Signal, Image and Video Processing
https://doi.org/10.1007/s11760-024-03500-x

ORIGINAL PAPER

Speech enhancement using deep complex convolutional neural network (DCCNN) model

Yasir Iqbal[1] · Tao Zhang[1] · Muhammad Fahad[1] · Sadiq ur Rahman[1] · Anjum Iqbal[2] · Yanzhang Geng[1] · Xin Zhao[1]

Received: 15 June 2024 / Revised: 26 July 2024 / Accepted: 2 August 2024

Abstract

In cases with highly non-stationary noise, single-channel speech enhancement is quite challenging, mainly when the noise includes interfering speech. In this situation, deep learning's success has contributed to speech enhancement to boost intelligibility and perceptual quality. Existing speech enhancement (SE) works in time–frequency domains only aim to improve the magnitude spectrum via neural network learnings; the latest research highlights the significance of phase in perceptual speech quality. Motivated by multi-task machines and deep learning this paper, proposes an effective and novel approach to the task of speech enhancement using an encoder-decoder architecture based on Deep Complex Convolutional Neural Networks. The proposed model takes input from the spectrograms of the noisy speech signals, consisting of real and imaginary components for complex spectral mapping, and it simultaneously enhances the magnitude and phase responses of speech. Considering unseen non-stationary noise categories, which interfere with speech, the proposed model enhances speech quality by approximately, 0.44 MOS points compared to state-of-the-art single-stage techniques. Moreover, it outperforms all reference techniques constantly and improves intelligibility under low-SNR settings. In contrast, against the baselines, we find an incredible enhancement of over 3 dB in SNR, and 0.2 in STOI. In addition, our method outperforms baseline SE techniques in low-SNR conditions in terms of STOI.

Keywords Background noise · Speech signals · Speech enhancement · Noise suppression · Deep learning · Deep complex convolutional neural networks

✉ Yanzhang Geng
gregory@tju.edu.cn

✉ Xin Zhao
zhaoxin_16@tju.edu.cn

Yasir Iqbal
iqbalyasir_0167@tju.edu.cn

Tao Zhang
zhangtao@tju.edu.cn

Muhammad Fahad
mfahadgull77@tju.edu.cn

Sadiq ur Rahman
sadiqkhahak@tju.edu.cn

Anjum Iqbal
anjumiqbal@mail.dlut.edu.cn

[1] School of Electrical and Information Engineering, Tianjin University, 92 Weijin Road, Nankai District, Tianjin 300072, China

[2] School of Software Technology, Dalian University of Technology, Dalian 116024, China

Abbreviations

DCCNN	Deep complex convolutional neural network
Mag-Phase	Magnitude-phase
Real-Imag	Real-imaginary
SE	Speech enhancement
CED	Convolutional encoder-decoder network
cMSA	Complex masked spectrum approximation
CNN	Convolutional neural network
CRM	Complex ratio mask
CSA	Complex spectrum approximation
DFT	Discrete Fourier transform
DNN	Deep neural network
IRM	Ideal ratio mask
LSTM	Long short-term memory
MA	Mask approximation
MMSE-LSA	Minimum mean-square error log-spectral amplitude
MOS-LQO	Mean opinion score for listening quality objective

Published online: 14 August 2024

MSA	Masked spectrum approximation
PESQ	Perceptual evaluation of speech quality
RNN	Recurrent neural network
ReLU	Rectified linear unit
SG-jMAP	Super-Gaussian joint maximum a posteriori
SNR	Signal-to-noise ratio
SNRI	Signal-to-noise ratio improvement
STFT	Short-time Fourier transform
STOI	Short-time objective intelligibility
T-F	Time–frequency

1 Introduction

In everyday circumstances, noise inevitably distorts speech waveforms that are recorded by equipment, which has a significant impact on real-world applications like telecommunication and hearing aid devices. Speech enhancement (SE) approaches based on deep learning are being developed to recover clean waveforms from degraded ones using neural networks, thereby improving speech perceived quality and mitigating the impact of noise.

One can broadly categorize the current state of SE approaches into two categories: time-domain methods and temporal-frequency (TF) domain systems. Neural networks were used by the time-domain SE approaches [1–5] to figure out how to map from noisy waveforms into cleaner versions and these approaches directly use audio signals to train the neural network [6]. Unluckily, because high-resolution waveforms were directly generated, this kind of approach continued to exhibit inefficiencies and quality constraints. TF-domain SE approaches performed superior performance. A deep neural network TF-domain SE methods aim to predict clean frame-level TF-domain representations and then reconstruct the enhanced waveforms [7–11]. In [12], a TF-domain based model called TFADCSU-Net was presented and it improves information flow inside the model and prevents a notable increase in computing complexity as the number of network layers rises.

Generally, phase is not included in commonly utilized representations since the challenge is tremendous to enhance phase spectra directly, given its wrapping and nonstructural properties. However, recent research has shown how phase information is crucial to the speech perception quality of SE approaches, particularly when signal-to-noise (SNR) is low [13]. In previous studies, the researchers merely improved the magnitude spectra and used the noisy phase and enhanced magnitude spectra to reconstruct the waveforms utilizing an inverse short-time Fourier transform (ISTFT) [7–10]. In [14] TFA-S-TCN model proposed which primarily concentrate on improving the magnitude spectrum and making use of the noise mixture's phase for reconstructing the signal. In absence of phase spectrum enhancement inevitably has led to degradation of enhanced speech quality. To solve above issues, several approaches concentrated on enhancing short-time complex spectra, that quietly restored jointly clean magnitude and phase spectra [15, 16]. A recent study also suggested refining the complex spectrum after enhancing its magnitude [17, 18]. This can help to mitigate the unbounded estimation issue [19, 20] that was present in the methods that solely improved the complex spectra.

Still, there was an imperfect phase estimation due to the compensating effect [21] between the phase and magnitude. In order to achieve this goal, a large number of DNN-based phase-in algorithms are proposed going forward and may be broadly classified into two groups: complex-domain-based [21–23]and time-domain-based [24, 25]. For complex domains, the implicit relative relationship between the real and imaginary (RI) elements contains phase information. For instance, in [26], authors suggested using fully-connected (FC) layers stacked one on top of the other for estimating complex ratio mask (CRM), it is then given individually to RI portions of the spectrum in order to recover phase and magnitude concurrently. Nevertheless, the goal dynamic range is typically compressed using the nonlinear function, which somewhat impairs network training.

In supervised speech enhancing, deep neural networks (DNNs) showed remarkable efficacy [27] and use of DNNs for SE has shown tremendous improvement over the classical methods [28]. Although effective in noise-independent SE, deep neural network (DNN) methods are not very good at generalizing speaker features [29]. Even though vanilla DNN is a strong model, its efficacy in mismatched scenarios, such as speaker-independent or noise-independent circumstances, may be limited since the interdependence between nearby temporal frames is not explicitly taken into account [29]. A convolutional recurrent network (CRN) utilizes to directly map RI components in Tan et al.'s more modern [15] complicated spectrum mapping approach, which produces empirically better results than CRM. Fu and others. Convolutional neural networks, or CNNs employed for speech augmentation recently. T-F illustration of speech mixed noise is used as input for CNN in speech improvement, which is driven by CNN-based image processing techniques, just to estimate target speech [30]. The performance of CNN used by the authors in [31] for estimating clean complex spectrograms straight from noisy spectrograms was better than that of the DNN-based magnitude processing technique.

Convolutional encoder-decoder (CED) is a principle that has been received from computer vision research and forms the foundation of several successful CNN model designs [32, 33] and recently CED mechanism has been utilized in Speech Enhancement techniques to enhance feature information [34, 35]. An encoder and decoder were employed to preserve the original information of audio signals in the [36] DCTDCCGRU based deep learning model. However, this model concentrates on frequency information and might not adequately capture time-domain properties. The temporal information in speech signals is essential for preserving the speech's naturalness and comprehensibility [37].

When compared to cutting-edge deep learning techniques for complicated spectrum mapping, the fully convolutional neural network (CNN) presented by the authors in [38] to process complex spectrogram in noise reduction has shown a significant improvement. Zhao et al. have also employed a CED network in a post-processing step to improve encoded and subsequently decoded speech, demonstrating impressive generalization capabilities even to unknown codecs [39]. In order to estimate the target voice, an auto-encoder convolutional neural network (AECNN) is presented [40], with mean absolute error (MAE) serving as a cost function. Two streams are employed within the PHASEN network, which uses phase information in improving performance of amplitude-based SE [41]. The authors of [42] suggested a CFN-based encoder–decoder features with numerous skip connections enabling monaural speech improvement, in contrast to traditional CNN architectures that simply use pooling layers to compress the feature dimension. Using strides deconvolutions or upsampling layers, the feature dimension in this model decompresses in the decoder section and compresses in the encoder section [43]. High-resolution structural information is preserved with skip connections integration going through layers of equal size within the encoder towards the decoder. This is particularly crucial for regression tasks like noise reduction when learning to map using a noisy speech spectrum towards a target clean speech spectrum of identical size is necessary. To increase the receptive fields and capture the interdependence between various frames, dilated CNN has been employed. In contrast to RNN-based techniques, gated residual networks (GRN) [44] have demonstrated superior performance when employed in conjunction with dilated CNN for speech enhancement. However there are still drawbacks to the previously listed strategies. For example, in a standard convolutional encoder-decoder network, using a high kernel size might boost the model's receptive fields, but at the expense of increased computational cost [40]. There was still potential for improvement in the speech quality because these techniques could not explicitly and precisely anticipate the clean phase spectra. Hence, for TFdomain SE techniques, it is imperative to do explicit prediction and optimization on the phase spectra.

