Speech Communication 156 (2024) 103008

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Compact deep neural networks for real-time speech enhancement on resource-limited devices

Fazal E Wahab [a], Zhongfu Ye [a,*], Nasir Saleem [b], Rizwan Ullah [c]

[a] National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, China
[b] Department of Electrical Engineering, Faculty of Engineering and Technology, Gomal University, D.I. Khan 29050, Pakistan
[c] Department of Electrical Engineering, Chulalongkorn University, Bangkok 10330, Thailand

ARTICLE INFO

Keywords:
Deep learning
Speech enhancement
Convolutional encoder–decoder
Recurrent networks
Quality and intelligibility
Phase estimation
Causal processing

ABSTRACT

In real-time applications, the aim of speech enhancement (SE) is to achieve optimal performance while ensuring computational efficiency and near-instant outputs. Many deep neural models have achieved optimal performance in terms of speech quality and intelligibility. However, formulating efficient and compact deep neural models for real-time processing on resource-limited devices remains a challenge. This study presents a compact neural model designed in a complex frequency domain for speech enhancement, optimized for resource-limited devices. The proposed model combines convolutional encoder–decoder and recurrent architectures to effectively learn complex mappings from noisy speech for real-time speech enhancement, enabling low-latency causal processing. Recurrent architectures such as Long-Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Simple Recurrent Unit (SRU), are incorporated as bottlenecks to capture temporal dependencies and improve the performance of SE. By representing the speech in the complex frequency domain, the proposed model processes both magnitude and phase information. Further, this study extends the proposed models and incorporates attention-gate-based skip connections, enabling the models to focus on relevant information and dynamically weigh the important features. The results show that the proposed models outperform the recent benchmark models and obtain better speech quality and intelligibility. The proposed models show less computational load and deliver better results. This study uses the WSJ0 database where clean sentences from WSJ0 are mixed with different background noises to create noisy mixtures. The results show that STOI and PESQ are improved by 21.1% and 1.25 (41.5%) on the WSJ0 database whereas, on the VoiceBank+DEMAND database, STOI and PESQ are improved by 4.1% and 1.24 (38.6%) respectively. The extension of the models shows further improvement in STOI and PESQ in seen and unseen noisy conditions.

1. Introduction

Speech enhancement for low-resource devices is an important research area aimed at improving the quality of speech in devices with limited computational capabilities or constrained resources. The goal of real-time SE is to achieve optimal results in enhancing speech intelligibility and quality while considering the constraints of low-resource platforms. In such situations, it becomes essential to develop efficient and lightweight speech enhancement models that can operate effectively on low-powered devices with limited memory and processing capabilities. These models need to strike a balance between computational efficiency and the ability to effectively reduce background noise and enhance speech. This paper focuses on enhancing single-channel noisy speech signals (Li et al., 2022a; Qiu et al., 2022; Lai and Zheng, 2019) to obtain optimal results without increasing the computational complexity of the models.

In recent decades, extensive research has been conducted by the speech-processing community on single-channel speech enhancement (SE). Drawing inspiration from the notion of time–frequency (T–F) masking in computational auditory scene analysis (CASA), SE has emerged as a supervised learning problem. However, for effective supervised speech enhancement, the selection of an appropriate training target is crucial. A well-defined training target may significantly enhance both speech quality and intelligibility. Conversely, the training target must be docile to supervised learning. In the T–F domain, numerous training targets have been devised, and they are primarily divided into two categories: masking and mapping targets. A masking-based target, like an ideal ratio mask (IRM) (Yang et al., 2023; Bao and Abdulla,

[*] Corresponding author.
E-mail address: yezf@ustc.edu.cn (Z. Ye).

https://doi.org/10.1016/j.specom.2023.103008
Received 10 July 2023; Received in revised form 11 October 2023; Accepted 9 November 2023
Available online 19 November 2023

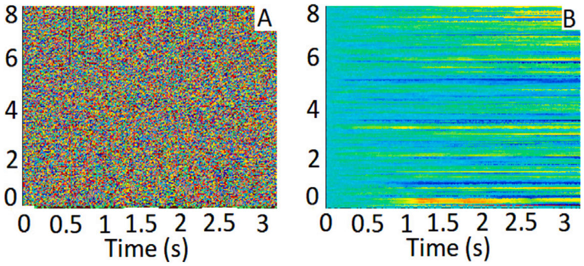

Fig. 1. Phase spectrum illustration of a speech signal.

2018; Saleem and Khattak, 2020) defines the T–F relationship between noise and clean speech. The target magnitude spectrum (TMS) (Tan and Wang, 2019) and log-power spectrum (LPS) are mapping-based targets describing clean speech in terms of spectral features.

Speech enhancement typically focuses on improving the magnitude spectrogram of noisy speech using the short-time Fourier transform (STFT) (Wang et al., 2014) and traditional methods use the noisy phase for waveform reconstruction (Boll, 1979; Xia and Bao, 2014; Ding et al., 2009). Conventionally, phase spectrogram enhancement has received less attention due to the challenging nature of direct clean speech phase estimation (Paliwal et al., 2011) and the perceived insignificance of its role in speech enhancement (Wang and Lim, 1982). Paliwal et al.'s study (Paliwal et al., 2011) emphasized the advantages of precise phase estimation, leading to the development of various phase enhancement techniques for speech enhancement (Yin et al., 2020; Choi et al., 2018; Zheng and Zhang, 2018; Saleem et al., 2021; Hasannezhad et al., 2022). These techniques include methods like mean squared error (MSE) estimation of phase spectra (Mowlaee et al., 2017), selective enhancement of voiced speech frames (Krawczyk and Gerkmann, 2014) and using phase decomposition and temporal smoothing for clean speech phase estimation (Mowlaee and Kulmer, 2015). Phase information integration can also be achieved via time–frequency (T–F) masking, as demonstrated by Wang et al. (2018), who employed a deep neural network (DNN) model to optimize masking and signal reconstruction. Phase-Sensitive Mask (PSM) (Lee et al., 2018), based on the difference between noisy and clean speech phases, enhances the signal-to-distortion ratio (SDR) compared to magnitude spectrum enhancement (Krawczyk and Gerkmann, 2014). Williamson et al. (2016) introduces the complex ideal ratio mask (cIRM) using a DNN to enhance both magnitude and phase (Wang and Bao, 2019), showing improved quality without significant effects on objective intelligibility compared to IRM estimates (Wang et al., 2018; Lee et al., 2018). In addition, CNN-based models (Williamson et al., 2016) estimate real and imaginary spectra from noisy speech, enhancing speech quality and intelligibility compared to DNN (Ouyang et al., 2019). Furthermore, in Xu et al. (2013) and Tan and Wang (2019), a DNN-based complex spectral mapping of noisy LSP features outperforms LPS spectral mapping, improving PESQ by 0.21 and STOI by 2.4

The use of recurrent neural networks (RNN) (Sun et al., 2017; Xia and Wang, 2015) and CNNs (Strake et al., 2020b; Pandey and Wang, 2019, 2021) has significantly improved supervised speech enhancement in the previous decade. In Liang et al. (2020), Gao et al. (2018), Liu et al. (2018), Wang et al. (2021b) and Takeuchi et al. (2020), speech enhancement is performed using RNNs with long short-term memory (LSTM) (Chen and Wang, 2017; Fu et al., 2016, 2019). Their experimental results show that for untrained test speakers, RNN generalizes very well, and in terms of STOI feedforward, DNN is substantially outperformed by RNN. Moreover, in Tan and Wang (2019) and Hasannezhad et al. (2020) CNNs have also been used for spectral mapping and mask estimation. Park and Lee (2017) performed spectrum mapping using a convolutional encoder–decoder network. With much fewer trainable parameters, such models show comparable denoising performance as compared to RNNs and DNNs. A similar encoder–decoder is presented by Grais et al. (2018). To capture long-term contexts, a gated residual network with dilated convolutions was proposed (Tan et al., 2018a). Another model combines CNN feature extraction and RNN temporal modeling capabilities. Naithani et al. (2017) stacked convolutional, recurrent, and fully connected layers successively to devise a similar model (Shi fas et al., 2020). In Strake et al. (2020a) an identical model was developed. A causal SE is developed by combining a convolutional encoder–decoder with LSTMs (Tan and Wang, 2018). Takahashi et al. (2009) devised a model that integrates recurrent and convolutional layers at several low scales.

