



OPEN End-to-end feature fusion for jointly optimized speech enhancement and automatic speech recognition

Mohamed Medani¹, Nasir Saleem², Fethi Fkih³✉, Manal Abdullah Alohal⁴, Hela Elmannai⁵ & Sami Bourouis⁶

Speech enhancement (SE) and automatic speech recognition (ASR) in real-time processing involve improving the quality and intelligibility of speech signals on the fly, ensuring accurate transcription as the speech unfolds. SE eliminates unwanted background noise from target speech in environments with high background noise levels, which is crucial in real-time ASR. This study first proposes a speech enhancement network based on an attentional-codec model. Its primary objective is to suppress noise in the target speech with minimal distortion. However, excessive noise suppression in the enhanced speech can potentially diminish the effectiveness of downstream ASR systems by excluding crucial latent information. While joint SE and ASR techniques have shown promise for achieving robust end-to-end ASR, they traditionally rely on using the enhanced features as inputs to the ASR systems. To address this limitation, our study uses a dynamic fusion approach. This approach integrates both the enhanced features and the raw noisy features, aiming to eliminate noise signals from the enhanced target speech while simultaneously learning fine details from the noisy signals. This fusion approach seeks to mitigate speech distortions, enhancing the overall performance of the ASR system. The proposed model consists of an attentional codec equipped with a causal attention mechanism for SE, a GRU-based fusion network, and an ASR system. The SE network uses a modified Gated Recurrent Unit (GRU), where the traditional hyperbolic tangent (*tanh*) is replaced by an attention-based rectified linear unit (AReLU). The SE experiments consistently obtain better speech quality, intelligibility, and noise suppression in matched and unmatched conditions than the baselines. With the LibriSpeech database, the proposed SE obtains better STOI (19.81%) and PESQ (28.97%) in matched conditions and unmatched conditions (STOI: 17.27% and PESQ: 27.51%). The joint training framework for robust end-to-end ASR evaluates the character error rate (CER). The ASR results find that the joint training framework reduces the error rate from 32.99% (average noisy signals) to 13.52% (with the proposed SE and joint training for ASR).

Keywords Speech enhancement, Speech recognition, Deep learning, End-to-end processing, Attentional GRU, Feature fusion, Joint optimization

Joint speech enhancement (SE) and speech recognition are vital for improving the accuracy and robustness of automatic speech recognition systems. By removing background noise and enhancing speech quality, the SE techniques enable ASR systems to better understand and transcribe spoken words, especially in noisy environments like crowded rooms or outdoor settings. This advancement is essential for the practical implementation of ASR in everyday applications, such as mobile devices and virtual assistants, ensuring reliable and effective communication with these technologies. Background noises frequently contaminate speech

¹Applied College of Muhayel Aseer, King Khalid University, Abha 62529, Saudi Arabia. ²Department of Electrical Engineering, FET, Gomal University, D.I. Khan 29050, Pakistan. ³Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia. ⁴Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. ⁵Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. ⁶Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia. ✉email: f.fkih@qu.edu.sa

signals, which can significantly impact speech-related applications, particularly ASR^{1–3}. Background noises and competing speakers are the primary sources of target signal distortion. To mitigate the impact of noise, a speech enhancement (SE) system can restore the quality and improve the intelligibility of degraded signals. A SE model performs well in various noisy backgrounds; however, developing a model that can handle different noisy backgrounds with minimal complexity and latency is a challenging task.

Traditional SE techniques, such as spectral subtraction⁴, Wiener filtering⁵, and statistical models^{6,7}, perform better in stationary noisy backgrounds. However, they perform poorly in nonstationary, noisy backgrounds. Deep learning (DL) has become a mainstream approach for speech enhancement⁸. Deep learning techniques learn to transform a noisy speech into a clean speech by training on a dataset of paired clean and noisy samples. These techniques may use mapping-based training objectives^{9–11} or masking-based training objectives to estimate the spectrum or time-frequency masks^{12–16}. The commonly used deep learning techniques for speech enhancement include fully connected networks (FCN)¹⁷, RNNs¹⁸, and CNNs¹⁹. To enhance noisy speech, Long Short Term Memory (LSTM) is used to develop a noise- and speaker-independent model²⁰. The model is trained using a four-layered LSTM network on speech samples from different speakers combined with various types of background noise. Another approach²¹ proposes a CNN architecture applying gated and dilated convolution. Another trend uses an attention mechanism to enhance noisy speech²². In^{23,24}, the LSTM model is proposed for speech enhancement by applying the attention gate to replace the forget gate. The study in²² proposes a self-attention dense CNN for better feature extraction and uses feature reusing. The study in²⁵ proposes a dual-path RNN with self-attention such that the processing of long sequences is improved. Several studies use attention mechanisms to enhance speech signals with promising results^{26–28}. These existing RNN models have good capability for noise suppression but suffer from their complex structure and long training times. GRU (Gated Recurrent Unit) is a recurrent neural network that can be used in speech enhancement for learning long-term temporal dependencies^{29–32}.