InceptionNet and MHCED use multiple kernels of varying sizes to increase model capacity; however, using high kernel sizes [45] is likely to reduce parameter efficiency and restrict the model's applicability in resource-constrained applications. Each group takes half of the input sequence in the two group convolutions that the AlexNet employs in parallel at each layer [46]. But just a portion of the input sequence is used for each convolution group, which can restrict each kernel to only extracting a portion of the information throughout the entire input sequence to downgrade the effectiveness of the model possibly. Group convolution channels within ShuffleNet are suggested to be rearranged using the channel shuffle [47], so that the channels relate to one another. Furthermore, ShuffleNet generates a single feature by sequentially applying ordinary convolution and depth-wise convolution. By keeping separate feature maps of conventional and depth-wise convolutions, this one characteristic can be improved much further. The AECNN architecture only uses the skip connections throughout the encoder as well as the decoder to feed data stored in encoder layers to the appropriate decoder layers [40]. Though it may help improve enhancement performance, the encoder/decoder's internal information flow reuse has not been investigated.

1.1 Contribution

We suggest a novel framework with a Complex spectral mapping called Deep Complex convolutional neural network (DCCNN), based on an encoder and decoder with parallel magnitude and phase or real-imaginary spectra denoising, to get around the performance limitations of previous SE techniques in difficult circumstances. In the proposed deep learning model encoder helps in encoding input noisy magnitude and phase spectrums to compressed TFdomain features for the upcoming decoding process while the corresponding decoder masks magnitude as well decodes phase and gives an output of enhanced mag-phase spectrum, respectively in the last iSTFT used on the enhanced mag-phase spectrum to reconstruct clean signal waveforms. The phase decoder incorporates the parallel phase estimation architecture to predict the clean phase spectra directly. In accordance with findings from experiments, our proposed DCCNN achieves explicit predictions and optimizations of the phase and magnitude spectrums, which reduces the compensating impact between them and surpasses state-of-the-art SE methods. The proposed deep learning model is unique to have achieved the direct enhancement of phase spectra.

1.2 Problem description

Time-domain noisy audio signal x(t) in real-time environments is a combination of additive noise n(t) and clean speech signal s(t), where t represents a discrete-time element. This noisy audio signal is calculated mathematically in Eq. (1).

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

This signal undergoes a transformation into the frequency domain using STFT that is utilized over consecutive frames. STFT of the noisy mixed signal is measured as:

$$X(k,\ l) = \sum_{\{n=-\infty\}}^{\{\infty\}} x(n)w(n\ -\ lH)e^{-j2\lambda kn/N} \tag{2}$$

In Eq. (2) window function is represented by w(n), hop size is H, N is FFT size, while k and l represent frequency bin and frame index, correspondingly. This gives a complex spectrogram:

$$X(k,\ l) = X_r((k,\ l)) +\ j\ X_i((k,\ l)) \tag{3}$$

In Eq. (3) $X_r((k,\ l))$ is real whereas $X_i((k,\ l))$ is imaginary component of the complex spectrogram, respectively.

1.3 Deep learning-based denoising process

A proposed neural network model is designed for estimating a clean complex spectrogram, $\widehat{S}(k,\ l)$, from this noisy spectrogram by effectively learning to reconstruct both the magnitude and phase components. The deep learning model is trained by utilizing a loss function of Mean Squared Error (MSE), specially formulated to handle the complex nature of the spectrogram. MSE loss is calculated as:

$$MSE = \frac{1}{KL}\sum\nolimits_{\{k=1\}}^{K}\sum\nolimits_{\{l=1\}}^{L} \times \left(\left(S_r(k,\ l) - \widehat{S}_r(k,\ l)\right)^2 + \left(S_i(k,\ l) - \widehat{S}_i(k,\ l)\right)^2\right) \tag{4}$$

where by L symbolizes the entire frames present while K entire quantity of frequency bins. This loss function ensures that the model will make accurate predictions for both real S_r and imaginary S_i portions, enabling a more accurate clean speech signal reconstruction. The proposed technique has the ability for maintaining well the natural dynamics and the timbre of the speech, resulting in an output that is less noisy and retains much of the original characteristics of the speech by addressing both components. After training, the estimated clean spectrogram $\widehat{S}(k,\ l)$ is processed by the inverse operation of the STFT, iSTFT, for getting a time-domain audio signal and iSTFT is illustrated by:

$$\widehat{s}(t) = \sum_{\{l=-\infty\}}^{\infty} \widehat{S}(k,\ l)w(t\ -\ lH)e^{j2\lambda kt/N} \tag{5}$$

In this way producing the denoised speech signal $\widehat{s}(t)$. This comprehensive method not limited to only enhancing denoised speech quality and intelligibility but also demonstrates significant enhancements in Signal-to-Noise Ratio (SNR), as confirmed through empirical research. Integration of simultaneous magnitude, as well as phase reconstructions in complex spectrogram processing, exemplifies a robust approach to managing real-world noisy speech signals.

2 Proposed methodology

The proposed model, DCCNN, is trained by supervised learning using features from the Fourier spectrum and its purpose is to give an estimation of clean audio signals from the noisy audio signals. The model inputs 13-time frame sequences at a time, all of which consist of real-imaginary spectrograms derived from audio signals. This dual-component approach has the phase information weighted for better quality of the reconstructed audio signal. The proposed model learns a mapping from a real-imaginary noisy signal feature X(k, l) to an estimated clean real-imaginary signal feature $\widehat{S}(k,\ l)$, as given in below Fig. 1.

In proposed methodology processing of data is prepared starting with clean audio signal (target) and noisy audio signal (source) x(t) as shown in Fig. 2 and below steps is followed in the methodology of deep learning Speech Enhancement model:

1. Given N raw waveform signals for clean and noisy speech, overlapped framing is applied.
2. Apply the windowing analysis w(n) to the framed signals.
3. Convert the framed and windowed signals into the required representation, with help of Short-Time Fourier Transform (STFT) and real-imaginary spectrogram X(k, l) is obtained.
4. Create an annotated data set with noisy and clean speech pairs features ($noisy\_speech\_real\text{-}img\_spectrograms_i$, $clean\_speech\_real\text{-}img\_spectrograms_i$) with $1\ \leq\ i\ \leq\ N$.

The training process for DCCNN follows below steps:

5. Train the DCCNN model in a way such that it minimizes an objective function while estimating clean features, from noisy features. $\widehat{S}(k,\ l)X(k,\ l)$

For denoising:

6. As shown in the figure a new noisy feature X(k, l) is applied to the proposed trained DCCNN giving an estimated clean feature $\widehat{S}(k,\ l)$.
7. iSTFT is used to estimate clean feature $\widehat{S}$, and time domain frames of the denoised signal obtained.
8. Then, synthesis windowing w(n) is applied to the time-domain denoised frames of $\widehat{s}(t)$ to reduce the spectral leakage.

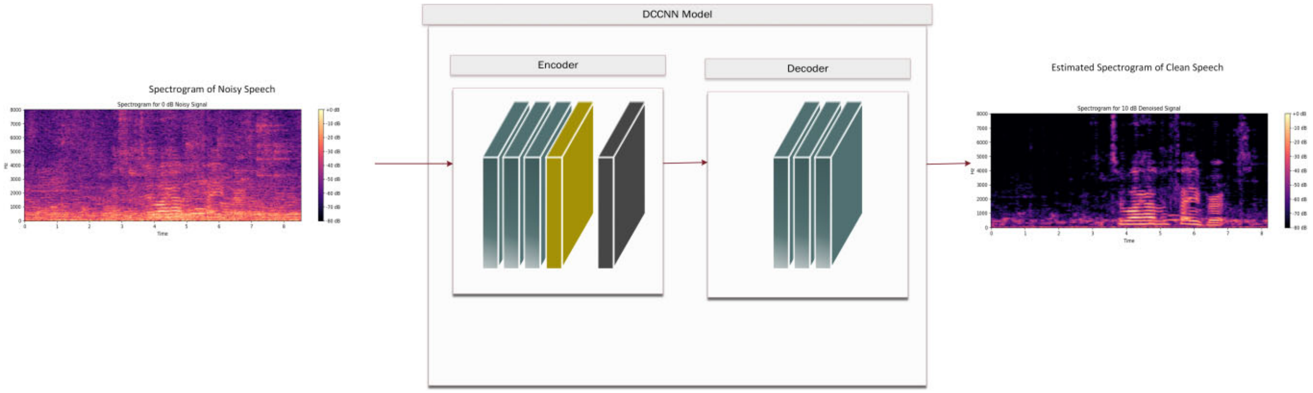

DCCNN Model
Encoder
Decoder
Spectrogram of Noisy Speech
Spectrogram for 0 dB Noisy Signal
Estimated Spectrogram of Clean Speech
Spectrogram for 10 dB Denoised Signal
Time
Hz
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Fig. 1 DCCNN taking Noisy (mixed) spectrograms as input for supervised training and directly maps it out to a clean one during speech enhancement

9. Finally, the overlap-add method is used to obtain the final time-domain denoised signal s(t) from the estimated clean frames. This step ensures a smooth reconstructed audio signal.