Convolutional recurrent models (Tan et al., 2018b) have proven effective in improving speech quality and intelligibility, yet their computational requirements and extensive parameterization require further improvements for resource-constrained devices. To overcome these limitations, this study proposes a compact neural model designed in a complex frequency domain for speech enhancement to achieve optimal results in resource-limited devices. We examine the computational complexity of three recurrent models (LSTM, GRU Hasannezhad et al., 2020; Saleem et al., 2022, and SRU Hsieh et al., 2020; Cui et al., 2020) as bottlenecks in the convolutional encoder–decoder architecture to assess their suitability for real-time speech enhancement. We further extend the proposed models and integrate attention gates into the skip connections, enabling the selection of important spectral features by assigning attention weights. The proposed models demonstrate considerable improvements in speech intelligibility and quality while achieving high computational efficiency and reduced model size.

The remaining paper is organized as follows. Section 2 explains single-channel speech enhancement in the STFT domain. Section 3 presents the description of the proposed models. The experiments are presented in Section 4. Section 5 presents results and discussions. Section 6 concludes this research.

2. Speech enhancement in STFT domain

This section describes a single-channel speech enhancement (SCSE) in the STFT domain. Given a noisy speech $y(t)$, the SCSE aims to reduce the background noise $d(t)$ in a target speech $x(t)$. The noisy speech is given as:

$$y(t) = x(t) + d(t) \tag{1}$$

where t indicates the time sample index. The STFT gives the following expression:

$$Y_{t,f} = X_{t,f} + D_{t,f} \tag{2}$$

where Y, X, and D shows the STFT representation of y, x, and d, respectively, whereas t and f are the time and frequency bins. The polar representation is given as:

$$|Y_{t,f}|e^{(j\theta_{Y_{t,f}})} = |X_{t,f}|e^{(j\theta_{X_{t,f}})} + |D_{t,f}|e^{(j\theta_{D_{t,f}})} \tag{3}$$

where $|\cdot|$ and θ shows the magnitude and phase responses, respectively. In spectral mapping-based approaches, the magnitude spectrum of clean speech is often utilized as a training target. A mapping function is learned from noisy features in these approaches to estimate the corresponding clean magnitude. The estimated magnitude and noisy phase are combined to reconstruct the final speech waveform.

Fig. 1 shows the phase spectrograms of the speech signal, where the phase is covered between $[-\pi, \pi]$. Since the phase is wrapped (between $[-\pi, \pi]$), the phase spectrogram appears randomly organized. When the phase hops (at $\geq \pi$) around successive time–frequency units, multiples of 2π are used for phase adjusting, resulting in a smooth phase (as demonstrated by Fig. 1(B)). Both plots lack distinct structures; as a result, estimating the phase spectrum directly using supervised learning would be challenging.

The real and imaginary spectrograms exhibit spectrotemporal structures similar to the magnitude spectrogram, as illustrated in Fig. 2,

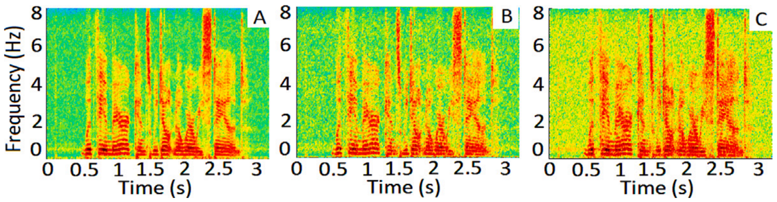

Fig. 2. Real and imaginary components of speech spectrum; (B) Real component, and (C) Imaginary component.

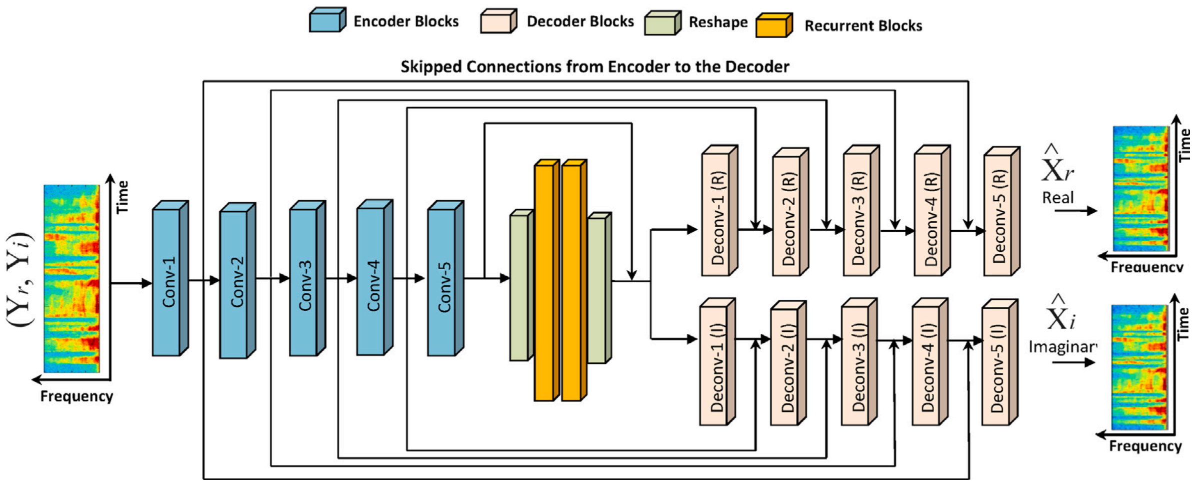

Fig. 3. The architecture of the proposed model for speech enhancement.

making them suitable for supervised learning. Consequently, we propose direct learning of spectral mapping from noisy speech (real and imaginary) to clean speech. While Williamson et al. (2016) indicated that using a deep neural network for estimating the real and imaginary parts of the STFT might be ineffective; however, recent studies (Wang et al., 2023) and our own study, have demonstrated the superior performance of complex spectral mapping over magnitude mapping and complex ratio masking.

3. Description of the proposed speech enhancement

The framework for the proposed speech enhancement depicted in Fig. 3, comprises an encoder that consists of Convolutional and pooling layers, responsible for extracting high-level features from the input speech data. In a similar fashion, the decoder, with a fundamentally equivalent structure as the encoder but in reverse order, maps the low-level features to the input feature size. This symmetrical structure of the encoder–decoder ensures consistent shaping of inputs and outputs. Causal convolutions are imposed on the encoder–decoder framework to design a real-time speech enhancement system.

Previously, two closely related architectures (Tan and Wang, 2019; Hu et al., 2020) were proposed in the literature; however, our proposed solution, while sharing some similarities with architectures (Tan and Wang, 2019; Hu et al., 2020), incorporates several key architectural differences that contribute to its compactness and memory efficiency. DCCRN (Hu et al., 2020) utilizes complex-valued DNNs with complex-valued operations, whereas our study uses real-valued DNNs. This choice is based on the findings of Wu et al. (2023), who examined that complex-valued DNNs, despite their higher computational complexity, exhibit similar performance to their real-valued DNNs. Similarly, GCRN (Tan and Wang, 2019) has utilized gated convolution operations in the architecture. A standard convolution operation efficiently captures local spectral features and nuances such as phonetic information, formants, and short-duration speech parts (such as consonants). In contrast, gated convolution operations, while effective at capturing long-range dependencies, may not perform well in preserving the local features and might be more prone to over-smoothing, potentially blurring fine details in the speech signal as well as introducing higher computational complexity (Yu and Koltun, 2015). Both studies (Tan and Wang, 2019; Hu et al., 2020) use LSTM as bottleneck layers and did not examine other recurrent layers which might be efficient both in terms of speech enhancement performance and computational load. However, our study deeply examines and compares multiple recurrent networks, including LSTM, GRU, and SRU as bottleneck layers. While conducting experiments, we found that SRU (as a bottleneck layer) is a better candidate as compared to LSTM in terms of SE performance and computational load. The studies (Tan and Wang, 2019; Hu et al., 2020) use simple skip connections to connect the encoder and decoder. However, the proposed model enhances the skip connections and adds attention gates to reduce redundancy and emphasize crucial spectral features. The attention gates allow our model to selectively focus on relevant spectral features, resulting in significant performance improvements with negligible complexity.