Nevertheless, speech enhancement focuses on refining the models to estimate the target speech, distinct from the speech recognition aspect. Consequently, speech enhancement approaches often do not align with the ultimate objective, resulting in suboptimal outcomes³³. Moreover, the output speech from these enhancement techniques tends to be excessively over-smoothed, leading to post-enhancement speech distortion. This distortion can significantly impact the effectiveness of ASR systems³⁴. Consequently, the success of this approach relies heavily on the success of the front-end enhancement³⁵. To enhance the noise robustness of ASR, three primary approaches are commonly used. The first approach involves integrating a speech enhancement component at the front end of the ASR system. The second method employs multi-condition training to enhance the noise robustness of ASR. This involves training the speech recognition model using various types of data, including both clean and noisy speech. However, this approach leads to increased complexity and computational costs. Moreover, it often yields underwhelming results when faced with unmatched conditions³⁶, and its performance can be impacted by speech distortion³⁷. The third prevalent approach involves joint training techniques^{38,39}, which utilize a unified framework to optimize both speech enhancement and recognition simultaneously. The rationale behind this approach is that speech enhancement and recognition are intertwined tasks that can mutually enhance each other performance. For instance, to improve the noise robustness of end-to-end ASR, a joint adversarial enhancement training method was proposed in⁴⁰. This method leverages the joint training framework to refine both the mask-based enhancement network and the attention-based encoder-decoder speech recognition network. Furthermore, even on the noisy AISHELL-1⁴⁰ dataset, the CER remains above 50%, indicating a need for improvement. On the other hand, concerning E2E speech recognition, speech transformer models have demonstrated remarkable performance, achieving state-of-the-art results. The self-attention network^{41,42} stands out as a crucial element of the speech transformer, offering greater capability in capturing long-term dependencies compared to sequence-to-sequence models based on recurrent neural networks (RNNs). The more recent literature of the SE and ASR can be found in studies such as^{43–49}.

To understand the problem and the need to jointly optimize SE and ASR, we analyze the spectrograms illustrated in Fig. 1, which depicts an example spectrogram of a test speech sample. The spectrogram of the enhanced speech, processed by the enhancement network, exhibits notable leaks, as shown in Fig. 1 (right), indicated by the highlighted boxes, resulting in speech distortion. These boxes indicate significant leaks, primarily due to the dominance of noise signals in these time-frequency bins, overshadowing the target speech. Consequently, the enhancement network interprets these time-frequency bins as noise signals and eliminates

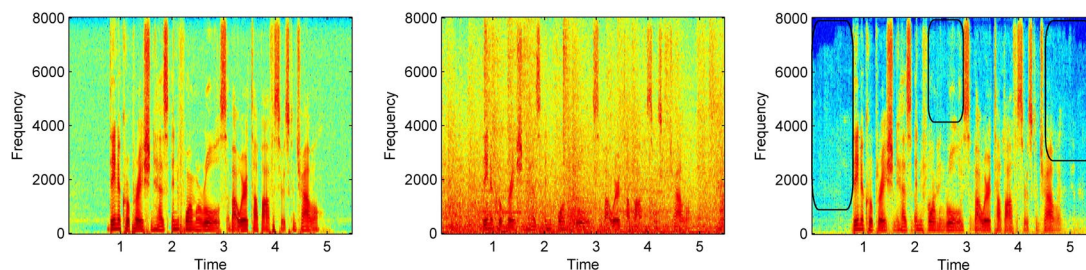


Fig. 1. An example spectrogram of a test speech sample. Clean speech (left), noisy speech (middle), and enhanced speech without joint optimization (right). The boxes highlight the spectral leaks (over-smoothed distortions).

the relevant information, such as formants. Although the enhancement network manages to reduce noise signals to some extent, these leaks remain unrecognized by the ASR system, leading to substantial loss of essential speech details. These factors explain how speech distortion adversely affects the performance of automatic speech recognition.