The steps in the suggested methodology are summarized in Fig. 2 where the input to the proposed DCCNN model is exploited in the form of time–frequency (spectrogram) of a noisy speech.

3 System overview

3.1 Network architecture

Recommended Deep Complex Convolutional Neural Network (DCCNN) is developed as a convolutional encoder-decoder architecture for processing of real-imaginary as well as magnitude-phase spectrograms of audio signals in SE. The components including different layers of DCCNN deep learning model are illustrated in Fig. 3. The complex structure provides a full representation of sound, which makes it better to differentiate between the two components speech and noise components of it. The proposed DCCNN takes the real-imag or mag-phase spectrum of a noisy mixture as input consists of 13 frames, and produces an estimate of the targeted speech's magnitude-phase spectrum or real-imaginary spectrum. Estimated target speech is reconstructed with the help of estimated real-imaginary or magnitude-phase of target speech. At the heart of the encoder, convolutional layers are stacked, with each having a LeakyReLU activation and batch normalization, starting from 16 filters and finally rising to 128. The encoder serves for obtaining major characteristics from the given input data, spatial dimension are preserved of the input sequence by using the strides and depth is expanded as the network goes deeper.

In proposed DCCNN structure decoder layers are reflected as part of the encoder where transposed convolution layers are applied that reconstruct the audio signal to form an enhanced output from the encoded features. Skip connections are further applied from layer to layer across the network to ensure that the flow and preservation of important features are maintained for high-quality reconstructed speech. More precisely, a skip connection is utilized for linking CCU's output to the matching symmetric CDU. These skip connections allow the correspondence between different convolutional units and deconvolutional units in the encoder and decoder, respectively, which becomes very important in holding the integrity of spatial and feature information.

3.2 Cluster convolutional units

The proposed DCCNN employs CCU which is encoder side of the proposed monaural speech enhancement model. In its design for a deep complex convolutional unit, it encodes input spectrograms efficiently for processing. The mathematical model applied for every convolutional layer associated with the encoder's 2D convolution is provided as follows:

$$\mathrm{C}(\mathrm{k},\ \mathrm{l},\ \mathrm{n}) = \sum_{\mathrm{i}=1}^{\mathrm{K}} \sum_{\mathrm{j}=1}^{\mathrm{L}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{F}\left(\mathrm{i},\ \mathrm{j},\ \mathrm{m},\ \mathrm{n}\right) \cdot \mathrm{X}\left(\mathrm{k}+\mathrm{i}-1,\ \mathrm{l}+\mathrm{j}-1,\ \mathrm{m}\right) \tag{6}$$

where C(k, l, n) is the output feature map at position (k, l) for the *n*-th output channel. This equation summarizes the idea of convolution as a way of simply transforming the input feature matrix X across its spatial dimensions while modifying the channel depth from M to N thus encoding rich and complex patterns from the noisy speech inputs. Following the convolution, the output feature map is then passed through a LeakyReLU activation function [48] for introducing non-linearity. Each unit integrates a LeakyReLU

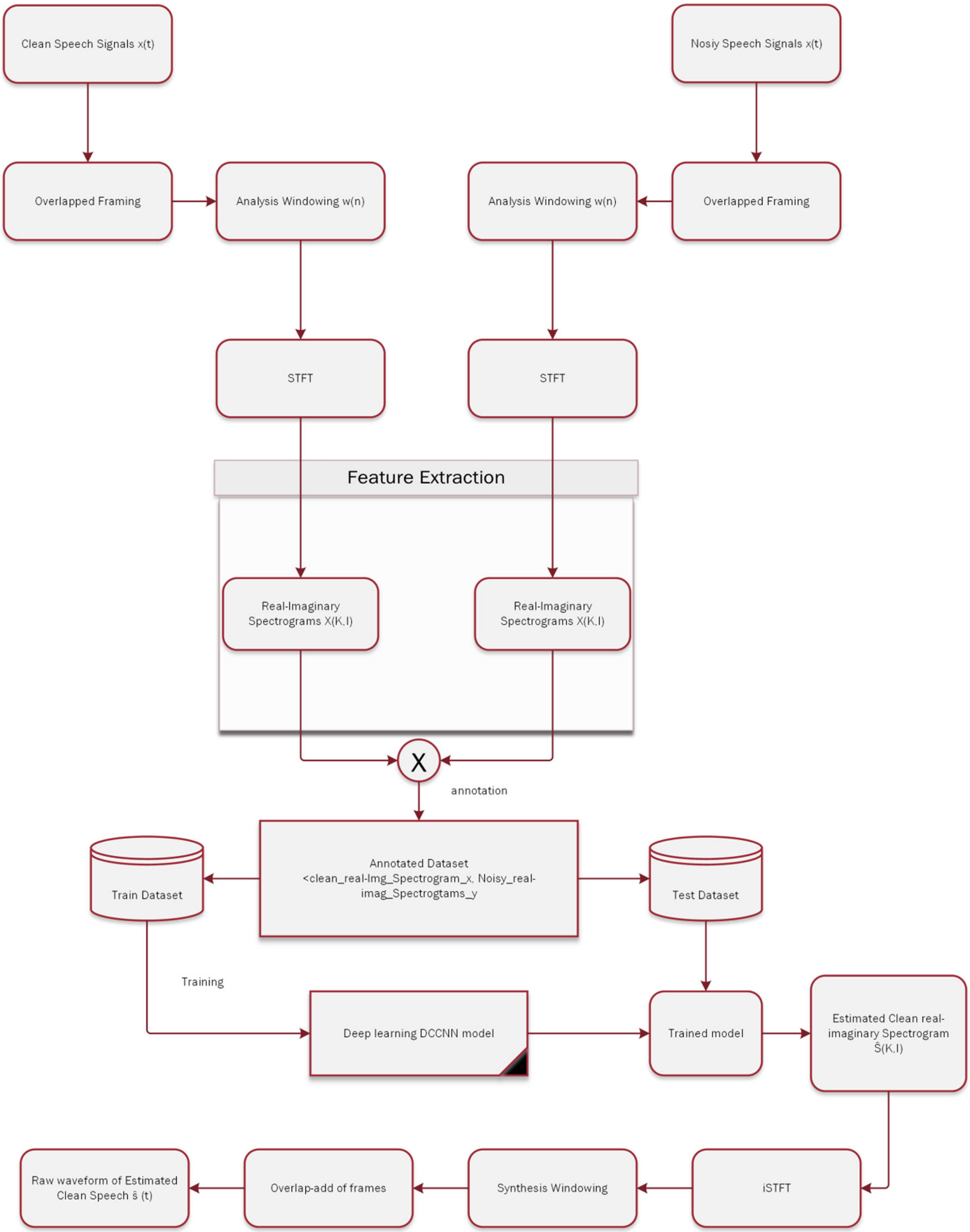

Clean Speech Signals x(t)
Nosiy Speech Signals x(t)
Overlapped Framing
Analysis Windowing w(n)
Analysis Windowing w(n)
Overlapped Framing
STFT
STFT
Feature Extraction
Real-Imaginary Spectrograms X(K.l)
Real-Imaginary Spectrograms X(K.l)
X
annotation
Train Dataset
Annotated Dataset <clean_real-Img_Spectrogram_x, Noisy_real-imag_Spectrogtams_y
Test Dataset
Training
Deep learning DCCNN model
Trained model
Estimated Clean real-imaginary Spectrogram Ŝ(K.l)
Raw waveform of Estimated Clean Speech ŝ (t)
Overlap-add of frames
Synthesis Windowing
iSTFT

Fig. 2 Demonstrating the procedures for DCCNN SE model training for fitting a noisy prediction function

activation function given by:

$$y = \max(0.01x, \; x) \tag{7}$$

This ensures that there is still small gradient flow even when the neuron is at rest, thus improving network's capability to learn nuanced characteristics during training. Another is activation function batch normalization, which is used for stabilization to quicken the learning by normalizing the output over each convolutional layer. For the encoder of