Fig. 4 illustrates the incorporation of causal convolutions with a time dimension. In this representation, we consider the inputs as a sequence of feature vectors, while ensuring that the outputs remain independent of future sequences of feature vectors. This approach allows for a causal encoder–decoder framework to be established. The architecture of the causal encoder–decoder consists of five Convolutional (Conv-2D) and Deconvolutional (Deconv-2D) layers. All layers, except the output layer, use exponential linear rectified unit (ELU) activation and batch normalization (BN). ELU activation is chosen for its quick convergence, improved generalization, and reduced complexity. Batch normalization is applied after each convolution (or deconvolution) layer and before the ELU activation. The number of kernels gradually increases in the encoder while steadily decreasing in the decoder, ensuring symmetric kernel numbers. To capture larger contexts, a stride of 2 is employed along the frequency direction in all convolutional (or deconvolutional) layers, while the time dimension of the features remains unchanged. To facilitate the flow of gradients and information throughout the network, skip connections are incorporated. Recently a

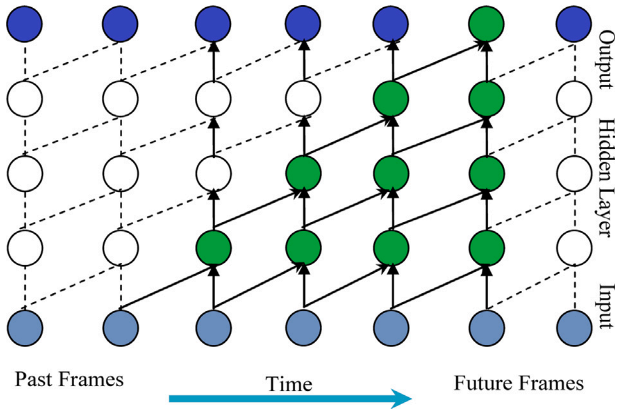

Fig. 4. Example causal convolutions. The convolution outputs do not depend on future inputs.

study (Yamanaka et al., 2017) adopts a better representation of skip connections.

3.1. Temporal modeling with gated RNNs

It is crucial to make use of the long temporal contexts to target the required speech. However, recurrent neural networks (RNNs) face a challenge when it comes to capturing vital information from lengthy sequences during backpropagation, primarily due to the vanishing gradient problem. Gradients, which are values to update neural network weights, can significantly diminish as they propagate backward in time. This issue arises when gradients become excessively small, resulting in minimal contributions to the learning process. Consequently, RNNs encounter short-term memory issues, leading to difficulties in retaining information from longer sequences. To address this short-term memory problem, advanced architectures such as LSTM, GRU, and SRU were developed. They feature internal systems known as gates that allow them to control the flow of information.

Causal LSTM, GRU, and SRU are selected, which process a contextual window composed of 11 frames (10 previous and 1 current) for estimating a single frame of the target speech. A concatenated vector comprising 11 frames of feature vectors is utilized during all-time steps. The study follows the same architecture for LSTM, GRU, and SRU; that is, from input to the output layer, network architectures are [(128 × T × 4), 512, 512] units where each layer contains 512 neurons and (128 × T × 4) shows the input size with T number of time-frames. For causal speech processing (suitable for real-time applications), no future information is included. The RNN-based models can achieve good SE (Saleem et al., 2020); however, the computational training load is large. Therefore, three gated recurrent models are used to examine the SE performance when integrated into the convolutional encoder–decoder framework. This study has examined all three models individually. LSTM involves the following computation:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{4}$$

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{5}$$

$$O_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{6}$$

$$g_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c) \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{8}$$

where W_i, W_f, W_o, are weight matrices of input, forget, and output gate associated with hidden states, x_t is input to the current timestamp, h_{t-1} is hidden state of the previous timestamp, c_{t-1} and c_t shows the previous and current timestamp respectively whereas ⊙ denotes element-wise multiplication. b_i, b_f, and b_o are the biased terms of input, forget, and output gate, respectively.

Unlike LSTM, the GRU has three gates and does not keep an internal cell state. The information, that is kept in an LSTM recurrent unit's internal cell state, is integrated into the gated recurrent unit's hidden state. The next GRU receives this information. GRU involves the following computations:

$$z_t = \sigma(W_z[x_t, h_{t-1}] + b_z) \tag{9}$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \tag{10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + (z_t) \odot \tilde{h_t} \tag{11}$$

$$\tilde{h_t} = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h) \tag{12}$$

where x_t is the input at time step t, h_{t-1} is the hidden state at the previous time step, and h_t is the updated hidden state at time step t. z_t is the update gate, r_t is the reset gate, and $\tilde{h_t}$ is the candidate hidden state. W_z, W_r, W_h, b_z, b_r, and b_h are the model parameters that are learned during training.

SRU is a recurrent neural unit with the same parallelism as convolution and feedforward networks. This is accomplished by striking a balance between sequential dependence and independence: although SRU's state computation is time-dependent, each state dimension is independent. The SRU also enhances deep recurrent model training by using highway connections and a parameter initialization technique designed for gradient propagation in deep architectures. SRU involves the following computation:

$$f_t = \sigma(W_f x_t + b_f) \tag{13}$$

$$r_t = \sigma(W_r x_t + b_r) \tag{14}$$

$$c_t = f_t \odot c_{t-1} + (1 - f_t) \odot (W x_r) \tag{15}$$

$$h_t = r_t \odot g(c_t) + (1 - r_t) \odot x_t \tag{16}$$

where W_f, W_r and b_f, b_r are learnable weight matrices and bias terms to be learned during training. In SRU, the forget and reset gates are computed independently and do not depend on each other, which simplifies the gating mechanism and allows for faster training. Additionally, the candidate hidden state is computed using element-wise multiplication of the reset gate with the previous hidden state, which enables SRU to capture long-term dependencies more effectively than traditional RNNs. The three models are denoted as CDNN-LSTM, CDNN-GRU, and CDNN-SRU, respectively.

3.2. Proposed model extension

The Convolutional Encoder–Decoder (CED) framework for speech enhancement shown in Fig. 3 is further extended and an attention mechanism is added to the skip connections. Attention gates (AGs) are introduced between the convolutional encoder and the decoder architecture to further increase the performance of the proposed models, shown in Fig. 5(A). The AGs in skips can effectively reduce redundant regions while emphasizing the important spectral features. Given the large number of frequency components in the spectra, the formant frequencies are typically dominating in low-frequency areas and high-frequency regions have a sparse distribution. Hence, it is important to differentiate distinct spectral locations with varying weights. Assume the inputs to AGs are k and l, where k and l represent a decoding

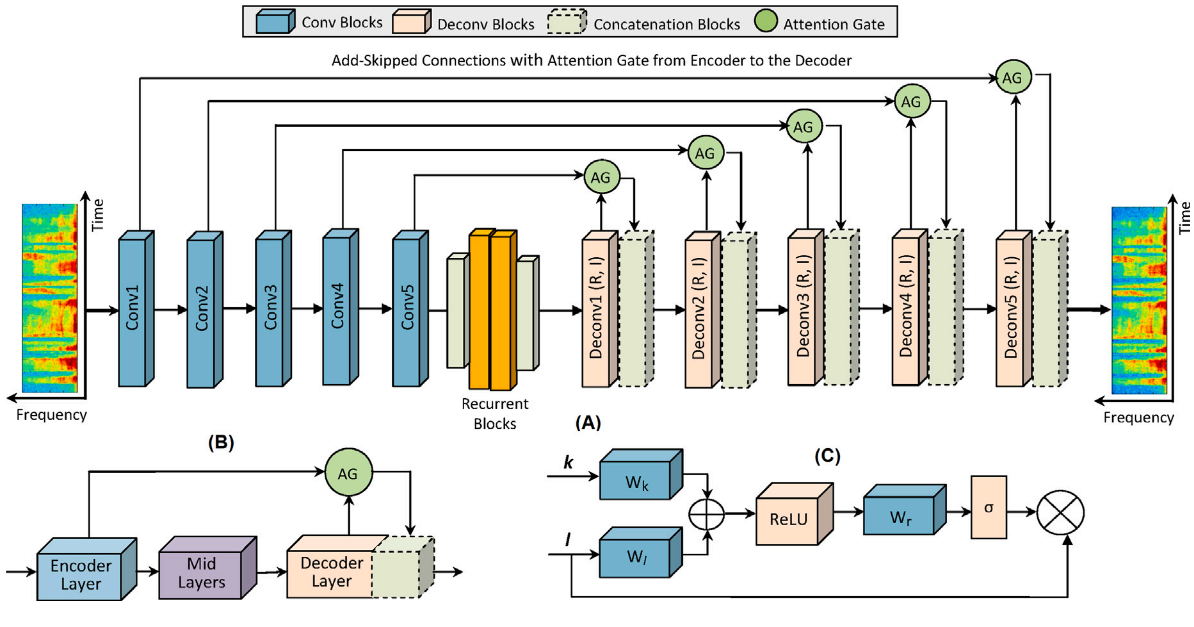

Fig. 5. The architecture of the extended proposed model.

layer feature and its equivalent encoding layer feature, respectively. The result may be computed as follows:

$$AG_{y_{out}} = l \odot \sigma(W_r \circledast \mathrm{ReLU}(W_k \circledast k + W_l \circledast l)) \tag{17}$$

where W_r, W_k, and W_l are the convolution kernels whereas $\odot$ represents element-wise multiplication and $\circledast$ represents the Hadamard product. It is important to note that the AG unit is divided into two branches, one of which combines the information from both inputs and creates the attention coefficients using a sigmoid function, and the other of which copies the information from l and multiplies the coefficients. As the subsequent decoding layer's input, the output of AG has combined with the features from the next decoding layer along the channel dimension. Fig. 5(C) demonstrates the structure of AG integrated between the encoder–decoder structure. The three extended versions are denoted as E-CDNN-LSTM, E-CDNN-GRU, and E-CDNN-SRU, respectively.

