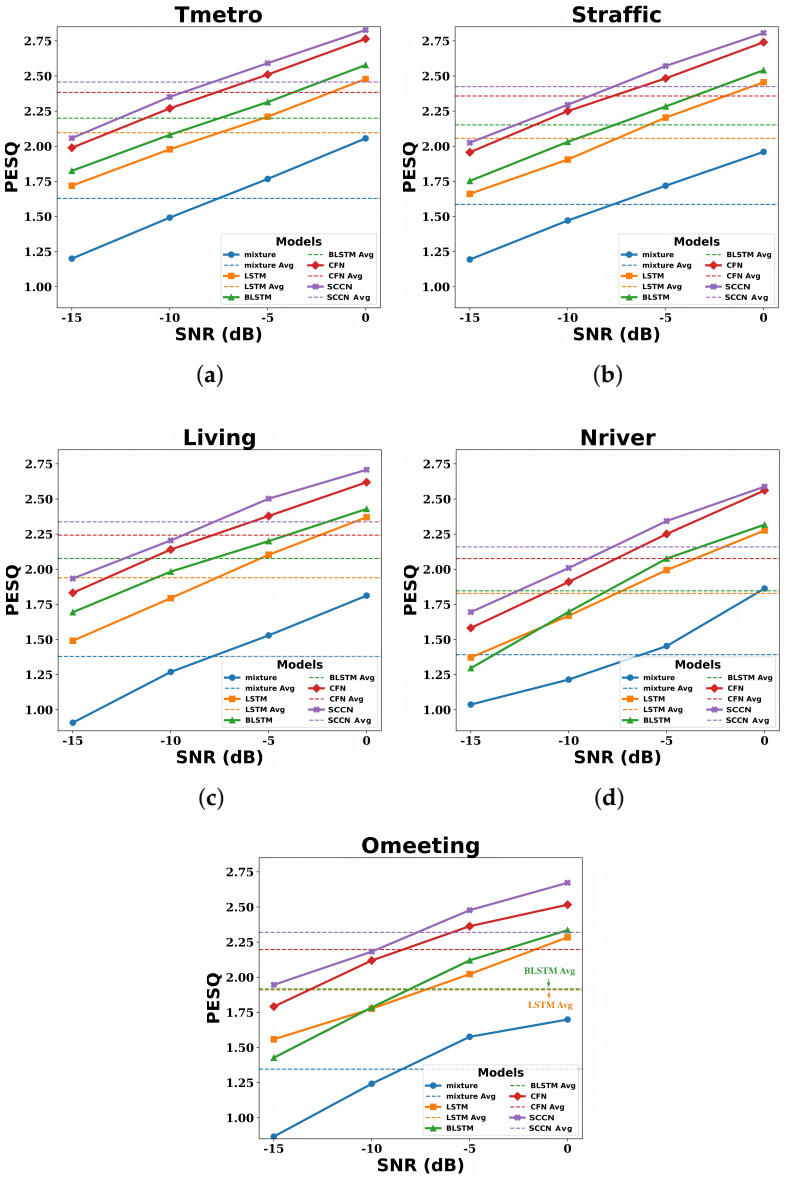


Figure 6. *Cont.*



(**e**)

Figure 6. STOI scores for different speech enhancement models on the Tmetro, Straffic, Living, Nriver, and Omeeting noises, with subplots (**a**–**e**) corresponding to each noise.



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

(**c**) (**d**)

(**e**)

Figure 7. PESQ scores for different speech enhancement models on the Tmetro, Straffic, Living, Nriver, and Omeeting noises, with subplots (**a**–**e**) corresponding to each noise.

More specifically, the proposed SCCN model provides on average 84.14% of STOI and 2.34 of PESQ score, demonstrating the best enhancement performance among all baseline models. The SCCN is composed of multiple layers of CCBs and CDBs, which incorporate a

parallel convolutional structure with varied configurations, including kernel size, depth multiplier, and dilation rates, to capture multiscale features. Therefore, the SCCN utilizes these multi-scale features to facilitate the interpretation and processing of noisy mixtures. Moreover, the inclusion of grouped ConvLSTM layers enables the SCCN to realize the effective temporal and spatial interdependency extraction within the information flow.