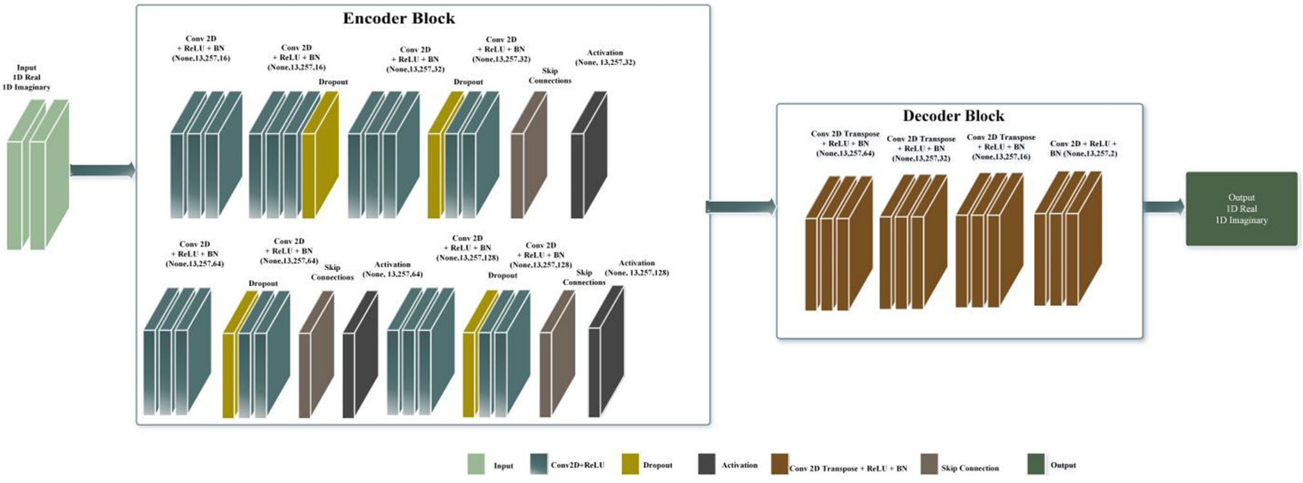


Fig. 3 Illustration of Proposed DCCNN architecture including all layers with dimensions. The labels of components are given at bottom of Fig. 3

the proposed DCCNN, this becomes a robust skeleton with which, in a complicated and efficient manner, it is possible to code audio signals; this is fundamental both for further decoding and for a subsequent improved speech output whenever background noise is present. Dropout layers are applied inside the proposed DCCNN architecture after certain convolutional layers in the encoder to randomly deactivate a fraction of units during training, reducing overfitting and encouraging robust feature learning. Unlike standard CNNs in proposed DCCNN's encoder either real and imaginary spectrums or magnitude and phase spectrums process to estimate clean data in the architecture.

3.3 Cluster deconvolutional units

The decoder segment of suggested DCCNN incorporating single channel speech enhancement is an assembly of cluster deconvolutional units carefully structured for reconstructing denoised audio signals from encoded characteristics. Presented operation is done using transposed convolutional layers (Conv2DTranspose), mathematically doing the operation:

$$\mathrm{C}'(\mathrm{k},\ \mathrm{l},\ \mathrm{n}) = \sum_{\mathrm{i}=1}^{\mathrm{K}} \sum_{\mathrm{j}=1}^{\mathrm{L}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{F}'(\mathrm{i},\ \mathrm{j},\ \mathrm{m},\ \mathrm{n})\ \ \mathrm{X}'(\mathrm{k}-\mathrm{i}+1,\ \mathrm{l}-\mathrm{j}+1,\ \mathrm{m}) \tag{8}$$

where X′ is input to layer, F′ denotes the filter matrix used for deconvolution, and C′ is the output feature map. This process spatially undoes the down-sampling that happened during the encoding—for the motive to upscale again, to the original dimensions of the spatial feature maps. Followed by each transposed convolution, it applies Leaky ReLU [48] using the DCNN:

$$\mathrm{y}' = \max\left(0.01\mathrm{x}',\ \mathrm{x}'\right) \tag{9}$$

This function reintroduces non-linearity into the upsampled outputs and ensures the maintenance of small gradients when units are inactive, aiding in the deep network's learning process. Batch normalization is then applied to normalize the outputs of each deconvolutional unit:

$$\hat{\mathrm{x}}' = \frac{\mathrm{x}' - \mu_{\mathrm{B}}^{'}}{\sqrt{\left\{\sigma_{\mathrm{B}'}^{2} + \epsilon\right\}}} \tag{10}$$

Here, $\mu_{\mathrm{B}'}$ and $\sigma_{\mathrm{B}'}^{2}$ represent the batch-wise mean and variance, respectively, ensuring that the learning process remains stable by maintaining consistent activation distributions. These cluster deconvolutional units (CDU) play a crucial role in restoring the detailed and nuanced audio features, ensuring the output audio signals are clear and closely resemble the original pre-noise conditions. Using batch normalization and Leaky ReLU in the decoder would stabilize the training, and the network will learn more advanced features when reconstructing the enhanced audio signal. Unlike standard CNNs the Decoder or CDU in the proposed DCCNN process either real-imaginary or mag-phase spectrums of noisy mixture audio signals to learn about speech enhancement.

3.4 Skip connections within encoder and decoder

A convolutional encoder-decoder processes the series of inputs through a number of layers. Certain details might get wasted because variation within dimensions of signal characteristic representations [40, 49]. In an effort to enhance reusing features, skip connections connecting encoder and decoder are implemented for overcoming this problem and skip connections in the architecture of proposed Deep Complex CNN played a vital role for Complex Spectrograms as a conduit for the denoising of the features of both real and imaginary spectrograms. On the contrary, this model diverts from conventional design ones and does not apply dense layers; rather, it makes use of convolutional layers, which are densely and unambiguous connected through skip connections. The connection hooks the direct pass in important paths from the encoder to the decoder without breaking them, resulting in this way for the passing of important information for the speech enhancement. In spectrogram denoising, the decoder has access to the information at the spectrum level from fine to high level, which has been extracted by the encoder using skip connections. Such a holistic perspective shall help the decoder to build denoised complex spectrograms of enhanced quality by using the power of both the richness of real and imaginary parts in a more effective manner. Such a strategically designed deployment of skip connections should optimally handle the flow of information; therefore, the proficiency of the model can be improved to disentangle complex relationships present in the spectrogram data for optimum denoising.

4 Selection and processing of data

4.1 Databases and preprocessing

During training and assessment of DCCNN clean speech data use is collected from TIMIT [50] and CSTR VCTK Corpus [51] and noise speech data from DEMAND [52] databases and all audios down sampled to 8 kHz. CSTR VCTK Corpus consists of 110 English speakers with 400 sentences. To start our framework for desired noise reduction against baseline methods initially cafeter, traffic, metro, bus noise files from DEMAND database and café noise type from the QUT [53] noise database whereas babble and restaurant noises from AURORA-2 [54] database mix with training, development and testing utterances. For each clean and noisy speeches we combine the two databases into a single, sizable set that has an overall amount of 45,150 audios of equal number of males and females speakers. We create unique test, development, and training setups utilizing 70%, 15%, and 15% of entire data, respectively. A total of $4 \times 7 = 28$ training conditions are produced for each batch of data by combining all files with a selected portion from each of the 7 noise samples and apply SNRs of 0, 5, 10, as well as 15 dB. In total of 45,150 speeches per clean and noisy set ($30{,}102 \times 6 \div 60 \div 60 = 50.17$) 50.17 h noisy audio mixtures use for training proposed DCCNN framework, and approximately 7,524 audios of 12.54 h noisy speech mixtures for each development and test set to evaluate our model. For development and test data SNRs of $-$ 5 and 15 dB unseen noise mix with clean data to analyze model Speech Enhancement performance.

4.2 Training and network parameters

The CNN-based Deep Complex network, which is an encoder-decoder model, is trained using standard backpropagation [55], employing a Mean Squared Error (MSE) loss function as defined in Sect. 1.4. During model training the Adam optimizer [56] with a 0.001 learning rate initially, 1 batch size, and other parameters like window length of 512, window shift of 256 for STFT providing frequency content and temporal smoothness, respectively, number of DFT set to 512 controlling frequency resolution, and a context window width of 13 frames are set to provide temporal context. The proposed model performance is tested with different values for optimizer, batch size, window length, window shift, DFT set and context window width. It is noted that high learning rate can cause divergence, smaller batch size avoid randomness and memory efficient for proposed model. The width of the context window for complex spectrograms in current network significantly affects the ability to capture the temporal dependencies in both clean speech and noise. A wider context window width allows the model to better differentiate and separate noise from the clean signal and improving the denoising process. During training proposed DDCCNN, learning rate adjusts using learning rate scheduler that decreases it by a factor of 0.9 after the 5th epoch to avoid underfitting and overfitting of model. If there is no decrease in loss after two epochs, training network resumes following epoch featuring most favorable development set loss. If the learning rate falls below 0.00001, the training is terminated. To get favorable speech enhancement performance proposed Deep Complex Convolutional Neural Network (DCCNN) model is tested with different number of layers and kernel size which have big impact on model's noise reduction capacity. It is noted large number of layers causes high computational resources and less number of layer has issue of underfitting of the proposed network during training. So finally the proposed network is designed with 4 encoder blocks and 4 decoder blocks ($L = 4$), each using a kernel size of 2×2. Number of convolutional filters starts at 16 in the first layer and doubles with each subsequent layer, up to a maximum of 128 filters. This structure of layers in encoder and decoder as well as other parameters mentioned above maintain stability between computational resources and model complexity and this settings give favorable speech enhancement results in terms of SNRI, PESQ and STOI.