This paper presents a jointly optimized speech enhancement and automatic speech recognition model that aims to automatically acquire more robust representations that are well-suited for the recognition task. The contributions of this study are twofold.

- This study proposes a speech enhancement based on an attentional-codec model to effectively reduce noise in the target speech while minimizing distortions, such as over-smoothed spectrograms. The proposed Speech Enhancement (SE) network improves noisy speech using an attention process that mirrors human focus on specific speech components amidst surrounding noise. By employing this attention process within the codec (coder-decoder), the model achieves enhanced sequential modelling, allowing learned weights from past input features to predict current features accurately. This attention mechanism actively manages the correlation between preceding and current frames, assigning attention weights to earlier speech frames. Experimental results demonstrate that the proposed SE model surpasses baseline methods in terms of speech quality, intelligibility, noise reduction, and speech distortion.
- Traditionally, ASR systems have often depended on utilizing enhanced features as inputs. In our study, however, we use a dynamic fusion approach to overcome this limitation. This approach integrates both the enhanced features and the raw noisy features to filter out noise signals from the enhanced target speech while simultaneously capturing fine details from the noisy signals. By employing this fusion approach, we aim to reduce speech distortions and enhance the overall performance of the ASR system.

The paper is structured as follows: “**Proposed speech enhancement**” presents the proposed speech enhancement approach. “**Speech enhancement experiments**” details the experiments, results, and discussions about speech enhancement. “**Joint optimization and ASR**” discusses jointly optimized speech enhancement and automatic speech recognition with corresponding results. Finally, “**Summary and conclusion**” provides the conclusion of this study.

Proposed speech enhancement

Figure 2 shows the diagram of the proposed SE. A clean speech and background noise can be represented by $s(t)$ and $d(t)$. The resulting noisy speech $y(t)$ is obtained by mixing $s(t)$ and $d(t)$, given as:

$$y(t) = s(t) + d(t) \quad (1)$$

where $\{y, s, d\} \in \mathbb{R}^{M \times 1}$ and M shows speech samples. The speech enhancement network recovers the estimate $\hat{s}(t)$ of underlying clean speech $s(t)$ from a noisy speech $y(t)$. The SE network is fed with inputs $Y = [y_1, y_2, \dots, y_t, \dots, y_N]$ and $X = [x_1, x_2, \dots, x_t, \dots, x_N]$, where Y and X represent the magnitudes of the noisy mixture and underlying clean speech at frame t . The encoder extracts features h , given as:

$$h^K, h^Q = \text{Encoder}(y, x) \quad (2)$$

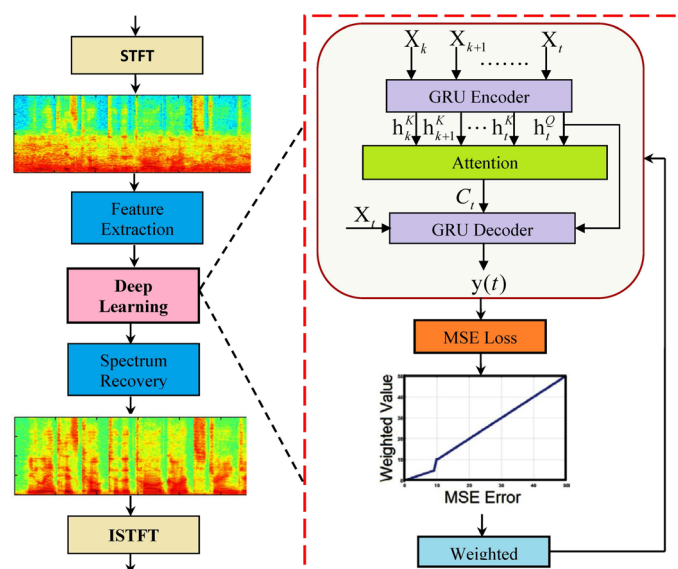


Fig. 2. The proposed speech enhancement pipeline.

whereas the parameters Q and K represent the query and key. Our study uses a Gated Recurrent Unit (GRU) encoder-decoder, showing an ability to model sequential information, resulting in lower computational costs and improved performance as compared to LSTM, as reported in ⁵⁰. To generate fixed-length context vectors, the attention process is applied to the key and query inputs.