The CFN offers an average STOI of 82.50% and a PESQ score of 2.25, demonstrating the second best performance among all baseline models. The CFN model leverages the advantages of the standard convolution and depthwise separable convolution to strengthen the model's capacity. Furthermore, the channel shuffle mechanism is employed to extract inter-channel interdependency, while the inter skip connections within the encoder and decoder facilitate feature reutilization.

The LSTM achieves an average STOI of 78.89% and a PESQ score of 1.97, representing the lowest improvements over SCCN, CFN, and BLSTM models, while the LSTM yields significant improvements over the noisy mixtures. By applying gate mechanisms to control the input, output, memory, and forget states, the LSTM is capable of utilizing the hidden state to model relationships among the past and current temporal frames, thereby increasing the generalization ability of the speech enhancement method. The BLSTM shows advantages over the LSTM, since it employs the forward LSTM to model relationships among the past and current temporal frames, and the backward LSTM within the BLSTM is applied to extract the interdependency among the current and future temporal frames.

Additionally, *t*-test is conducted between the proposed SCCN model and LSTM, BLSTM, CFN, and unprocessed noisy mixture to further evaluate whether the SCCN provides significant improvements based on PESQ and STOI. The significance level is set to 0.05, meaning that a *p*-value smaller than 0.05 indicates a statistically significant improvement. Table 2 shows the *t*-Test results, all *p*-values are below 0.05, and all H_0 are $+$. These results demonstrate SCCN offers statistically significant improvements compared to baseline approaches.

Table 2. The *p*-value of the *t*-test at 5% significance level, and comparison of the proposed approach with noisy mixtures, LSTM, BLSTM, and CFN models. H_0 denotes the null hypothesis, and $(+)$ indicates the improvement of two pairs is statistically significant at the 95% confidence level.

Measures	**STOI**		**PESQ**	
	***p*-Value**	H_0	***p*-Value**	H_0
Noisy	6.66×10^{-13}	(+)	6.33×10^{-20}	(+)
LSTM	1.18×10^{-15}	(+)	6.29×10^{-21}	(+)
BLSTM	9.85×10^{-14}	(+)	4.26×10^{-14}	(+)
CFN	3.49×10^{-10}	(+)	1.42×10^{-10}	(+)

4.4. Results and Analysis of the Second Experiment

The experimental results, evaluated using STOI and PESQ, are summarized in Figures 8 and 9. Figure 8 presents the STOI performance under noise conditions, while Figure 9 shows the PESQ results for noise scenarios.

Figures 8 and 9 illustrate the performance comparisons among the proposed SCCN model, GRN model, and AECNN model in terms of STOI and PESQ for three speaker-independent scenarios, including the seen noise, semi-unseen noise and unseen noise. The speaker-independent for seen noise scenario involves Machine noise, while the semi-unseen scenario involves Machinegun, Traffic and Car noises. The Crowd noise is fully unseen. As shown in Figures 8 and 9, SCCN, AECNN, and GRN models exhibit improvements over the noisy mixture, which demonstrates their capability to suppress environmental noise to different degrees.

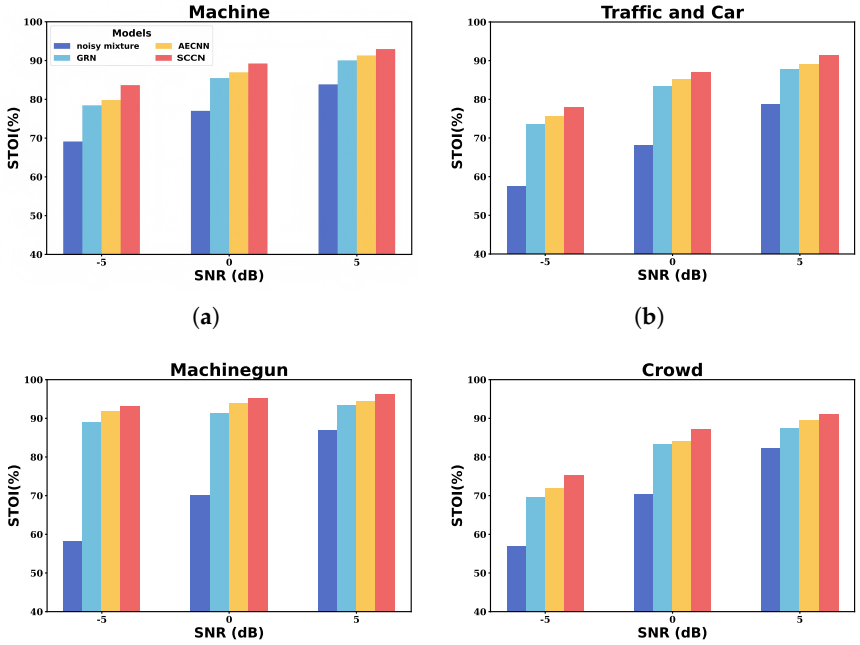


Figure 8. STOI scores for different speech enhancement models on the Machine, Traffic and Car, Machinegun, and Crowd noises, with subplots (**a**–**d**) corresponding to each noise.