4.3 Instrumental evaluation metrics

We decide using solely instrumental measurements on noisy speech x(t), clean speech reference s(t), and the enhanced speech $\widehat{s}(t)$. As a measure of the system's ability to suppress noise, signal-to-noise ratio improvement (SNRI) offered during network testing assesses in accordance with ITU-T G.160

[57]. SNRI choice for performance evaluation indicator is due to the fact as it reads how much noise has been reduced by proposed Deep Complex Convolutional Neural Network (DCCNN) network. In addition, we apply perceptual evaluation of speech quality (PESQ) [58] for obtaining mean opinion score for listening quality objective (MOS-LQO). In proposed network PESQ as a performance evaluation indicator tells approximate values to human listening objective metrics MOS for enhancement in speech quality in real time. Short-time objective intelligibility (STOI) metric [59] is utilized for evaluating the boosted speech's intelligibility. STOI metric, which has values in the range [0, 1], is especially intended for assessing noise suppression techniques therewith high values closely correlated with high intelligibility. By evaluation of proposed DCCNN with STOI scores indicates that after noise reduction the enhanced speech can be hearable or not for humans which is critical for every communication system.

5 Results and analysis

Several representative SE methods MMSE-LSA and SG-jMAP [60], LSTM-IRM [29], LSTMMSA, LSTMcMSA, CEDcSA-du, LSTMcMSA + DNNcSA, LSTMcMSA + CEDcSA-du and LSTM-cMSA + CEDcSAtr [61] and LSTM-cMSA + CLED-cSA-du [62] were selected to compare with proposed DCCNN. Which are discussed below based on Seen noise types and Unseen Noise types.

5.1 Seen noise types results

The results achieved from processing data of development class employing noise types that were observed throughout training are displayed in Table 1. The proposed deep learning-based approach DCCNN performs significantly better than the existing deep learning methods and the traditional MMSE-LSA and SG-jMAP [60] in context of PESQ, STOI, as well as SNRI, not only on average measurements but under any single SNR situation too. Most remarkably, the raw noisy speech owns an average achieved across STOI values of 0.75, and conventional procedures cannot increase intelligibility in context of objective metric STOI. Suggested deep learning-based approach, on the other hand, improves on that value by as much as 0.16 points (0.91) on average across SNR conditions.

Further, for the extremely difficult − 5 dB condition, PESQ throughout raw noisy speech is only marginally improved by MMSE-LSA and SG-jMAP, while PESQ is greatly enhanced by the proposed deep learning-based method DCCNN, despite not seeing a comparable low SNR during training. By contrasting the LSTM-IRM along with LSTM-MSA single-stage baselines, we find that DCCNN consistently outperforms them in context of PESQ, alongside average enhancements about 0.65 MOS points respectively. This supports the idea that optimization is more beneficial in the speech spectral domains than masking domains. In comparison to LSTM-MSA (17.89 dB) at an average across entire SNR circumstances, proposed DCCNN exhibits superior noise cancellation capabilities than multiple stages filtering networks.

Nonetheless, DCCNN outperforms all standard approaches in terms of PESQ values in low-SNR situations of − 5 and 0 dB. This could occur since by employing the real along with imaginary elements of clean speech spectrum $\widehat{s}(t)$ as target, DCCNN implicitly estimates the clean magnitude and phase of noisy mixture complex spectrum. Particularly during low-SNR situations, DCCNN nevertheless offers far more noise reduction in terms of SNRI than any other single-stage and multiple-stage technique. We decide to use the suggested DCCNN because of this finding as well as the fact that it offers most favored average Speech Enhancement (SE) ability on the development class. It obtains an average PESQ improvement of 0.41 points (3.04), average SNRI of 0.12 points (25.46), and average STOI of 0.03 points (0.91) across the development set over all SNR conditions in comparison to best baseline method LSTMcMSA + CEDcSA-tr. It is understandable that this discovery to mean that, even with similar extra noise suppression, the DCCNN network is more effective at reconstructing lost or damaged speech segments, leading to improved total speech quality in terms of PESQ. The DCCNN approach considerably enhances STOI throughout all SNR levels with regards to intelligibility. In case of lower SNR level of − 5 dB, where upgrading intelligibility is particularly essential it delivers gain of up to 0.13 points (0.82) in STOI scores against best baseline approach.

When comparing the outcomes from the test class and the development class, the examination for each set produces similar judgments regarding system rank as well as efficiency patterns across each of the factors as indicated in Table 2. Each model, even those using traditional methods which are not using development class during parameter adjustment, perform somewhat worse overall on the test class. This indicates, test class is marginally higher challenging in processing for various speech enhancement techniques, and deep learning techniques perform better when applied during processing test class samples. In comparison to the development class, the average results of the suggested technique DCCNN slightly little poorer in context of PESQ and STOI; nevertheless, in terms of SNRI, they are even marginally better than those of the high complexity referencing, LSTMcMSA + CLEDcSA − du.

Table 1 Development class seen noise comparison of proposed with baseline methods

SNR	Methods	PESQ	STOI	SNRI
- 5	Unprocessed	1.35	0.53	0
	MMSE-LSA	1.39	0.5	3.51
	SG-JMAP	1.38	0.49	4.13
	LSTM-IRM	1.61	0.65	11.76
	LSTM-MSA	1.65	0.66	12.61
	LSTM-cMSA	1.6	0.64	15.83
	CEDcSA-du	1.52	0.65	11.75
	LSTM-cMSA + DNN-cSA	1.61	0.68	17.07
	LSTM-cMSA + CED-cSA-du	1.63	0.69	17.12
	LSTM-cMSA + CED-cSA-tr	1.63	0.69	17.27
	Proposed-DCCNN	2.87	0.82	19.02 dB
0	Unprocessed	1.52	0.65	0
	MMSE-LSA	1.64	0.63	4.37
	SG-JMAP	1.63	0.63	5.19
	LSTM-IRM	1.99	0.79	17.14
	LSTM-MSA	2.08	0.8	19.21
	LSTM-cMSA	2.03	0.8	24.59
	CED-cSA-du	1.92	0.81	19.15
	LSTM-cMSA + DNN-cSA	2.06	0.83	26.55
	LSTM-cMSA + CED-cSA-du	2.12	0.85	26.1
	LSTM-cMSA + CED-cSA-tr	2.12	0.84	26.5
	Proposed-DCCNN	2.95	0.88	27.11 dB
5	Unprocessed	1.77	0.76	0
	MMSE-LSA	1.97	0.75	5.1
	SG-JMAP	1.98	0.75	6.14
	LSTM-IRM	2.42	0.87	18.39
	LSTM-MSA	2.53	0.88	20.2
	LSTM-cMSA	2.55	0.89	26.13
	CED-cSA-du	2.43	0.9	21.98
	LSTM-cMSA + DNN-cSA	2.61	0.91	28.6
	LSTM-cMSA + CED-cSA-du	2.7	0.92	27.79
	LSTM-cMSA + CED-cSA-tr	2.7	0.92	28.47
	Proposed-DCCNN	3.01	0.94	28.61
10	Unprocessed	2.11	0.86	0
	MMSE-LSA	2.36	0.84	5.54
	SG-JMAP	2.41	0.85	6.88
	LSTM-IRM	2.84	0.92	17.92
	LSTM-MSA	2.93	0.93	19.22
	LSTM-cMSA	3.01	0.93	25.51
	CED-cSA-du	2.97	0.95	24.09
	LSTM-cMSA + DNN-cSA	3.08	0.95	28.29
	LSTM-cMSA + CED-cSA-du	3.17	0.96	27.17
	LSTM-cMSA + CED-cSA-tr	3.19	0.96	28.16
	Proposed-DCCNN	3.27	0.96	28.48
15	Unprocessed	2.53	0.93	0

Table 1 (continued)

SNR	Methods	PESQ	STOI	SNRI
	MMSE-LSA	2.76	0.92	5.53
	SG-JMAP	2.87	0.91	7.22
	LSTM-IRM	3.24	0.92	16.64
	LSTM-MSA	3.32	0.96	18.2
	LSTM-cMSA	3.39	0.96	23.47
	CED-cSA-du	3.42	0.96	25.58
	LSTM-cMSA + DNN-cSA	3.43	0.97	26.39
	LSTM-cMSA + CED-cSA-du	3.51	0.97	25.03
	LSTM-cMSA + CED-cSA-tr	3.54	0.97	26.31
	Proposed-DCCNN	3.65	0.96	28.91
Mean	Unprocessed	1.86	0.75	0
	MMSE-LSA	2.02	0.73	4.81
	SG-JMAP	2.05	0.73	5.91
	LSTM-IRM	2.42	0.84	16.37
	LSTM-MSA	2.5	0.85	17.89
	LSTM-cMSA	2.52	0.85	23.11
	CED-cSA-du	2.45	0.85	20.51
	LSTM-cMSA + DNN-cSA	2.56	0.87	25.38
	LSTM-cMSA + CED-cSA-du	2.62	0.88	24.64
	LSTM-cMSA + CED-cSA-tr	2.63	0.88	25.34
	Proposed-DCCNN	3.04	0.91	25.46

5.2 Unseen noise types results

Table 3, displays results after analyzing the unseen noise test class in regard to baseline techniques, in which results average across the several forms of noise (traffic, cafeter, metro, bus, café, babble and restaurant noise). Observations of similar patterns and model ranks to the assessment for seen noise varieties are made once more, indicating that the techniques based on deep learning are generally well-suited to these very non-stationary unseen noise patterns. Particularly, when averaged across all SNR levels, the recommended DCCNN network outperform the current speech enhancement techniques and gives PESQ 0.08 MOS points (2.70), STOI 0.01 (0.90) and SNRI 0.93 points (23.40 dB) for PESQ, STOI, and SNRI, respectively. The proposed Deep Complex CNN (DCCCN) network is able to improve by 0.21 MOS points (2.26) for lower SNR conditions, such as -5, while the baselines method does not improve speech quality in the context of PESQ. Using the encoder and decoder of our proposed DCCNN system increases each of the quality indicators for all four analyzed noise types.