$$C^t = \text{Attention}(h^K, h^Q) \quad (3)$$

With the context vectors C^t and h^Q , the output of the decoder $\bar{y}(t)$ recovers the noisy enhanced speech $\hat{x}(t)$.

$$\bar{y}(t) = \text{decoder}(C^t, h_t^Q, y_t) \quad (4)$$

Figure 3 shows the attentional-GRU codec. The encoder extracts features from the speech spectrum. To accomplish this, the extracted features are provided to the input layer. Where $f(\cdot)$ represents a neural network (GRU) function, and h_t^K is the GRU output, respectively.

$$\bar{x}(t) = \tanh(W_s x_t + b_s) \quad (5)$$

The $\bar{x}(t)$ is the input to the GRU cell as:

$$h_t^K = f(\bar{x}_t) \quad (6)$$

where $f(\cdot)$ represents the GRU function, and h_t^K is the GRU output, respectively. The h_t^Q can be computed as:

$$h_t^Q = f(h_t^K) \quad (7)$$

Unidirectional attentional-GRU encoder

A gated recurrent unit (GRU) is an RNN type that includes a gating mechanism for controlling the flow of information. The unidirectional GRU processes the sequence in one direction and is commonly used for sequence-to-sequence learning. Since the GRU has fewer parameters to optimize, it can mitigate the gradient vanishing problem and train faster than LSTM. This study employs a modified GRU in which the classical hyperbolic tangent (\tanh) is replaced with an attention-based ReLU (ARELU)⁵¹, a learnable activation function that leverages an element-wise attention approach. The hyperbolic tangent shows high complexity because of dense activation computations. ARELU employs learned data-adaptive parameters to amplify positive elements and diminish negative elements. The training process remains robust towards vanishing the gradient since the attention mechanism in ARELU activation learns element-wise residues of the active region. The attention activation learning through ARELU leads to well-focused activations in significant areas of the feature map. Having additional learnable parameters (α and β) per layer enables fast network training at low learning rates. According to study⁵¹, ARELU is denoted as $f(x_i, \alpha, \beta)$ using a combination of an element-by-element sign-based attention approach $l(x_i, \alpha, \beta)$ and the classical ReLU $R(x_i)$, as follows:

$$f(x_i, \alpha, \beta) = l(x_i, \alpha, \beta) + R(x_i) \quad (8)$$

$$f(x_i, \alpha, \beta) = \begin{cases} D(\alpha)x_i & \text{when } x_i < 0 \\ (1 + \sigma(\beta))x_i & \text{when } x_i \geq 0 \end{cases} \quad (9)$$

where $X = x_i$ is input to the activation layer, $[\alpha, \beta] \in \mathbb{R}^2$, $D(\cdot)$ grasps the input variables to $[0.01, 0.99]$ such that preventing α to be zero, and σ shows the sigmoid activation. The inclusion of ARELU in GRUs can assist in capturing long-term contextual dependencies between features, which is critical in SE. As a result, in addition to preventing gradient vanishing, the use of ARELU in the GRU can aid in capturing these long-term dependencies and improving the performance of SE. The attention-ReLU-based GRU cell structure is depicted in Fig. 4.

Attentional process

The attention process plays a crucial role in creating fixed-length context vectors by processing information about the key and query inputs. An attention mechanism can process preceding and future speech frames.

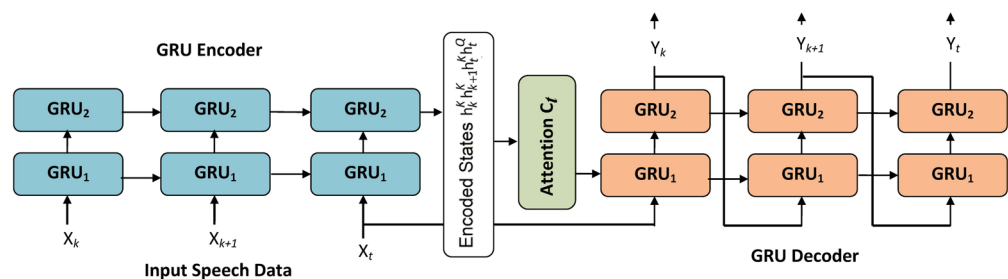


Fig. 3. The Architecture of the proposed stacked attention encoder-decoder unidirectional GRU.