3.3. Network architecture

The proposed models comprise a convolutional encoder–decoder with LSTM/GRU/SRU bottlenecks, and are illustrated in Figs. 3 and 5. The real and imaginary spectrograms of noisy speech are treated as two independent input channels. Figs. 3 and 5 show shared encoder and recurrent modules (LSTM/GRU/SRU) whereas two decoders are used to estimate real and imaginary spectra, respectively. Such architecture is inspired by multi-task learning, in which numerous related prediction tasks are concurrently learned with information exchanged across tasks. The estimate of the imaginary component and real component are two related subtasks in complex mapping. As a result, parameter sharing is projected to have a regularization impact amongst the subtasks, perhaps leading to improved generalization. Furthermore, parameter sharing may promote learning, especially when two subtasks are closely connected. Excessive parameter sharing across subtasks, on the other hand, may hinder learning, particularly if the two subtasks are poorly connected. As a result, the right selection of parameter sharing becomes crucial for achieving optimal performance. Sharing the encoder and LSTM/GRU/SRU modules but not sharing the decoder module results in optimum performance. All signals are sampled at 16 kHz where a 20 ms Hamming window generates time frames series following 50% overlapping between subsequent frames. The input to the model is 161-dimensional spectra, corresponding to a 320-point (16 kHz × 20 ms) STFT.

Table 1
Architecture details.

Layer name	Input size	Output size	Hyperparameters
Conv2D-1	2 × T × 161	8 × T × 80	(1 × 3), (1,2), 8
Conv2D-2	8 × T × 80	16 × T × 39	(1 × 3), (1,2), 16
Conv2D-3	16 × T × 39	32 × T × 19	(1 × 3), (1,2), 32
Conv2D-4	32 × T × 19	64 × T × 9	(1 × 3), (1,2), 64
Conv2D-5	64 × T × 9	128 × T × 4	(1 × 3), (1,2), 128
Reshape-1	128 × T × 4	T × 512	–
LSTM/GRU/SRU-1	T × 512	T × 512	512
LSTM/GRU/SRU-2	T × 512	T × 512	512
Reshape-2	T × 512	128 × T × 4	–
Deconv2D-5 (×2)	256 × T × 4	64 × T × 9	(1 × 3), (1,2), 64
Deconv2D-4 (×2)	128 × T × 9	32 × T × 19	(1 × 3), (1,2), 32
Deconv2D-3 (×2)	64 × T × 19	16 × T × 39	(1 × 3), (1,2), 16
Deconv2D-2 (×2)	32 × T × 39	8 × T × 80	(1 × 3), (1,2), 8
Deconv2D-1 (×2)	16 × T × 80	1 × T × 161	(1 × 3), (1,2), 1
Linear (×2)	1 × T × 161	1 × T × 161	161
Concatenation	1 × T × 161 (×2)	2 × T × 161	–

Table 1 describes the architecture of the proposed network. The input and output sizes of each layer are specified in the (**FeatureMaps**× **TimeSteps** × **FrequencyChannels**) format. Furthermore, the layer hyperparameters are supplied in the format (Kernel Size, Strides, Out Channels). By using skip connections, the number of feature mappings in each decoder layer is doubled. Rather than using the kernel size of (2 × 3) corresponding time and frequency, we use (1 × 3), which we observed does not really decrease performance. Following each convolutional or deconvolutional block is a batch normalization operation and an exponential linear unit (ELU) activation function. To project the learned features to real or imaginary spectrograms, a linear layer is placed on top of each decoder. No future frames are included during the processing. This amounts to a causal processing strategy. Fig. 6 demonstrates the causal process where 11 frames create the context window used as input to the model. With causal processing, 10 previous and 1 current frame will create 1 present frame (t−1). After processing the (t−1) frame, the next frame will start from (t−10) for which the current frame will t, and so on.

Table 2
STOI scores in four example noise with three proposed SE models.

CDNN-LSTM																				
Noise	Babble noise					Exhibition noise					Restaurant noise					Street noise				
SNR	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB
Noisy	55.5	62.6	67.1	72.1	78.8	58.1	65.7	70.8	75.5	82.1	54.9	61.7	66.5	70.7	76.5	58.3	65.8	70.3	75.5	82.1
Enhanced	72.9	82.1	83.7	86.6	90	72.2	80.9	84.5	86.9	89.8	78.7	84.4	87	88.9	91.1	79	84.6	87.1	89.6	92.1
CDNN-GRU																				
Noise	Babble noise					Exhibition noise					Restaurant noise					Street noise				
SNR	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB
Noisy	55.5	62.6	67.1	72.1	78.8	58.1	65.7	70.8	75.5	82.1	54.9	61.7	66.5	70.7	76.5	58.3	65.8	70.3	75.5	82.1
Enhanced	73.1	79.5	82.7	85.1	87.8	80.6	84.5	86.3	87.7	88.7	75.91	81.6	84.6	86.8	89	76.5	82.1	85	87.1	89.7
CDNN-SRU																				
Noisy	Babble noise					Exhibition noise					Restaurant noise					Street noise				
SNR	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	OdB	2 dB	5 dB
Noisy	55.5	62.6	67.1	72.1	78.8	58.1	65.7	70.8	75.5	82.1	54.9	61.7	66.5	70.7	76.5	58.3	65.8	70.3	75.5	82.1
Enhanced	79.6	85	88.6	90	92.6	84.4	89	91.2	92.3	92.9	84.9	89.8	92.1	94.4	95.8	82.1	86.9	89.5	91.6	93.9

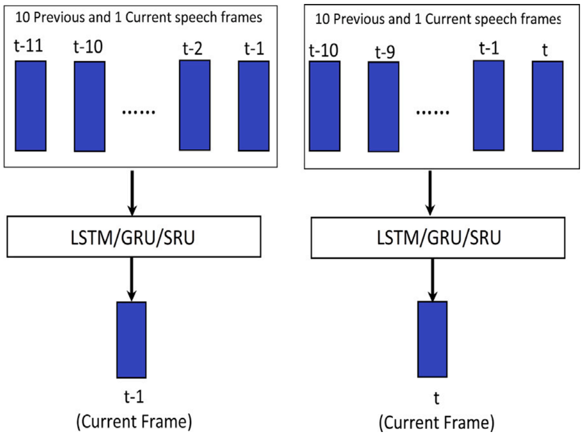

Fig. 6. Causal LSTM/GRU/SRU using a feature window of 11 frames.

4. Experiments and setup

4.1. Speech database

In the experiments, we evaluated SE networks on the Wall Street Journal (WSJ0-S184) dataset (Garofolo et al., 1993). The dataset comprises read material that is not dependent on specific speakers. This material is divided into different sets, including training, development test, and evaluation test sets. For the training of speech recognition algorithms, there are 90 utterances available from each of the 92 speakers. Additionally, for testing purposes, 48 additional speakers were chosen. They each read 40 sentence utterances, which only contain words from a fixed vocabulary of 5000 words out of a larger vocabulary of 64,000 words. These 40 sentences will serve as the testing material. Furthermore, all 140 speakers recorded a set of 18 adaptation sentences. To train the networks, we used a collection of 60 different noise types sourced from the Perception and Neurodynamics Laboratory and Laboratory for Recognition and Organization of Speech and Audio. For model testing, we specifically selected four challenging noise types, namely multi-talker babble, exhibition, street, and restaurant. To create a training set, we randomly extracted utterances from the 60 training noise types, employing a random cut approach. The selected utterances were sampled at various signal-to-noise ratios (SNRs) such as −5 dB, −2 dB, 0 dB, 2 dB, and 5 dB. During the testing, the same SNRs were used. To assess generalization, we subjected the models to two unseen noise types, namely factory2 and cafeteria. Both the speech utterances and noise types were sampled at a rate of 16 kHz.