5.3 Model evaluation on environmental noises

In Table 4, the results were obtained for evaluating proposed deep learning model in contents of PESQ, STOI and SNRI for low SNR levels considering unseen environmental noises. With current encoder and decoder parameters complexity the proposed Deep Complex Convolutional Neural Network model highly suppress the traffic noise and gives high improvements in SNRI, PESQ and STOI up to 25.76 dB, 3.74 MOS, 0.94 compared to other noise types. For bus noise it is slightly worst performance due to the high interference of noise in speech signal in bus.

5.4 Analyzing DCCNN model

Improved speech spectrograms achieved from the proposed deep complex network were investigated using a case study of test class utterance in traffic noise at varied dB SNR circumstances in order to investigate more the root causes of the quality gains that have been noticed using the suggested deep complex network. The spectrograms for clean speech signal s(t) along with its corresponding noisy speech signal x(t), and denoised speech signal $\widehat{s}(t)$ are present in Fig. 4.

When examining the outputs, the spectrogram enrichment demonstrates how much greater noise reduction is made possible when using Deep Complex Convolutional Neural Networks (DCCNN) with current model complexity. Owing to intricate spectrogram processing, which improved the parallel noisy signal's phase and magnitude and assisted in reconstructing improved audio signals. Additionally, by

Table 2 Test class seen noise comparison of proposed with baseline methods

SNR	Methods	PESQ	STOI	SNRI
- 5	Unprocessed	1.35	0.53	0
	MMSE-LSA	1.38	0.49	3.34
	SG-JMAP	1.37	0.49	3.92
	LSTM-IRM	1.6	0.64	11.68
	LSTM-MSA	1.64	0.65	12.51
	LSTM-cMSA	1.58	0.64	15.56
	CED-cSA-du	1.52	0.64	11.45
	LSTM-cMSA + DNN-cSA	1.59	0.67	16.75
	LSTM-cMSA + CED-cSA-du	1.61	0.69	16.85
	LSTM-cMSA + CED-cSA-tr	1.61	0.69	16.99
	Proposed -DCCNN	2.88	0.84	17.94 dB
0	Unprocessed	1.52	0.65	0
	MMSE-LSA	1.63	0.62	4.18
	SG-JMAP	1.62	0.62	4.96
	LSTM-IRM	1.98	0.78	17.02
	LSTM-MSA	2.07	0.8	19.16
	LSTM-cMSA	2.02	0.79	24.44
	CED-cSA-du	1.92	0.8	18.74
	LSTM-cMSA + DNN-cSA	2.05	0.83	26.38
	LSTM-cMSA + CED-cSA-du	2.11	0.84	25.92
	LSTM-cMSA + CED-cSA-tr	2.11	0.84	26.32
	Proposed-DCCNN	2.97	0.91	27.31 dB
5	Unprocessed	1.77	0.76	0
	MMSE-LSA	1.96	0.74	4.9
	SG-JMAP	1.97	0.75	5.91
	LSTM-IRM	2.41	0.87	18.24
	LSTM-MSA	2.52	0.88	20.11
	LSTM-cMSA	2.54	0.88	26.04
	CED-cSA-du	2.42	0.9	21.65
	LSTM-cMSA + DNN-cSA	2.6	0.91	28.5
	LSTM-cMSA + CED-cSA-du	2.68	0.92	27.65
	LSTM-cMSA + CED-cSA-tr	2.69	0.92	28.34
	Proposed-DCCNN	3.19	0.94	29.12 dB
10	Unprocessed	2.11	0.86	0
	MMSE-LSA	2.35	0.84	5.35
	SG-JMAP	2.39	0.85	6.66
	LSTM-IRM	2.83	0.92	17.83
	LSTM-MSA	2.93	0.93	19.15
	LSTM-cMSA	3	0.93	25.52
	CED-cSA-du	2.96	0.94	23.91
	LSTM-cMSA + DNN-cSA	3.07	0.94	28.31
	LSTM-cMSA + CED-cSA-du	3.17	0.95	27.15
	LSTM-cMSA + CED-cSA-tr	3.18	0.95	28.14
	Proposed -DCCNN	3.29	0.96	28.62 dB
15	Unprocessed	2.53	0.93	0

Table 2 (continued)

SNR	Methods	PESQ	STOI	SNRI
	MMSE-LSA	2.75	0.9	5.38
	SG-JMAP	2.85	0.91	7.04
	LSTM-IRM	3.24	0.96	16.63
	LSTM-MSA	3.32	0.96	18.17
	LSTM-cMSA	3.39	0.96	23.5
	CED-cSA-du	3.41	0.97	25.57
	LSTM-cMSA + DNN-cSA	3.43	0.96	26.42
	LSTM-cMSA + CED-cSA-du	3.51	0.97	25.03
	LSTM-cMSA + CED-cSA-tr	3.54	0.97	26.3
	Proposed-DCCNN	3.63	0.97	29.02 dB
Mean	Unprocessed	1.86	0.75	0
	MMSE-LSA	2.01	0.72	4.63
	SG-JMAP	2.04	0.72	5.7
	LSTM-IRM	2.41	0.83	16.28
	LSTM-MSA	2.5	0.84	17.82
	LSTM-cMSA	2.51	0.84	23.01
	CED-cSA-du	2.44	0.85	20.26
	LSTM-cMSA + DNN-cSA	2.55	0.86	25.27
	LSTM-cMSA + CED-cSA-du	2.62	0.87	24.52
	LSTM-cMSA + CED-cSA-tr	2.62	0.87	25.22
	LSTM-cMSA + CLED-cSA-du	2.66	0.88	25.13
	Proposed -DCCNN	3.12	0.92	25.73

means of its skip connection, the suggested encoder and decoder architecture directly utilize high-resolution speech data intrinsic to the noise characteristics, which can also help with a more thorough reconstruction. This demonstrates that our recently suggested approach may accomplish similar speech restoring and noise reduction characteristics with far lower model parameters and computing resource consumption.

6 Conclusion

In this paper, we introduced a Deep Complex Convolutional Neural Network (DCCNN) which is a Speech Enhancement (SE) architecture using an encoder-decoder structure, with network arrangement specifically selected for the tasks of reducing noise and realistic sound speech restoration. DCCNN supervised model employing a complex network for complex-valued spectra mapping. The proposed model takes complex-valued input from the spectrograms of the noisy speech signals, consisting of real and imaginary components incorporating complex spectral mapping, which simultaneously perform enhancement of magnitude and phase dynamics of speech signals. The encoder encodes noisy magnitude and phase spectra, and the corresponding magnitude mask decoder and phase decoder decode out the enhanced magnitude and phase spectrums, respectively. The direct improvement of phase spectra to enhance PESQ and STOI of speech signals was the primary innovation of the DCCNN. In contrast against the baselines, we find an incredible enhancement of over 3 dB in SNR, 0.2 in STOI, and 0.5 in PESQ. In addition, our method outperforms baseline SE techniques in low-SNR conditions in terms of STOI. Moreover, it consistently surpasses all reference approaches and improves intelligibility in low-SNR environments. In future this proposed deep learning model can be evaluated by different datasets and can integrate alongside other deep learning models to gain better MOS points in terms of PESQ and SNRI.

Plan for future enhancements of the proposed DCCNN is to add more robust loss functions and use of depth-wise separable as well as vanilla convolutions with adaptive weights to train model on complex spectrograms with fewer parameters, introducing spatial information (up-down) and contextual dynamic changes.