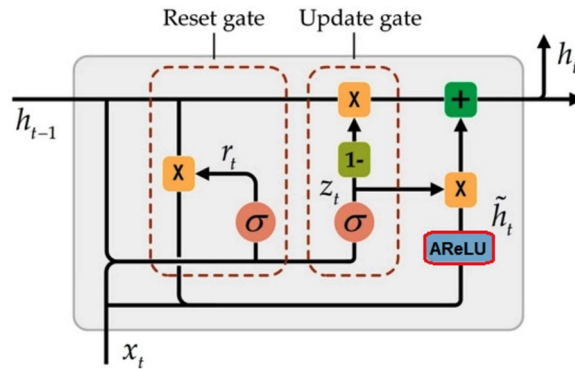


Fig. 4. Attention-ReLU in GRU cell structure.

However, speech enhancement in this study tends to avoid processing latency and therefore only uses previous speech frames. To achieve this, the model uses both causal dynamic and causal local attention approaches. In causal dynamic attention, the model uses the entire previous speech sequence $Y = [y_1, \dots, y_t]$, and the input sequence $X = [x_1, \dots, x_t]$ for computing attention weights. This indicates that all preceding frames are utilized to enhance the present frame. However, for long speech sequences, the attention weights of many preceding frames may become almost zero. To address this, the model uses the causal local attention process, where $Y = [y_{t-z}, \dots, y_t]$, and $X = [x_{t-z}, \dots, x_t]$ are utilized for computing attention weights. The model learns the attention weights κ as:

$$\kappa_{tk} = \frac{\exp(h_k^K, h_t^Q)}{\sum_{k=l}^t \exp(h_k^K, h_t^Q)} \quad (10)$$

where $l = 1$ is causal dynamic attention and $l = (t - z)$ denotes causal local attention with z is a constant. According to the correlation computation:

$$\exp(h_k^K, h_t^Q) = h_k^{KT} W h_t^Q \quad (11)$$

The attention-weighted context vectors are given as:

$$C^t = \sum_{m=l}^{\tau} (\kappa_{tm} h_m^K) \quad (12)$$

The proposed SE model determines the attention process for a speech frame with attention-weighted context vectors.

Unidirectional attentional-GRU decoder

The GRU decoder reconstructs the speech samples with input features, encoder outputs, and attention-weighted context vectors, respectively. The enhanced vector E_t , is computed using the attention-weighted context vectors and features.

$$E_t = \tanh(W_E [C^t; h_t^Q] + b_E) \quad (13)$$

The context vectors and feature vectors are concatenated as $[C^t; h_t^Q]$. A time-frequency mask is finally estimated from the final vectors, given as:

$$M_{t,f}^{IRM} = \sqrt{\frac{|S_{t,f}|^2}{|S_{t,f}|^2 + |D_{t,f}|^2}} \quad (14)$$

where $M_{t,f}^{IRM}$ shows an ideal ratio mask. The magnitude spectrums of clean speech $s(t)$ and noise signals $d(t)$ are denoted as $|S_{t,f}|$ and $|D_{t,f}|$ respectively. To reconstruct the noisy enhanced speech, multiply the noisy features by the enhanced vectors, and obtain the resulting signal by performing the inverse Short-time Fourier Transform (STFT).

$$\bar{y}_t = y_t \odot \text{sigmoid}(W_m E_t + b_m) \quad (15)$$

The enhanced features along with raw noisy features are further used for joint SE and ASR optimization.

Speech enhancement experiments
Dataset and data generation

The experiments utilized the LibriSpeech dataset⁵², and noise sources were selected from the AURORA database⁵³. LibriSpeech is a database of approximately 1000 h of English speech. The database includes 16,000 audio files, each 10 s in length, derived from public-domain audiobooks from the LibriVox project. The database is divided into several subsets, including test-clean, test-other, train-clean-100, train-clean-360, and train-other-500, intended for development, testing, and training purposes. The speech dataset D includes training and testing sentences, denoted as M_{tr} and M_{te} , respectively. The training and testing sentences are labelled as D_{tr} and D_{te} . The noisy sentences are created by mixing background noises with D_{tr} and D_{te} .

$$y_{tr}^i = s_{tr}^i + d_{tr}^i, i = 1, 2, 3, \dots, M^{tr} \tag{16}$$

$$y_{te}^j = s_{te}^j + d_{te}^j, j = 1, 2, 3, \dots, M^{tr} \tag{17}$$

where y_{tr}^i and y_{te}^j denote training and testing noisy data.