4.2. Evaluation metrics

The experiments employ a pair of commonly utilized objective measurements to assess the proposed SE models. These metrics consist of the STOI (Short-Time Objective Intelligibility) and the PESQ (Perceptual Evaluation of Speech Quality). The STOI determines the intelligibility of the enhanced speech signals, while the PESQ evaluates their quality. PESQ, referenced as ITU-T P.862 recommendation (Rix et al., 2001), assigns perceptual speech quality scores between −0.5 and 4.5. Similarly, STOI (Taal et al., 2010) quantifies speech intelligibility using a scale of 0 to 100 for its output values.

5. Results and discussions

This section discusses the results of this study. We examined the proposed speech enhancement models objectively, as indicated in the following subsections.

5.1. Speech enhancement in seen noises

Tables 2–3 shows the STOI and PESQ scores with the proposed SE models for four example noises and five SNR levels. All three SE models enhanced the noisy speech and obtained better speech intelligibility and quality. Table 4 provides the average scores across all noise types and compares the proposed SE methods for the seen noises using STOI and PESQ. It is observed that better intelligibility and quality are achieved when the proposed SE models are applied to noisy speech. The three proposed SE models (CDNN-LSTM, CDNN-GRU, and CDNN-SRU) improved the noisy speech. For example, CDNN-LSTM improved the STOI score (by 19%) and the PESQ score (by 43.18%) over the unprocessed noisy speech at −5 dB SNR. Similarly, with CDNN-GRU, notable improvements in STOI score (by 19.83%) and PESQ score (by 40.47%) are observed over the noisy speech at −5 dB SNR. CDNN-SRU performed better and improved the STOI score by 20.05% and the PESQ score by 42.96% over noisy speech at −5 dB SNR, respectively. For all SNR levels and noise types, the CDNN-LSTM improved the overall average STOI score by 16.08% and the PESQ score by 1.20 (40.95%). The overall average improvement in STOI scores over the noisy speech by the other two SE models (CDNN-GRU and CDNN-SRU) is 15.19% and 20.08%, respectively. Further, the overall average improvement in PESQ scores over the noisy speech for CDNN-GRU and CDNN-SRU is 1.16 (40.20%) and 1.26 (42.14%), respectively. The highest average STOI score (89.33%) is obtained by CDNN-SRU. Similarly, the highest average PESQ score (2.99) is again obtained by CDNN-SRU. The average scores for CDNN-LSTM are STOI (84.61%) and PESQ (2.93) whereas the average scores for CDNN-GRU are: STOI (83.72%) and PESQ (2.89).

Table 3
PESQ scores in four example noise with three proposed SE models.

CDNN-LSTM																				
Noise	Babble noise					Exhibition noise					Restaurant noise					Street noise				
SNR	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	OdB	2 dB	5 dB
Noisy	1.46	1.61	1.71	1.84	2.04	1.56	1.63	1.72	1.83	2.14	1.39	1.49	1.58	1.73	1.87	1.59	1.64	1.72	1.86	2.13
Enhanced	2.63	2.76	2.86	2.96	3.03	2.57	2.80	2.82	2.94	2.99	2.80	3.13	3.16	3.25	3.35	2.58	2.87	2.94	3.00	3.09
CDNN-GRU																				
Noise	Babble noise					Exhibition noise					Restaurant noise					Street noise				
SNR	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB
Noisy	1.46	1.61	1.71	1.84	2.04	1.56	1.63	1.72	1.83	2.14	1.39	1.49	1.58	1.73	1.87	1.59	1.64	1.72	1.86	2.13
Enhanced	2.46	2.65	2.72	2.82	2.89	2.69	3.01	3.16	3.30	3.40	2.63	2.90	3.01	3.15	3.26	2.30	2.69	2.81	2.92	3.01
CDNN-SRU																				
Noisy	Babble noise					Exhibition noise					Restaurant noise					Street noise				
SNR	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB
Noisy	1.46	1.61	1.71	1.84	2.04	1.56	1.63	1.72	1.83	2.14	1.39	1.49	1.58	1.73	1.87	1.59	1.64	1.72	1.86	2.13
Enhanced	2.43	2.61	2.70	2.78	2.82	2.64	2.99	3.13	3.24	3.29	3.05	3.34	3.54	3.69	3.81	2.39	2.70	2.80	2.87	2.93

Table 4
STOI and PESQ scores in seen noises.

Models	CDNN-LSTM					CDNN-GRU					CDNN-SRU				
STOI (in%) in seen background noises															
SNRs	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB	−5 dB	−2 dB	0 dB	2 dB	5 dB
Noisy	56.70	63.95	68.68	73.45	79.88	56.70	63.95	68.68	73.45	79.88	56.70	63.95	68.68	73.45	79.88
Enhanced	75.70	83.00	85.58	88.00	90.75	76.53	81.93	84.65	86.68	88.80	82.75	87.68	90.35	92.08	93.80
PESQ in seen background noises															
Noisy	1.50	1.59	1.68	1.82	2.05	1.50	1.59	1.68	1.82	2.05	1.50	1.59	1.68	1.82	2.05
Enhanced	2.64	2.89	2.95	3.04	3.12	2.52	2.81	2.93	3.05	3.14	2.63	2.91	3.04	3.14	3.21

Table 5
PESQ and STOI score in unseen background noise.

Models	CDNN-LSTM				CDNN-GRU				CDNN-SRU			
PESQ and STOI in cafeteria unseen background noise												
Metric	PESQ		STOI		PESQ		STOI		PESQ		STOI	
SNR	−5 dB	−2 dB	−5 dB	−2 dB	−5 dB	−2 dB	−5 dB	−2 dB	−5 dB	−2 dB	−5 dB	−2 dB
Noisy	1.30	1.35	54.9	61.7	1.30	1.35	54.9	61.7	1.30	1.35	55.9	61.7
Enhanced	2.22	2.42	71.4	75.3	2.11	2.34	70.1	76.1	2.26	2.48	72.4	78.1
PESQ and STOI in Factory2 unseen background noise												
Noisy	1.48	1.58	58.3	65.7	1.48	1.58	58.3	65.7	1.48	1.58	58.3	65.7
Enhanced	2.24	2.43	71.8	76.3	2.19	2.39	70.7	77.2	2.29	2.51	72.5	78.8

5.2. Speech enhancement in unseen noises

Table 5 compares the proposed SE models for unseen noises at two SNRs (−5 dB and −2 dB). The noise types are not included in the model training. When applying the proposed SE models, it is noted that better intelligibility (STOI) and quality (PESQ) scores are achieved as compared to noisy speech in unseen noisy conditions. Factory 2 and the cafeteria are the two unseen noises. In cafeteria noise, the STOI at −5 dB is improved from 54.9% (noisy speech) to 71.4% with CDNN-LSTM, 54.9% (noisy speech) to 70.1% with CDNN-GRU, whereas STOI with CDNN-SRU is improved from 54.9% (noisy speech) to 72.4%, respectively. Similarly, the PESQ at −2 dB is improved from 1.35 to 2.42 with CDNN-LSTM, from 1.35 to 2.34 with CDNN-GRU, and from 1.35 to 2.48 with CDNN-SRU, respectively. Further, in factory2 noise, the STOI at −5 dB is improved from 58.3% (noisy speech) to 71.8% with CDNN-LSTM, 58.3% (noisy speech) to 70.7% with CDNN-GRU, whereas STOI with CDNN-SRU is improved from 58.3% (noisy speech) to 72.5%, respectively. Also, in factory2 noise, the PESQ at −5 dB is improved from 1.48 (noisy speech) to 2.24 (33.92% improvement over noisy speech) with CDNN-LSTM, from 1.35 (noisy speech) to 2.19 (32.42% improvement) with CDNN-GRU, whereas the PESQ with CDNN-SRU is improved from 1.48 (noisy speech) to 2.29 (35.37% improvement). The CDNN-SRU achieved the highest overall improvements (both in STOI and PESQ).