Table 3 Test class unseen noise comparison of proposed with baseline methods

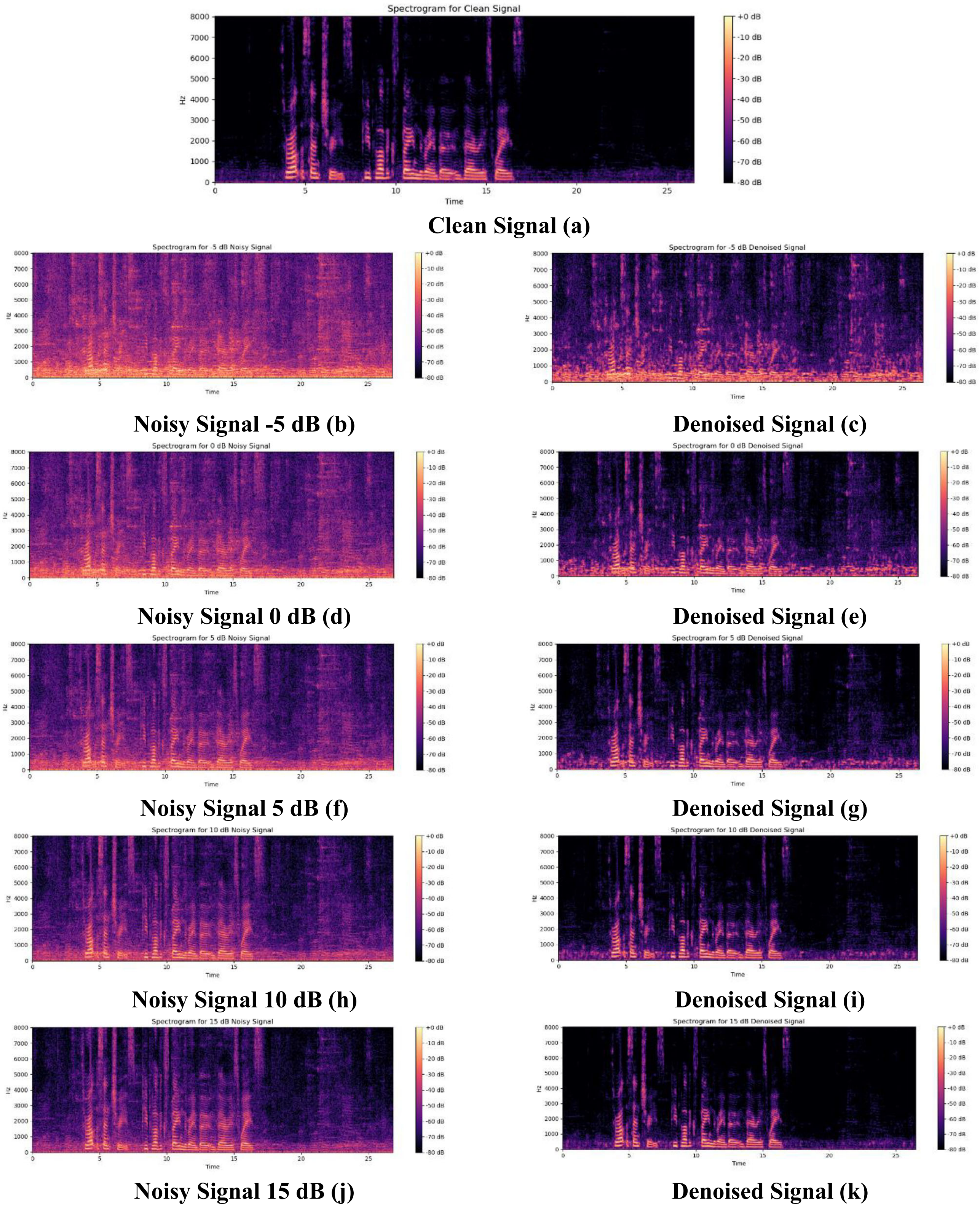
SNR	Methods	PESQ	STOI	SNRI
- 5	Unprocessed	1.4	0.55	0
	MMSE-LSA	1.39	0.51	3.01
	SG-JMAP	1.38	0.51	4.13
	LSTM-IRM	1.6	0.63	9.42
	LSTM-MSA	1.62	0.65	11.05
	LSTM-cMSA	1.57	0.64	13.65
	CED-cSA-du	1.55	0.67	11.59
	LSTM-cMSA + DNN-cSA	1.57	0.67	14.85
	LSTM-cMSA + CED-cSA-du	1.59	0.69	14.98
	LSTM-cMSA + CED-cSA-tr	1.59	0.69	15.15
	Proposed -DCCNN	2.10	0.79	16.13 dB
0	Unprocessed	1.59	0.68	0
	MMSE-LSA	1.65	0.65	3.74
	SG-JMAP	1.64	0.65	4.47
	LSTM-IRM	1.95	0.78	13.44
	LSTM-MSA	2.01	0.8	14.96
	LSTM-cMSA	1.98	0.8	19.12
	CED-cSA-du	1.89	0.82	16.33
	LSTM-cMSA + DNN-cSA	1.99	0.83	20.94
	LSTM-cMSA + CED-cSA-du	2.04	0.85	20.94
	LSTM-cMSA + CED-cSA-tr	2.05	0.85	21.27
	Proposed -DCCNN	2.26	0.88	23.51 dB
5	Unprocessed	1.86	0.8	0
	MMSE-LSA	1.98	0.77	4.41
	MMSE-LSA	1.98	0.77	4.41
	SG-JMAP	1.98	0.78	5.35
	LSTM-IRM	2.39	0.88	15.7
	LSTM-MSA	2.48	0.89	17.35
	LSTM-cMSA	2.49	0.89	22.07
	CED-cSA-du	2.39	0.91	19.73
	LSTM-cMSA + DNN-cSA	2.53	0.91	24.45
	LSTM-cMSA + CED-cSA-du	2.61	0.92	24.11
	LSTM-cMSA + CED-cSA-tr	2.63	0.93	24.71
	Proposed -DCCNN	2.70	0.94	26.41 dB
10	Unprocessed	2.22	0.89	0
	MMSE-LSA	2.37	0.86	4.85
	SG-JMAP	2.4	0.87	6.06
	LSTM-IRM	2.84	0.93	16.58
	LSTM-MSA	2.93	0.94	18.2
	LSTM-cMSA	2.97	0.94	23.19
	CED-cSA-du	2.96	0.95	23.4
	LSTM-cMSA + DNN-cSA	3.03	0.95	25.88
	LSTM-cMSA + CED-cSA-du	3.12	0.96	25.11
	LSTM-cMSA + CED-cSA-tr	3.15	0.96	26.06
	Proposed -DCCNN	3.18	0.96	27.21 dB

Table 3 (continued)

SNR	Methods	PESQ	STOI	SNRI
15	Unprocessed	2.65	0.95	0
	MMSE-LSA	2.77	0.92	4.88
	SG-JMAP	2.86	0.93	6.36
	LSTM-IRM	3.27	0.97	15.76
	LSTM-MSA	3.35	0.97	17.92
	LSTM-cMSA	3.39	0.97	22.66
	CED-cSA-du	3.44	0.97	25.69
	LSTM-cMSA + DNN-cSA	3.41	0.97	25.42
	LSTM-cMSA + CED-cSA-du	3.5	0.97	24.24
	LSTM-cMSA + CED-cSA-tr	3.54	0.98	25.55
	Proposed -DCCNN	3.58	0.98	27.01 dB
Mean	Unprocessed	1.94	0.77	0
	MMSE-LSA	2.03	0.74	4.18
	SG-JMAP	2.05	0.75	5.27
	LSTM-IRM	2.41	0.84	14.19
	LSTM-MSA	2.48	0.85	18.12
	LSTM-cMSA	2.48	0.85	20.14
	CED-cSA-du	2.44	0.86	19.35
	LSTM-cMSA + DNN-cSA	2.51	0.87	22.3
	LSTM-cMSA + CED-cSA-du	2.57	0.88	21.88
	LSTM-cMSA + CED-cSA-tr	2.59	0.88	22.54
	LSTM-cMSA + CLED-cSA-du	2.62	0.89	22.47
	Proposed-DCCNN	2.70	0.90	dB

Table 4 Results and Analysis for speech signals with different noise types

Noise Types	PESQ	STOI	SNRI									
SNR	− 6	−3	0	2	− 6	− 3	0	2	− 6	− 3	0	2
Unprocessed	1.13	1.27	1.59	1.88	0.58	0.62	0.78	0.85	0	0	0	0
Traffic	3.27	3.33	3.35	3.74	0.92	0.92	0.93	0.94	16.52	19.56	25.57	25.76
Cafeter	3.07	3.03	3.20	3.38	0.88	0.90	0.92	0.92	15.00	18.22	24.17	24.51
Metro	2.73	3.17	3.25	3.47	0.89	0.90	0.91	0.91	15.90	17.81	24.14	24.42
Bus	3.10	3.12	3.15	3.12	0.73	0.79	0.86	0.90	15.68	17.78	24.21	23.36



Clean Signal (a)

Noisy Signal -5 dB (b)

Denoised Signal (c)

Noisy Signal 0 dB (d)

Denoised Signal (e)

Noisy Signal 5 dB (f)

Denoised Signal (g)

Noisy Signal 10 dB (h)

Denoised Signal (i)

Noisy Signal 15 dB (j)

Denoised Signal (k)

Fig. 4 Spectrogram analysis for unseen Traffic noise at different dB SNR levels

Acknowledgements The authors acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 62271344).

Author contributions Conceptualization: YI, TZ, MF; implemented the System, YI, XZ, MF, SR wrote the paper, YI, SR, YG, AI and MF completed the experiments and YI, YG, AI collected the dataset and all authors reviewed the paper.

Funding This research was supported by National Natural Science Foundation of China (No.62271344).

Data availability All data created or investigated throughout this study are comprised in this available article.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable.