Feature extraction

The noisy-clean pairs $y(t)$, $s(t)$ are transformed to the frequency domain by applying the Short-Time Fourier Transform (STFT), given as:

$$Y = STFT(y); S = STFT(s) \tag{18}$$

where $\{S, Y\} \in \mathbb{Z}^{(T \times F)}$, F is the number of frequency bins, and T shows the number of frames. This study has used the STFT magnitude $|Y|$ as the input feature.

SE network architecture

The architecture of the proposed GRU-based codec consists of the input layer, three GRU layers each containing 256 units, and an output layer containing 257 sigmoidal-activated units. The hyperparameters include epochs (160), learning rate (0.0001), and weights (randomly initialized), respectively. The training process utilized mini-batches of 32 sequences, employing back-propagation through time with an Adam optimizer. The GRU layer configuration is given as (161/256/256/256/257) units. Table 1 provides the details of the hyperparameters. Noisy sentences are generated using -5dB , 0dB , and $+5\text{dB}$ SNRs. The sentences of both genders are repeated for all SNRs and mixed with all noises, resulting in 21,600 sentences (approximately 18 h). These sentences are used to train the proposed SE model. During testing, half of the speech sentences are used in matched and half are used in mismatched noisy conditions. All noises are tested with distinct sentences. Sentences are sampled at a 16 kHz rate, and a Hanning window (512 points) with 75% overlapping is used in experiments. Usually, a noisy phase is used during speech reconstruction; however, this study uses an estimated phase⁵⁴ to reconstruct the final speech. A loss function quantifies the differences between a predefined mask and an estimated mask in masking-based SE. The goal of the loss function is to minimize errors. Typically, MSE is used as a loss function in TF-masking-based SE, defined as

$$MSE_l[f(S), Y] = [f(S) - Y]^2 \tag{19}$$

where $f(S)$, S , and Y represent the model output, input, and ground truth label. The MSE in Eq. (19) can be expressed as:

Component	Parameter	Value/range	Description
STFT preprocessing	Window Size	512 points	Hanning window
	Hop Length	128 points (75% overlap)	Frame overlap
	FFT Bins	257	Frequency bins
SE Network	GRU Layers	3	Stacked layers
	Units per Layer	256	Hidden units
	Attention Window (Z)	5 frames	Local attention span
	AReLU Init (α, β)	$\alpha \in [0.01, 0.99], \beta \in \mathbb{R}$	Learned parameters
Training	Batch Size	32	Mini-batch size
	Learning Rate	0.0001	Adam optimizer
	Epochs	160	Training iterations
	Loss Weight (γ)	Adaptive (init = 1.0)	SE-ASR balance
	WMSE Threshold (B)	10	Dynamic weighting
Fusion Network	GRU Hidden Units	256	Fusion layer size
	Fusion Steps (p)	3	Iteration count

Table 1. Hyperparameters and initial conditions.

$$MSE_l(\hat{M}(s), M(s)) = (\hat{M}(s) - M(s))^2 \quad (20)$$

The estimated mask and predefined mask are represented as $\hat{M}(s)$ and $M(s)$, respectively. A dynamic-weighted loss function is employed to enhance network learning. This loss function multiplies weighted values by the learning errors. With such a process, the loss function emphasizes larger errors, enhancing overall performance. The weighted Mean Squared Error (WMSE) is calculated by multiplying the MSE function by a weight variable λ .

$$MSE_l(f(s), y) = \lambda \otimes (f(s) - y)^2 \quad (21)$$

To give more importance to the situations with significant errors, the weighting variable λ in Eq. (21) is modified based on the following formula:

$$MSE_l(f(x), y) = \lambda[(f(x) - y) \otimes (f(x) - y)^2] \quad (22)$$

The following conditions are applied to select the weights, given as:

$$\lambda(f(x), y) = \begin{cases} \frac{|f(x), y|}{2}, & |f(x), y| < B \\ |f(x), y|, & |f(x), y| \geq B \end{cases} \quad (23)$$

When the absolute divergence falls below a constant value of λ (experimentally set to 10; since the model performs better at $B=10$), the weighting is reduced by half. When the error is smaller than B , the weighting factor is reduced by half, indicating the model does not focus as much on small errors. Similarly, When the error is greater than or equal to B , the weight λ is set equal to the error magnitude, meaning the model will focus strongly on these larger errors.