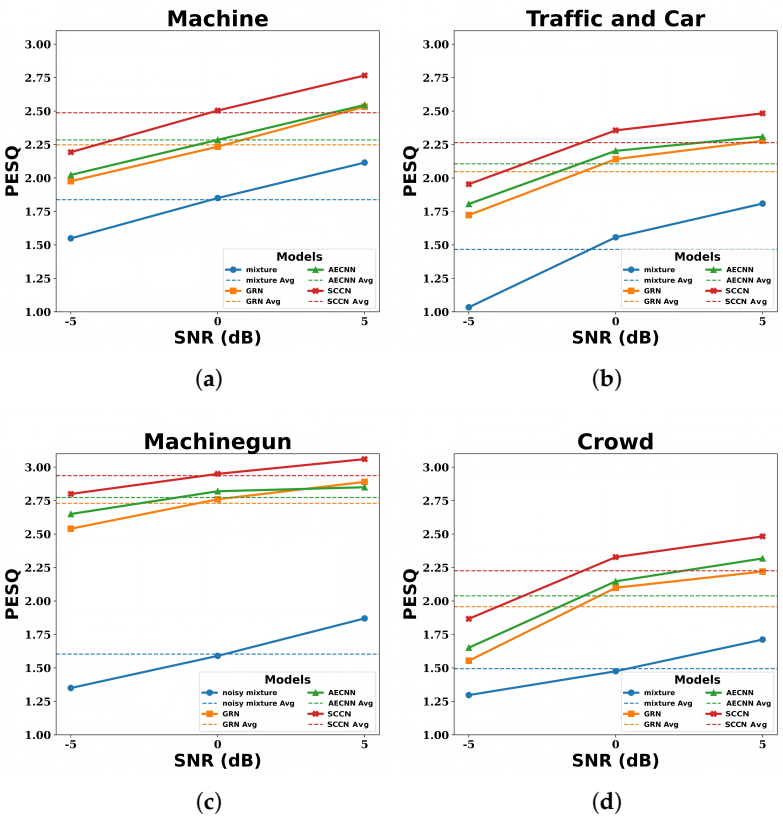


Figure 9. PESQ scores for different speech enhancement models on the Machine, Traffic and Car, Machinegun, and Crowd noises, with subplots (**a**–**d**) corresponding to each noise.

The proposed SCCN yields an average STOI score of 87.86% and a PESQ score of 2.51, achieving the best enhancement performance over the GRN and AECNN models. The SCCN model features several innovative structural designs to promote the model's capacity, with its core being the CCB and CDB. These blocks combine three convolution strategies:

regular convolution for feature extraction, DCLK to expand the receptive field, and DCLM to enhance channel representation while maintaining efficiency. Its multi-branch design captures features at various scales, improving detailed understanding. Additionally, the convolutional layers with exponentially growing dilation rates further promote global context recognition without extra computational cost. Moreover, the grouped ConvLSTM layers between the encoder and decoder are utilized to facilitate temporal feature extraction.

The AECNN attains average STOI and PESQ scores of 85.01% and 2.28, respectively, making it a more effective model than GRN model. AECNN improves speech enhancement performanceby using a convolutional encoder decoder neural network with skip connections, which utilizes MAE as the loss function. It further improvesthe speech clarity and quality in noise suppression compared to the MSE-based speech magnitude estimation method. Additionally, the model learns the speech structure during optimization, ensuring high-quality output and avoiding invalid STFT issues.