References

1. Saleem, N., Gunawan, T.S., Dhahbi, S., Bourouis, S.: Time domain speech enhancement with CNN and time-attention transformer. Digital Signal Process. **147**, 104408 (2024)
2. Kolbæk, M., Tan, Z.-H., Jensen, S.H., Jensen, J.: On loss functions for supervised monaural time-domain speech enhancement. IEEE/ACM Transactions Audio, Speech, Lang. Process. **28**, 825–838 (2020)
3. Pandey, A., Wang, D.: TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6875–6879, IEEE (2019)
4. Kong, Z., Ping, W., Dantrey, A., Catanzaro, B.: Speech denoising in the waveform domain with self-attention. In: ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) pp. 7867–7871, IEEE (2022)
5. Patel, A., Prasad, G.S., Chandra, S., Bharati, P., Das Mandal, S.K.: Speech enhancement using linknet architecture in speech and computer, pp. 245–257. Springer, Cham (2023)
6. Jannu, C., Vanambathina, S. D.: An attention based densely connected U-NET with convolutional GRU for speech enhancement. In 2023 3rd international conference on artificial intelligence and signal processing (AISP), 1–5, https://doi.org/10.1109/AISP57993.2023.10134933 (2023)
7. Pascual, S., Bonafonte, A., Serra, J.: SEGAN: Speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452, (2017)
8. Wahab, F.E., Ye, Z., Saleem, N., Ullah, R.: Compact deep neural networks for real-time speech enhancement on resource-limited devices. Speech Commun (2024). https://doi.org/10.1016/j.specom.2023.103008
9. Yang, Y., et al.: Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion. Biomed. Signal Process. Control **80**, 104279 (2023)
10. Shi, S., Paliwal, K., Busch, A.: On DCT-based MMSE estimation of short time spectral amplitude for single-channel speech enhancement. Appl. Acoust. **202**, 109134 (2023)
11. Kantamaneni, S., Charles, A., Babu, T.R.: Speech enhancement with noise estimation and filtration using deep learning models. Theoret. Comput. Sci. **941**, 14–28 (2023)
12. Parisae, V., Nagakishore Bhavanam, S.: Stacked U-Net with time–frequency attention and deep connection net for single channel speech enhancement. Int. J. Image Gr. p. 2550067 (2024)
13. Mamun, N., Hansen, J. H.: Speech enhancement for cochlear implant recipients using deep complex convolution transformer with frequency transformation. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
14. Jannu, C., Vanambathina, S.D.: Multi-stage progressive learning-based speech enhancement using time-frequency attentive squeezed temporal convolutional networks. Circuits, Syst., Signal Process. (2023). https://doi.org/10.1007/s00034-023-02455-7
15. Tan, K., Wang, D.: Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Transactions Audio, Speech, Lang. Process. **28**, 380–390 (2019)
16. Wang, Z.-Q., Wang, P., Wang, D.: Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. IEEE/ACM transactions audio, speech, lang process **28**, 1778–1787 (2020)
17. Yu, G., Li, A., Zheng, C., Guo, Y., Wang, Y., Wang, H.: Dual-branch attention-in-attention transformer for single-channel speech enhancement. In: ICASSP 2022–2022 IEEE International conference on acoustics, speech and signal processing (ICASSP) pp. 7847–7851, IEEE (2022)
18. Cao, R., Abdulatif, S., Yang, B.; Cmgan: Conformer-based metric gan for speech enhancement. arXiv preprint arXiv:2203.15149, 2022.
19. Wang, Z.Q., Wang, P., Wang, D.: Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR. IEEE/ACM Transactions Audio, Speech, Lang. Process. **28**, 1778–1787 (2020). https://doi.org/10.1109/TASLP.2020.2998279
20. Zheng, C., et al.: Sixty years of frequency-domain monaural speech enhancement: from traditional to deep learning methods. Trends in Hearing **27**, 23312165231209910 (2023)
21. Wang, Z.-Q., Wichern, G., Le Roux, J.: On the compensation between magnitude and phase in speech separation. IEEE Signal Process. Lett. **28**, 2018–2022 (2021)
22. Xu, R., Wu, R., Ishiwaka, Y., Vondrick, C., Zheng, C.: Listening to sounds of silence for speech denoising. Adv. Neural. Inf. Process. Syst. **33**, 9633–9648 (2020)
23. Zheng, C., et al.: Sixty years of frequency-domain monaural speech enhancement: from traditional to deep learning methods. Trends Hearing (2023). https://doi.org/10.1177/23312165231209913
24. Stoller, D., Ewert, S., Dixon, S.: Wave-u-net: a multi-scale neural network for end-to-end audio source separation. arxiv preprint arXiv:1806.03185, (2018)
25. Luo, Y., Mesgarani, N.: Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions audio, speech, language process. **27**(8), 1256–1266 (2019)
26. Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE/ACM transactions audio, speech, lang. process. **24**(3), 483–492 (2015)
27. Jannu, C., Vanambathina, S.D.: An overview of speech enhancement based on deep learning techniques. Int J Image Gr. (2023). https://doi.org/10.1142/s0219467825500019
28. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., Sainath, T.: Deep learning for audio signal processing. IEEE J. Selected Topics Signal Process. **13**(2), 206–219 (2019)
29. Chen, J., Wang, D.: Long short-term memory for speaker generalization in supervised speech separation. J Acoustical Society America **141**(6), 4705–4714 (2017)

30. Park, S. R., Lee, J.: A fully convolutional neural network for speech enhancement arXiv preprint arXiv:1609.07132, (2016)
31. Fu, S.-W., Hu, T.-y., Tsao, Y., Lu, X.: Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP) pp. 1–6 IEEE (2017)
32. Mao, X., Shen, C., Yang, Y.-B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Advances in neural information processing systems 29 (2016)
33. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
34. Zeng, J., Yang, L.: Speech enhancement of complex convolutional recurrent network with attention. Circuits, Syst, Signal Process (2023). https://doi.org/10.1007/s00034-022-02155-8
35. Parisae, V., Nagakishore Bhavanam, S.: Multi scale encoder-decoder network with time frequency attention and S-TCN for single channel speech enhancement. J. Intell Fuzzy Syst (2024). https://doi.org/10.3233/JIFS-233312
36. Jannu, C., Vanambathina, S.D.: DCT based densely connected convolutional GRU for real-time speech enhancement. J Intell Fuzzy Syst **45**(1), 1195–1208 (2023)
37. Parisae, V., Bhavanam, S.N.: Adaptive attention mechanism for single channel speech enhancement. Multimed Tools Applications (2024). https://doi.org/10.1007/s11042-024-19076-0
38. Ouyang, Z., Yu, H., Zhu, W.-P., Champagne, B.: A fully convolutional neural network for complex spectrogram processing in speech enhancement. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP) pp. 5756–5760, IEEE (2019)
39. Zhao, Z., Liu, H., Fingscheidt, T.: Convolutional neural networks to enhance coded speech. IEEE/ACM Transactions Audio, Speech, Lang Process **27**(4), 663–678 (2018)
40. Pandey, A., Wang, D.: A new framework for CNN-based speech enhancement in the time domain. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(7), 1179–1188 (2019)
41. Yin, D., Luo, C., Xiong, Z., Zeng, W.: Phasen: a phase-and-harmonics-aware speech enhancement network. In: Proceedings of the AAAI conference on artificial intelligence 34(05) 9458–9465 (2020)
42. Xian, Y., Sun, Y., Wang, W., Naqvi, S.M.: Convolutional fusion network for monaural speech enhancement. Neural Netw. **143**, 97–107 (2021)
43. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp. 1520–1528 (2015)
44. Tan, K., Chen, J., Wang, D.: Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ACM transactions audio, speech, lang process **27**(1), 189–198 (2018)
45. Grais, E. M., Ward, D., Plumbley, M. D.: Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders. In: 2018 26th european signal processing conference (EUSIPCO) pp. 1577–1581 IEEE (2018)
46. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 25 (2012)
47. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848–6856 (2018)
48. Maas, A. L., Hannun, A. Y., Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models. In Proc. Icml 30(1): Atlanta, GA, (p. 3). (2013)
49. Lan, C., Wang, Y., Zhang, L., Yu, Z., Liu, C., Guo, X.: Speech enhancement algorithm combining cochlear features and deep neural network with skip connections. J. Signal Process. Syst. (2023). https://doi.org/10.1007/s11265-023-01891-7
50. Garofolo, J. S.: Timit acoustic phonetic continuous speech corpus, Linguistic Data Consortium 1993.
51. Junichi Yamagishi, C. V., MacDonald, K.: CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). [Online]. Available: https://doi.org/10.7488/ds/2645
52. Thiemann, N. I. J., Vincent. E.: DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments (1.0). [Online]. Available: https://doi.org/10.5281/zenodo.1227121
53. Dean, D., Sridharan, S., Vogt, R., Mason, M.:The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms. In: Proceedings of the 11th annual conference of the international speech communication association: International speech communication association, pp. 3110–3113 (2010)
54. Pearce, D. J. B., Hirsch, H.-G.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Interspeech (2000)
55. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
56. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization arXiv preprint arXiv:1412.6980, (2014)
57. I. T. U.-T. S. S. (ITU-T) G.160 Appendix II, Objective measures for the characterization of the basic functioning of noise reduction algorithms, ITU (2012)
58. Recommendation I. T: Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU, Geneva (2001)
59. Taal, R. C. H. C. H., Heusdens, R., Jensen, J.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. Presented at the Proceedings of ICASSP, pp. 4214–4217. (2010)
60. Yu, H.: *Post-filter optimization for multichannel automotive speech enhancement*. Shaker (2013)
61. Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T.: Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. EURASIP J. Adv. Signal Process **2020**, 1–26 (2020)
62. Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T.: Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages. In: 2019 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), pp. 239–243. IEEE (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.