SE evaluation metrics and related models

To examine the proposed SE, this study uses well-adopted metrics including Short-time Objective Intelligibility (STOI)⁵⁵, Perceptual Evaluation of Speech Quality (PESQ)⁵⁶, and Source-to-Distortion Ratio (SDR). In this study, we chose LSTM²⁰ and DNN¹⁷ as baseline models for estimating Ideal Ratio Mask (IRM). The baseline models are represented as **LSTM+IRM** denotes that LSTM is used to estimate IRM; **DNN+IRM** indicates that a fully-connected DNN estimates IRM, and **GRN+IRM** indicates that the proposed SE estimates IRM as a training objective.

SE results and discussions

The study first compares the performance of the proposed SE against the baselines. Tables 2 and 3 displays the average test results of three metrics (STOI, PESQ, and SDR) across four testing noises and three SNRs for both matched and unmatched noisy conditions. It is important to highlight that, unlike the baselines, the proposed GRN+IRM consistently outperforms them across all noisy testing scenarios.

Table 2 displays the results of speech enhancement under matched conditions, where the proposed GRN+IRM demonstrates better values for all objective measures in all background noises. Specifically, at low SNR (−5dB), the proposed SE network achieves the highest STOI ($\geq 86.25\%$) and PESQ (≥ 2.19) values for airport noise,

Noise	Metric	STOI in %				PESQ				SDR in dB			
Types	SNR	− 5dB	0dB	5dB	Avg	− 5dB	0dB	5dB	Avg	− 5dB	0dB	5dB	Avg
Airport Noise	Noisy Speech	63.05	69.76	83.95	72.25	1.64	1.86	2.14	1.88	−4.78	0.11	5.07	0.13
	DNN+IRM	80.85	84.54	90.74	85.38	1.83	2.21	2.59	2.21	3.98	6.86	8.38	6.41
	LSTM+IRM	83.55	87.65	92.36	87.85	2.01	2.34	2.67	2.34	4.09	7.1	9.54	6.91
	GRN+IRM	86.25	89.58	94.55	90.13	2.19	2.47	2.78	2.48	4.21	7.33	10.7	7.41
Babble noise	Noisy speech	57.75	68.05	79.67	68.52	1.52	1.75	2.07	1.78	−4.73	0.13	5.08	0.16
	DNN+IRM	74.22	78.35	85.93	79.51	1.91	2.24	2.48	2.22	3.82	6.31	8.74	6.29
	LSTM+IRM	76.85	80.64	87.45	81.65	2.04	2.35	2.66	2.35	3.95	6.58	9.05	6.52
	GRN+IRM	80.22	82.25	89.27	83.91	2.15	2.46	2.75	2.45	4.08	7.4	9.36	6.94
Car Noise	Noisy Speech	58.84	68.92	79.60	69.12	1.37	1.62	1.92	1.63	−4.85	0.08	5.05	0.09
	DNN+IRM	78.65	81.74	86.77	82.38	1.74	2.18	2.48	2.13	3.81	6.42	8.83	6.35
	LSTM+IRM	80.23	84.47	89.18	84.62	1.99	2.39	2.55	2.31	4.2	7.21	9.92	7.11
	GRN+IRM	85.48	86.56	92.68	88.24	2.09	2.47	2.71	2.42	4.56	7.59	10.8	7.65
Factory Noise	Noisy Speech	58.44	67.44	78.80	68.22	1.31	1.61	1.92	1.61	−4.69	0.12	5.07	0.17
	DNN+IRM	78.25	80.71	85.25	81.41	1.66	2.15	2.47	2.09	3.66	6.34	8.72	6.24
	LSTM+IRM	79.45	82.45	88.17	83.35	1.89	2.33	2.55	2.26	3.85	5.53	9.52	6.3
	GRN+IRM	81.62	82.15	90.44	84.73	2.11	2.45	2.77	2.44	4.01	6.69	10.3	6.99

Table 2. SE performance in matched testing conditions.