5.3. Evaluation of extended model

To examine the extended version of the proposed model, the same datasets and metrics are employed in experiments. To train the models, we used 20 noises from the Perception and Neurodynamics Laboratory and the Laboratory for Recognition and Organization of Speech and Audio. To test the models, we used three challenging noise types (multi-talker babble, street, and restaurant). The training set is created by randomly selecting utterances with an indiscriminate cut from the 20 training noise types at −5 dB and 0 dB SNRs. In experiments, the cafeteria noise is used as an unseen noise type. STOI and PESQ are used to examine the extended versions of the proposed model.

Table 6 shows the PESQ and STOI scores with three extended proposed SE models across two SNR levels in seen noisy conditions. The extension of three proposed SE models obtained better speech intelligibility and quality as compared to the previous version after adding attention gates to the skip connections. For a demonstration at −5 dB SNR, the STOI scores with E-CDNN-LSTM (76.60%), E-CDNN-GRU (76.90%), and E-CDNN-SRU (83.40%) are improved by 21.7%, 22.0%, and 28.5% over the noisy speech. Further, PESQ scores with E-CDNN-LSTM (2.68), E-CDNN-GRU (2.57), and E-CDNN-SRU (2.72) are improved with factors of 1.18 (44.02%), 1.07 (41.63%), and 1.22 (44.85%) over the noisy speech at −5 dB SNR. The proposed E-CDNN-SRU outscored its two counterparts at low SNRs. The improvements in

Table 6
Average PESQ and STOI scores in seen noisy conditions (Extended model).

Models	E-CDNN-LSTM				E-CDNN-GRU				E-CDNN-SRU			
Metric	PESQ		STOI		PESQ		STOI		PESQ		STOI	
SNR	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB
Noisy	1.50	1.68	56.70	68.68	1.50	1.68	56.70	68.68	1.50	1.68	56.70	68.68
Enhanced	2.68	3.02	76.60	86.10	2.57	2.97	76.90	85.40	2.72	3.12	83.40	91.20

Table 7
Average PESQ and STOI scores in cafeteria unseen noisy condition (Extended model).

Models	E-CDNN-LSTM				E-CDNN-GRU				E-CDNN-SRU			
Metric	PESQ		STOI		PESQ		STOI		PESQ		STOI	
SNR	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB
Noisy	1.30	1.61	54.9	64.9	1.30	1.61	54.9	64.9	1.30	1.61	54.9	64.9
Enhanced	2.49	2.76	76.60	82.11	2.31	2.70	73.90	81.40	2.52	2.81	77.40	86.20

Table 8
Parameter efficiency.

Model	Para#	MACs	Param size
CDNN-LSTM	2.488 M	0.658 G/s	9.95 MB
CDNN-GRU	1.961 M	0.600 G/s	7.85 MB
CDNN-SRU	1.173 M	0.427 G/s	4.69 MB
E-CDNN-LSTM	2.533 M	0.705 G/s	10.13 MB
E-CDNN-GRU	2.007 M	0.648 G/s	8.03 MB
E-CDNN-SRU	1.218 M	0.475 G/s	4.88 MB

the STOI and PESQ results indicate the success of attention gates in the skip connections.

Table 7 shows the PESQ and STOI scores with three proposed SE models across two SNRs in a cafeteria unseen noisy condition. The proposed models perform better in terms of speech intelligibility and quality in unseen noise at challenging SNR (−5 dB). The STOI scores with E-CDNN-LSTM (76.60%), E-CDNN-GRU (73.90%), and E-CDNN-SRU (77.40%) are improved by 21.7%, 19%, and 22.5% over the noisy speech at −5 dB SNR. In addition, the PESQ scores with E-CDNN-LSTM (2.49), E-CDNN-GRU (2.31), and E-CDNN-SRU (2.52) are improved by factors of 1.19 (47.79%), 1.01 (43.72%), and 1.22 (48.41%) over the noisy speech at −5 dB SNR. The proposed E-CDNN with SRU module outscored the two counterparts and the noisy speech by significant margins. The addition of an attention gate between the encoder–decoder structure significantly increased speech intelligibility and quality. The AGs in E-CDNN-SRU improved the average perceived quality of the noisy speech (at −5 dB and 0 dB) by 1.21 (45.31%) and the intelligibility is improved by 21.90%, respectively. Further, the AGs in E-CDNN-GRU improved the perceived quality of the noisy speech by 1.05 (41.83%) whereas, the intelligibility is improved by 17.75% over noisy speech. Finally, the addition of AGs in E-CDNN-LSTM increased the STOI and PESQ by 1.17 (44.44%) and 19.45% over noisy speech, respectively.

5.4. Computational complexity and parameter efficiency

Since computational resources are often constrained in real-world applications, it is necessary to establish an appropriate trade-off between the model's enhancement performance and parameter efficiency. Table 8 shows the parameters efficiency of the proposed speech enhancement models. The parameters efficiency of the proposed SE models shows that integrating the SRU into the proposed SE framework (without attention gates in skips) significantly reduced the parameters count (1.173M) and parameter size (4.69 MB) as compared to the LSTM (2.488M, 9.95 MB), and GRU (1.961M, 7.85 MB). After adding attention gates into the skips, the results show a marginal increase in parameter count, that is, 2.5338M with LSTM, 2.007M with GRU, and 1.218M with SRU, respectively. Recent DNN research for SE attempts to increase model performance on hardware accelerators that execute these SE models rapidly and efficiently. In order to use the suggested SE models in mobile or embedded systems, hardware memory consumption must be reduced. Therefore, we provide an overview of the multiply–accumulate operations (MACs). It is clear that the suggested model with SRU has the lowest MACs (0.427 G/s without attention gates and 0.475 G/s with attention gates) efficiently without sacrificing the SE performance. The proposed approach with SRU integration has greatly reduced the parameter size and MACs.

Furthermore, in our proposed model, we assessed the Real-Time Factor (RTF). The RTF quantifies the processing time to input audio data duration ratio, serving as a critical metric for real-time applications (Ivanov et al., 2016). We conducted this assessment using a single core of the Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz processor for measurement, resulting in a remarkable RTF score of 0.18.

5.5. Comparison with related models

The proposed models for SE are measured against multiple recent baseline studies. We examined the proposed and baseline models with three noise types (multi-talker babble, street, and cafeteria) at +5 dB SNR. Table 9 shows the SE model comparisons over three test noises at 5 dB SNR in terms of STOI and PESQ where the symbol "↑" indicates the improvements in PESQ and STOI. The results indicate that the proposed SE models outperformed the benchmarks with reasonable margins. For demonstration, CDNN-LSTM improved the STOI (by 2.76%) and PESQ score (by 13.44%) over the DDAEC (Pandey and Wang, 2020). Similarly, with CDNN-GRU, notable improvements in STOI (by 0.77%) and PESQ score (by 22.95%) are observed over DEMUCAS (Défossez et al., 2020). Further, CDNN-SRU outperformed and improved the STOI score by 10.62% and the PESQ score by 40% over GCRN (Tan and Wang, 2019), respectively. Furthermore, E-CDNN-LSTM improved the STOI (by 8.32%) and PESQ score (by 34.64%) over DCCRN (Hu et al., 2020). Similarly, with E-CDNN-GRU, considerable improvements in STOI (by 7.13%) and PESQ score (by 20.16%) are observed over PHASEN (Yin et al., 2020). In addition, E-CDNN-SRU improved the STOI score by 12.01% and PESQ by 33.87% over FullSubNet (Chen et al., 2022). Finally, the E-CDNN-SRU improved the PESQ score by 11.71%, 11.71%, and 10.15% and whereas STOI by 7.95%, 7.23%, and 6.73% over CTSNet (Li et al., 2021), GaGNet (Li et al., 2022a), and MDNet (Li et al., 2022b), respectively.

The proposed CDNN, while sharing some similarities with architectures DCCRN (Hu et al., 2020), incorporates several key architectural differences to achieve better SE performance. While using complex-valued DNN with complex-valued arithmetic, DCCRN shows higher parameter (3.67M) counts and MACs (11.13 G/s) as compared to the proposed CDNN, provided in Table 9 for reference. Secondly, DCCRN uses simple skip connections whereas the proposed CDNN enhances the skip connections with attention mechanisms to reduce redundancy and emphasize crucial spectral features in complex-domain speech enhancement. Thirdly, While DCCRN primarily relies on LSTM layers in the

Table 9
Comparison with related models.