GRN achieves average STOI and PESQ scores of 83.10% and 2.23, respectively, indicating the lowest enhancement performance among CFN and AECNN models. The GRN applies dilated convolutions to enlarge the receptive field, enabling the capture of long-term contextual information while preserving resolution. The integration of skip connections that aggregate the outputs of all residual blocks facilitates the effective combination of features learned across layers. Additionally, the redesigned frequency-dilated module enables the GRN to learn local patterns across the temporal and spectral dimensions simultaneously. Lastly, replacing ReLU with ELU accelerates convergence and significantly enhances the model's generalization ability.

The *t*-test is conducted to assess whether the proposed SCCN model significantly outperforms GRN, AECNN, and the unprocessed noisy mixture in terms of PESQ and STOI. According to Table 3, all *p*-values are below 0.05 and all H_0 are rejected, confirming the statistical significance of SCCN's improvements over the baselines.

Table 3. The *p*-value of the *t*-test at 5% significance level, and comparison of the proposed approach with the baseline approaches for Noisy mixture, GRN and AECNN models.

Measures	**STOI**		**PESQ**	
	***p*-Value**	H_0	***p*-Value**	H_0
Noisy	4.84×10^{-6}	(+)	2.91×10^{-7}	(+)
GRN	1.78×10^{-9}	(+)	1.77×10^{-10}	(+)
AECNN	6.43×10^{-7}	(+)	3.15×10^{-10}	(+)

4.5. Spectrum Analysis

Single-channel speech enhancement has achieved notable progress, but its performance remains limited due to the lack of spatial cues, which hinders the ability to separate speech from noise. To further evaluate the efficacy of the proposed SCCN for the single-microphone recordings, we employ real room impulse responses (RIRs) [35] to obtain the single microphone recorded noisy mixture. For RIRs, the positions of the sound source (noisy mixtures) and two microphones are fixed in real rooms. The noisy mixtures are convolved with the RIRs to generate the reverberant noisy mixtures. Although both left and right microphones are used to capture the noisy mixtures, only those captured by the left microphone are used to train and test the SCCN model, reflecting a typical monaural speech enhancement scenario where spatial information is unavailable. Moreover, the reverberations of speech and noise increase the difficulty of monaural speech enhancement. We feed these single-channel recorded noisy mixtures as input to SCCN. After training, the model demonstrates significant noise reduction, successfully removing most of the noise while preserving the speech components. As illustrated in Figure 10, the enhanced speech

demonstrates a substantial suppression of noise. The experiments were conducted using F16 noise. The results indicate that our proposed SCCN can effectively enhance speech in difficult single-channel reverberant environments without spatial cues.

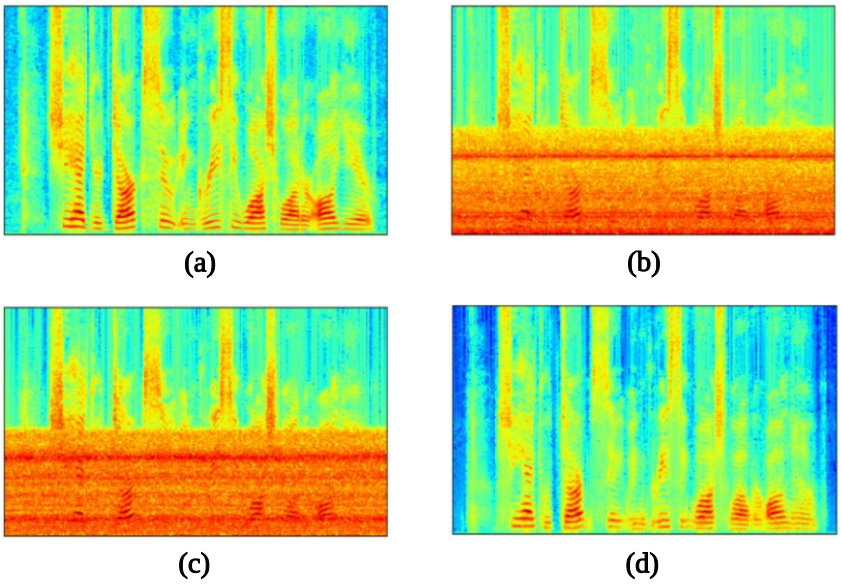


Figure 10. Spectra of clean speech, noisy mixtures, single-channel recorded noisy mixtures, and enhanced speech with F16 noise, with subplots (**a**–**d**) corresponding to each.