Models	Year	Para# [M]	MACs [G/s]	PESQ	STOI	↑PESQ	↑STOI
Noisy speech	–	–	–	1.73	68.25	0.00	0.00
DDAEC (Défossez et al., 2020)	2020	4.82	36.56	2.76	82.84	1.03	14.59
DEMUCAS (Li et al., 2021)	2020	18.87	4.35	2.67	84.23	0.94	15.98
GCRN (Tan and Wang, 2019)	2020	9.77	2.42	2.48	78.68	0.75	10.43
DCCRN (Zhao et al., 2022)	2020	3.67	11.13	2.54	78.58	0.81	10.33
PHASEN (Yin et al., 2020)	2020	8.76	6.12	2.73	79.77	0.99	11.52
FullSubNet (Luo and Mesgarani, 2019)	2022	5.64	31.35	2.55	73.89	0.82	05.64
CTSNet (Li et al., 2022b)	2021	4.35	5.57	2.86	82.15	1.13	13.90
GaGNet (Li et al., 2022a)	2022	5.94	1.63	2.86	82.87	1.13	14.62
MDNet (Fu et al., 2021)	2022	8.36	2.70	2.88	83.37	1.15	15.12
CDNN-LSTM (Proposed)	2023	2.49	0.66	2.92	84.65	1.19	16.41
CDNN-GRU (Proposed)	2023	1.96	0.60	2.88	83.71	1.15	15.47
CDNN-SRU (Proposed)	2023	1.17	0.43	2.98	89.33	1.25	21.09
E-CDNN-LSTM (Proposed)	2023	2.53	0.71	2.97	86.30	1.24	18.06
E-CDNN-GRU (Proposed)	2023	2.01	0.65	2.91	85.10	1.18	16.86
E-CDNN-SRU (Proposed)	2023	1.22	0.48	3.01	90.10	1.28	21.86

Table 10
The performance evaluation conducted on the VoiceBank+DEMAND database. The symbol "–" indicates that the original paper does not provide the corresponding result.

Models	Year	Para#	PESQ	STOI	SSNR
Noisy	–	–	1.97	91.6	1.69
SEGAN (Nikzad et al., 2020)	2017	97.5 M	2.16	93.1	7.66
DCCRN (Zhao et al., 2022)	2019	3.70 M	2.68	93.7	8.62
GAGNet (Li et al., 2022a)	2021	5.94 M	2.94	94.7	9.24
RDL-Net (Defossez et al., 2020)	2020	3.91 M	3.02	93.8	–
DEMUCS (Wang et al., 2021a)	2020	128.0 M	3.07	95.1	8.53
TSTNN (Kim and Seo, 2021)	2021	0.92 M	2.96	95.1	9.72
SE-Conformer (Ivanov et al., 2016)	2021	–	3.13	95.1	–
CDNN-SRU	2023	1.17 M	3.21	95.7	10.28
E-CDNN-SRU	2023	1.22 M	3.25	96.1	10.35

bottleneck, this study has investigated and compared multiple recurrent networks, such as LSTM, GRU, and SRU. The experiments show that the SRU-based model offers both lower computational complexity and superior performance compared to LSTM and GRU, as given in Table 9. SRU achieves this efficiency by simplifying gating mechanisms, relying on matrix multiplications and element-wise operations while effectively capturing temporal dependencies. This makes SRU an efficient and effective choice for speech enhancement.

5.6. Performance comparison using VoiceBank+DEMAND

In addition to the WSJ0-SI84 database, this study performs experiments on another publicly available benchmark database called VoiceBank+DEMAND. This set of experiments is performed to further confirm the effectiveness of the proposed SE approaches compared to other state-of-the-art benchmarks. The results obtained from these experiments are provided in Table 10. The following observations are made after analyzing the results (in terms of PESQ, STOI, and Segmental SNR (SSNR)) in Table 10. The higher values of these measures indicate better performance. Since the encoder–decoder with SRU performs better between counterparts, this set of experiments provides results for CDNN-SRU and E-CDNN-SRU in Table 10. The proposed CDNN-SRU and E-CDNN-SRU perform better on VoiceBank+DEMAND and obtain competitive results in terms of PESQ, STOI, SSNR, and parameter count. From GAGNet (Li et al., 2022a), the proposed CDNN-SRU and E-CDNN-SRU improve the metrics by 0.27 and 0.31 (PESQ), 1.0% and 1.4% STOI whereas 1.04 dB and 1.11 dB (SSNR), respectively. Further, from DCCRN (Hu et al., 2020), the proposed CDNN-SRU and E-CDNN-SRU improve the metrics by 0.53 and 0.57 (PESQ), 2.0% and 2.4% STOI whereas 1.66 dB and 1.73 dB (SSNR), respectively. The parameter count of TSTNN (Wang et al., 2021a) is better (0.92M) but the PESQ (2.96), STOI (95.1), and SSNR (9.72 dB) are less competitive as compared to the proposed SE models. The proposed SE models obtain better STOI, PESQ, and SSNR results with fewer parameter counts (1.17M and 1.22M).

6. Summary and conclusions

This study develops a convolutional encoder–decoder framework and deeply examines the inclusion of three different recurrent models including Long-Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Simple Recurrent Unit (SRU) in terms of speech intelligibility, speech quality, and model complexity. The proposed speech enhancement models demonstrate improvements in speech intelligibility, quality, and computational efficiency with reduced model size (in MBs). The proposed models are extended by integrating attention gates into the skip connections, enabling the selective emphasis of important spectral components through assigned attention weights. This integration leads to substantial performance improvements in the already high-performing models. The incorporation of attention gates and multitask learning in complex mapping-based models not only enhances model performance but also exhibits promising potential for improved generalization capabilities. To ensure suitability for real-time processing on resource-limited devices, the models' performance is evaluated based on metrics such as trainable parameter count (in millions), parameter size (in megabytes), and multiply–accumulate operations (MACs). The parameters efficiency of the proposed model concludes that integrating the SRU as a bottleneck into the encoder–decoder framework significantly reduced the parameters count (0.44M without attention gate and 0.45M with attention gate) number, parameter size (1.78 Mb without attention gate and 1.83 Mb with attention gate), and MACs (0.081 G/s and 0.092 G/s) respectively without deteriorating the speech quality and intelligibility. Additionally, the Real-Time Factor (RTF) achieved was 0.18. Based on the STOI and PESQ assessments, the analysis concludes that the proposed models exhibit substantial enhancements in both quality and intelligibility in both seen and unseen noisy scenarios when compared to the benchmark SE models. The proposed SE models obtain better STOI (95.7% and 96.1%), PESQ (3.21 and 3.25), and SSNR (10.28 dB and 10.35 dB) results on the VoiceBank+DEMAND dataset with fewer parameter counts (1.17M and 1.22M).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (no. 61671418).

References

Bao, F., Abdulla, W.H., 2018. A new ratio mask representation for CASA-based speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (1), 7–19.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27 (2), 113–120.

Chen, J., Wang, D., 2017. Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. 141 (6), 4705–4714.

Chen, J., Wang, Z., Tuo, D., Wu, Z., Kang, S., Meng, H., 2022. FullSubNet+: Channel attention fullsubnet with complex spectrograms for speech enhancement. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7857–7861.

Choi, H.S., Kim, J.H., Huh, J., Kim, A., Ha, J.W., Lee, K., 2018. Phase-aware speech enhancement with deep complex u-net. In: International Conference on Learning Representations.

Cui, X., Chen, Z., Yin, F., 2020. Speech enhancement based on simple recurrent unit network. Appl. Acoust. 157, 107019.

Défossez, A., Synnaeve, G., Adi, Y., 2020. Real time speech enhancement in the waveform domain. In: Proc. Interspeech 2020. pp. 3291–3295.

Defossez, A., Synnaeve, G., Adi, Y., 2020. Real time speech enhancement in the waveform domain. arXiv preprint arXiv:2006.12847.

Ding, H., Soon, Y., Koh, S.N., Yeo, C.K., 2009. A spectral filtering method based on hybrid wiener filters for speech enhancement. Speech Commun. 51 (3), 259–267.

Fu, S.W., Liao, C.F., Tsao, Y., 2019. Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality. IEEE Signal Process. Lett. 27, 26–30.

Fu, S.W., Tsao, Y., Lu, X., 2016. SNR-aware convolutional neural network modeling for speech enhancement. In: Interspeech. pp. 3768–3772.

Fu, S.W., Yu, C., Hsieh, T.A., Plantinga, P., Ravanelli, M., Lu, X., Tsao, Y., 2021. Metricgan+: An improved version of metricgan for speech enhancement. arXiv preprint arXiv:2104.03538.

Gao, T., Du, J., Dai, L.R., Lee, C.H., 2018. Densely connected progressive learning for lstm-based speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5054–5058.