4.6. Component Analysis

The component analysis in Table 4 is performed to demonstrate the contribution of different components. Full denotes that all components of the proposed SCCN model are utilized. No-CCB means ablating the Combined Convolution Block and only using the standard convolution to build the encoder. No-CDB denotes deleting the Combined Deconvolution Block and using standard deconvolution to build the decoder. No-CCB-CDB means that only standard convolution and deconvolution are utilized. No-DIA represents the non-dilation employed in the encoder and decoder. No GC denotes removing two grouped ConvLSTM layers.

Table 4. Component analysis for different components in the proposed SCCN method for monaural speech enhancement.

Component	STOI (%)	PESQ	Parameter (Millions)
Full	78.54	2.06	7.2
No-CCB	76.79	2.02	3.7
No-CDB	77.29	2.03	4.3
No-CCB-CDB	75.18	1.92	1.8
No-DIA	78.01	2.03	7.2
No-GC	78.06	2.04	5.6

As Table 4 shows, the CCB offers the highest STOI and PESQ improvements over all individual components, demonstrating the paralleled depthwise separable convolutions in CCB are critical. The separable convolution with a larger kernel size effectively captures contextual information, while the one with an increased depth multiplier helps extract diverse features per input channel. In comparison, the CDB contributes less than CCB, since the CCB-stacked encoder would provide more informative high-level features to feed the group ConvLSTM layers. Moreover, ablating CCB and CDB generates a larger performance decrease when compared to removing CCB or CDB, which confirms the importance of

separable convolutions in both encoder and decoder. Additionally, the dilated encoder and decoder enlarge the receptive field. Removing dilation would cause a slight performance decrease in terms of STOI and PESQ. Moreover, removing the grouped ConvLSTM layers causes a marginal performance decrease, reflecting long-term dependency captured by ConvLSTM improves the enhancement performance.

5. Conclusions and Future Work

We introduce a new architecture for monaural speech enhancement, where SCCN incorporates several novel techniques to enhance both performance quality and computational efficiency. We also employ the Combined Convolutional encoder to expand the receptive fields in two dimensions, which enables the network to capture features efficiently. Three convolution branches are utilized jointly to extract contextual and local information. Furthermore, the CCBs and CDBs exploit the exponentially increasing dilation kernels to expand the receptive fields of the encoder and decoder. Moreover, extensive experiments are performed based on the renowned, commonly used databases. In addition, the grouped ConvLSTM layers are employed to capture spatial and temporal features. The results again prove SCCN's superiority over leading baseline approaches, highlighting its effectiveness for audio-based speech enhancement. Building on this foundation, future work will consider extending the proposed framework to multi-channel speech enhancement, which may further improve robustness in challenging noisy environments. Meanwhile, the proposed SCCN will incorporate the complex spectrum and conformer blocks to boost enhancement performance.

Author Contributions: Conceptualization, Y.X., Y.F. and P.X.; methodology, Y.X., H.T. and Y.S.; software, Y.X.; validation, Y.X., Y.F. and Y.S.; formal analysis, Y.F.; investigation, Y.S.; resources, Y.X.; data curation, Y.S.; writing—original draft preparation, Y.X. and Y.F.; writing—review and editing, Y.X.; visualization, Y.F.; supervision, P.X., H.T. and Y.S.; project administration, Y.X.; funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by Science and Technology Project of Henan Province (232102210027) and Henan Youth Natural Science Foundation (242300420695).

Data Availability Statement: All original contributions of this study are provided within the article.

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

References

1. Yousif, S.T.; Mahmmod, B.M. Speech Enhancement Algorithms: A Systematic Literature Review. *Algorithms* **2025**, *18*, 272. [CrossRef]
2. Chang, R.J.; Chen, Z.; Yin, F.L. Distributed parameterized topology-independent noise reduction in acoustic sensor networks. *Appl. Acoust.* **2023**, *213*, 109649. [CrossRef]
3. Zheng, C.S.; Zhang, H.Y.; Liu, W.Z.; Luo, X.X.; Li, A.D.; Li, X.D.; Moore, B.C. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods. *Trends Hear.* **2023**, *27*, 23312165231209913. [CrossRef]
4. Zhang, J.; Li, C.H. Quantization-aware binaural MWF based noise reduction incorporating external wireless devices. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3118–3131. [CrossRef]
5. Amini, J.; Hendriks, R.C.; Heusdens, R.; Guo, M.; Jensen, J. Spatially correct rate-constrained noise reduction for binaural hearing aids in wireless acoustic sensor networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2731–2742. [CrossRef]
6. Natarajan, S.; Al-Haddad, S.A.R.; Ahmad, F.A.; Kamil, R.; Hassan, M.K.; Azrad, S.; Macleans, J.F.; Abdulhussain, S.H.; Mahmmod, B.M.; Saparkhojayev, N.; et al. Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Eng. J.* **2025**, *16*, 103405. [CrossRef]
7. Chen, J.D.; Benesty, J.; Huang, Y.T.; Doclo, S. New insights into the noise reduction Wiener filter. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2006**, *14*, 1218–1234. [CrossRef]
8. Ghael, S.P.; Sayeed, A.M.; Baraniuk, R.G. Improved wavelet denoising via empirical Wiener filtering. *Wavelet Appl. Signal Image Process. V* **1997**, *3169*, 389–399.

9. Guo, D.N.; Shamai, S.; Verdu, S. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory.* **2005**, *51*, 1261–1282. [CrossRef]
10. Martin, R. Spectral subtraction based on minimum statistics. *Power* **1994**, *6*, 1182–1185.
11. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process Lett.* **2013**, *21*, 65–68. [CrossRef]
12. Wang, D.L. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends Amplif.* **2008**, *12*, 332–353. [CrossRef]
13. Han, K.; Wang, Y.; Wang, D.; Woods, W.S.; Merks, I.; Zhang, T. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 982–992. [CrossRef]
14. Huang, P.-S.; Kim, M.J.; Hasegawa-Johnson, M.; Smaragdis, P. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 2136–2147. [CrossRef]
15. Weninger, F.J.; Erdogan, H.; Watanabe, S.; Vincent, E.; Le Roux, J.; Hershey, J.R.; Schuller, B.W. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *Latent Variable Analysis and Signal Separation, 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, 25–28 August 2015, Proceedings*; Springer: Cham, Switzerland, 2015; pp. 91–99.
16. Fu, S.-W.; Hu, T.-Y.; Tsao, Y.; Lu, X. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In Proceedings of the 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), Tokyo, Japan, 25–28 September 2017; pp. 1–6.
17. Park, S.R.; Lee, J.W. A fully convolutional neural network for speech enhancement. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1993–1997.
18. Tan, K.; Wang, D.L. A convolutional recurrent neural network for real-time speech enhancement. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3229–3233.
19. Tan, K.; Chen, J.T.; Wang, D.L. Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *27*, 189–198. [CrossRef]
20. Pandey, A.; Wang, D.L. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1179–1188. [CrossRef]
21. Pascual, S.; Bonafonte, A.; Serrà, J. SEGAN: Speech Enhancement Generative Adversarial Network. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 3642–3646.
22. Abdulatif, S.; Cao, R.; Yang, B. CMGAN: Conformer-Based Metric-GAN for Monaural Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 2477–2493. [CrossRef]
23. Zhang, X.Y.; Zhou, X.Y.; Lin, M.X.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
24. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
25. Xian, Y.; Sun, Y.; Wang, W.W.; Naqvi, S.M. Convolutional fusion network for monaural speech enhancement. *Neural Netw.* **2021**, *143*, 97–107. [CrossRef] [PubMed]
26. Strake, M.; Defraene, B.; Fluyt, K.; Tirry, W.; Fingscheidt, T. Fully convolutional recurrent networks for speech enhancement. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6674–6678.
27. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
28. Thiemann, J.; Ito, N.; Vincent, E. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. In Proceedings of the Meetings on Acoustics, Montreal, QC, Canada, 2–7 June 2013.
29. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM. In *NIST Interagency/Internal Report (NISTIR)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 1993.
30. Botinhao, C.V.; Wang, X.; Takaki, S.; Yamagishi, J. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In Proceedings of the 9th ISCA Workshop on Speech Synthesis (SSW 9), Sunnyvale, CA, USA, 13–15 September 2016; pp.146–152.
31. Varga, A.; Steeneken, H. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
32. Hu, G.N.; Wang, D.L. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2010**, *18*, 2067–2079.

33. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2008**, *16*, 229–238. [CrossRef]
34. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
35. Shinn-Cunningham, B.; Kopco, N.; Martin, T. Localizing nearby sound sources in a classroom: Binaural room impulse responses. *J. Acoust. Soc. Am.* **2005**, *117*, 3100–3115. [CrossRef]