Garofolo, J., Graff, D., Paul, D., Pallett, D., 1993. Csr-I (Wsj0) Complete Ldc93s6a. Web Download. Linguistic Data Consortium, Philadelphia, p. 83.

Grais, E.M., Ward, D., Plumbley, M.D., 2018. Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders. In: 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, pp. 1577–1581.

Hasannezhad, M., Ouyang, Z., Zhu, W.P., Champagne, B., 2020. An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp. 764–768.

Hasannezhad, M., Yu, H., Zhu, W.P., Champagne, B., 2022. PACDNN: A phase-aware composite deep neural network for speech enhancement. Speech Commun. 136, 1–13.

Hsieh, T.A., Wang, H.M., Lu, X., Tsao, Y., 2020. Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement. IEEE Signal Process. Lett. 27, 2149–2153.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., …, Xie, L., 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement.

Ivanov, A.V., Lange, P.L., Suendermann-Oeft, D., Ramanarayanan, V., Qian, Y., Yu, Z., Tao, J., 2016. Speed vs. accuracy: Designing an optimal asr system for spontaneous non-native speech in a real-time application. In: Proc. of the IWSDS, Saariselk, Finland.

Kim, E., Seo, H., 2021. SE-conformer: Time-domain speech enhancement using conformer. In: Interspeech. pp. 2736–2740.

Krawczyk, M., Gerkmann, T., 2014. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1931–1940.

Lai, Y.H., Zheng, W.Z., 2019. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. Biomed. Signal Process. Control 48, 35–45.

Lee, J., Skoglund, J., Shabestary, T., Kang, H.G., 2018. Phase-sensitive joint learning algorithms for deep learning-based speech enhancement. IEEE Signal Process. Lett. 25 (8), 1276–1280.

Li, A., Liu, W., Zheng, C., Fan, C., Li, X., 2021. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 1829–1843.

Li, A., Zheng, C., Zhang, L., Li, X., 2022a. Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Appl. Acoust. 187, 108499.

Li, A., Zheng, C., Zhang, Z., Li, X., 2022b. MDNet: Learning monaural speech enhancement from deep prior gradient. arXiv e-prints, arXiv-2203.

Liang, R., Kong, F., Xie, Y., Tang, G., Cheng, J., 2020. Real-time speech enhancement algorithm based on attention LSTM. IEEE Access 8, 48464–48476.

Liu, M., Wang, Y., Wang, J., Wang, J., Xie, X., 2018. Speech enhancement method based on LSTM neural network for speech recognition. In: 2018 14th IEEE International Conference on Signal Processing (ICSP). IEEE, pp. 245–249.

Luo, Y., Mesgarani, N., 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (8), 1256–1266.

Mowlaee, P., Kulmer, J., 2015. Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (9), 1521–1532.

Mowlaee, P., Stahl, J., Kulmer, J., 2017. Iterative joint MAP single-channel speech enhancement given non-uniform phase prior. Speech Commun. 86, 85–96.

Naithani, G., Barker, T., Parascandolo, G., Bramsl, L., Pontoppidan, N.H., Virtanen, T., 2017. Low latency sound source separation using convolutional recurrent neural networks. In: 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp. 71–75.

Nikzad, M., Nicolson, A., Gao, Y., Zhou, J., Paliwal, K.K., Shang, F., 2020. Deep residual-dense lattice network for speech enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 8552–8559, (05).

Ouyang, Z., Yu, H., Zhu, W.P., Champagne, B., 2019. 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). IEEE, pp. 5756–5760.

Paliwal, K., Wójcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53 (4), 465–494.

Pandey, A., Wang, D., 2019. A new framework for CNN-based speech enhancement in the time domain. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (7), 1179–1188.

Pandey, A., Wang, D., 2020. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6629–6633.

Pandey, A., Wang, D., 2021. Dense CNN with self-attention for time-domain speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 1270–1279.

Park, S.R., Lee, J.W., 2017. A fully convolutional neural network for speech enhancement. Evaluation 10 (5).

Qiu, Y., Wang, R., Hou, F., Singh, S., Ma, Z., Jia, X., 2022. Adversarial multi-task learning with inverse mapping for speech enhancement. Appl. Soft Comput. 120, 108568.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Vol. 2. IEEE, pp. 749–752.

Saleem, N., Gao, J., Khattak, M.I., Rauf, H.T., Kadry, S., Shafi, M., 2022. DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition. Knowl.-Based Syst. 238, 107914.

Saleem, N., Khattak, M.I., 2020. Deep neural networks for speech enhancement in complex-noisy environments. Int. J. Interact. Multimedia Artif. Intell. 6 (1), 84–91.

Saleem, N., Khattak, M.I., Al-Hasan, M.A., Jan, A., 2021. Multi-objective long-short term memory recurrent neural networks for speech enhancement. J. Ambient Intell. Humaniz. Comput. 12 (10), 9037–9052.

Saleem, N., Khattak, M.I., Al-Hasan, M., Qazi, A.B., 2020. On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks. IEEE Access 8, 160581-160595.

Shi fas, M.P., Claudio, S., Tsiaras, V., Stylianou, Y., 2020. A fully recurrent feature extraction for single channel speech enhancement. arXiv preprint arXiv:2006.05233.

Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T., 2020a. Fully convolutional recurrent networks for speech enhancement. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6674–6678.

Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T., 2020b. Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. EURASIP J. Adv. Signal Process. 2020 (1), 1–26.

Sun, L., Du, J., Dai, L.R., Lee, C.H., 2017. Multiple-target deep learning for LSTM-RNN based speech enhancement. In: 2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA). IEEE, pp. 136–140.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 4214–4217.

Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., Shikano, K., 2009. Blind spatial subtraction array for speech enhancement in noisy environment. IEEE Trans. Audio Speech Lang. Process. 17 (4), 650–664.

Takeuchi, D., Yatabe, K., Koizumi, Y., Oikawa, Y., Harada, N., 2020. Real-time speech enhancement using equilibriated RNN. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 851–855.

Tan, K., Chen, J., Wang, D., 2018a. Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (1), 189–198.

Tan, K., Chen, J., Wang, D., 2018b. Gated residual networks with dilated convolutions for supervised speech separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 21–25.

Tan, K., Wang, D., 2018. A convolutional recurrent neural network for real-time speech enhancement. In: Interspeech, Vol. 2018. pp. 3229–3233.

Tan, K., Wang, D., 2019. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 380–390.

Wang, X., Bao, C., 2019. Mask estimation incorporating phase-sensitive information for speech enhancement. Appl. Acoust. 156, 101–112.

Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S., 2023. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5.

Wang, Q., Du, J., Dai, L.R., Lee, C.H., 2018. A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (7), 1185–1197.

Wang, K., He, B., Zhu, W.P., 2021a. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7098–7102.

Wang, D., Lim, J., 1982. The unimportance of phase in speech enhancement. IEEE Trans. Acoust. Speech Signal Process. 30 (4), 679–681.

Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1849–1858.

Wang, Z., Zhang, T., Shao, Y., Ding, B., 2021b. LSTM-convolutional-BLSTM encoder–decoder network for minimum mean-square error approach to speech enhancement. Appl. Acoust. 172, 107647.

Williamson, D.S., Wang, Y., Wang, D., 2016. Complex ratio masking for joint enhancement of magnitude and phase. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5220–5224.

Wu, H., Tan, K., Xu, B., Kumar, A., Wong, D., 2023. Rethinking complex-valued deep neural networks for monaural speech enhancement. arXiv preprint arXiv:2301.04320.

Xia, B., Bao, C., 2014. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun. 60, 13–29.

Xia, Y., Wang, J., 2015. Low-dimensional recurrent neural network-based Kalman filter for speech enhancement. Neural Netw. 67, 131–139.

Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2013. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21 (1), 65–68.

Yamanaka, J., Kuwashima, S., Kurita, T., 2017. Fast and accurate image super resolution by deep CNN with skip connection and network in network. In: International Conference on Neural Information Processing. Springer, Cham, pp. 217–225.

Yang, Y., Zhang, H., Cai, Z., Shi, Y., Li, M., Zhang, D., ., Wang, J., 2023. Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion. Biomed. Signal Process. Control 80, 104279.

Yin, D., Luo, C., Xiong, Z., Zeng, W., 2020. Phasen: A phase-and-harmonics-aware speech enhancement network. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 9458–9465, (05).

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Zhao, S., Ma, B., Watcharasupat, K.N., Gan, W.S., 2022. FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 9281–9285.

Zheng, N., Zhang, X.L., 2018. Phase-aware speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (1), 63–76.