Noise	Metric	STOI in %				PESQ				SDR in dB			
Types	SNR	−5dB	0dB	5dB	Avg	−5dB	0dB	5dB	Avg	−5dB	0dB	5dB	Avg
Airport Noise	Noisy Speech	60.95	71.84	82.24	71.67	1.55	1.85	2.14	1.84	−4.77	0.11	5.07	0.14
	DNN+IRM	77.36	83.24	87.48	82.69	1.81	2.25	2.58	2.21	3.9	6.78	8.22	6.3
	LSTM+IRM	80.14	85.46	89.87	85.15	1.94	2.34	2.67	2.31	4.01	7.06	9.46	6.84
	GRN+IRM	82.77	87.91	92.37	87.68	2.08	2.42	2.75	2.42	4.13	7.34	10.7	7.39
Babble Noise	Noisy Speech	54.64	65.94	77.48	66.02	1.39	1.74	2.04	1.72	−4.69	0.16	5.1	0.19
	DNN+IRM	71.15	78.47	80.77	76.79	1.77	2.01	2.56	2.11	3.88	5.85	8.66	6.13
	LSTM+IRM	73.55	79.87	82.33	78.58	1.88	2.17	2.64	2.23	3.98	6.01	8.94	6.31
	GRN+IRM	75.96	81.26	85.84	81.02	2.01	2.33	2.71	2.35	4.09	6.18	9.23	6.5
Car Noise	Noisy Speech	57.27	67.48	78.44	67.73	1.39	1.63	1.93	1.65	−4.78	0.1	5.07	0.13
	DNN+IRM	75.38	80.29	83.21	79.62	1.72	2.22	2.44	2.12	3.99	6.83	8.31	6.38
	LSTM+IRM	78.19	83.37	87.55	83.03	1.95	2.31	2.67	2.31	4.14	7.11	9.51	6.92
	GRN+IRM	80.91	86.42	90.81	86.04	2.14	2.45	2.73	2.44	4.3	7.45	10.6	7.45
Factory Noise	Noisy Speech	55.24	65.93	77.24	66.13	1.31	1.6	1.92	1.61	−4.67	0.12	5.08	0.18
	DNN+IRM	70.74	75.39	81.25	75.79	1.71	2.09	2.44	2.08	3.43	5.98	8.34	5.92
	LSTM+IRM	73.92	77.41	84.47	78.61	1.84	2.25	2.59	2.22	3.68	6.22	9.27	6.39
	GRN+IRM	77.35	79.49	87.64	81.49	1.97	2.41	2.71	2.36	3.94	6.47	10.2	6.87

Table 3. SE performance in unmatched testing conditions.

Condition	GRN+IRM-Matched			GRN+IRM-Unmatched			Average		
	STOI	PESQ	SDR	STOI	PESQ	SDR	STOI	PESQ	SDR
Results	86.75	2.45	7.20	82.27	2.39	6.97	84.51	2.42	7.09

Table 4. The overall results (Matched and Unmatched), averaged over all testing SNRs and noises.

whereas the best SDR (≥ 4.56 dB) is achieved at -5 dB for babble noise. Taking the babble noisy case (matched condition) at -5 dB SNR, STOI improves from 63.05% with noisy speech to 86.25% with GRN+IRM, resulting in a 23.2% improvement in STOI. Furthermore, STOI improves from 80.85% with DNN+IRM to 86.25% with GRN+IRM, resulting in a 5.4% improvement in STOI. Similarly, in the case of factory noise (matched condition) at 0dB SNR, PESQ increases from 1.61 with UnP to 2.45 with the proposed GRN+IRM, achieving a 0.84 (34.28%) improvement. Additionally, PESQ increases from 2.18 with DNN+IRM to 2.47 with the proposed GRN+IRM in-car noise, resulting in a 0.29 (11.74%) improvement over the DNN+IRM. In a matched condition, consider street noise at 5dB as another case, where the SDR value increases from 0.11dB with UnP to 6.82dB with GRN+IRM, achieving an improvement of 6.71 dB. On average, at low SNR (-5 dB) in matched conditions, the proposed GRN+IRM increases STOI (by 16.34%), PESQ (by 0.71), and SDR (by 7.17dB) over noisy unprocessed speech, demonstrating the effectiveness of the proposed SE model.

Table 3 presents the results of speech enhancement conducted under unmatched conditions, where the proposed GRN+IRM with an IRM training objective achieves better average values for all objective measures in all background noises. Specifically, at low SNR (-5 dB), the proposed model achieves the highest STOI ($\geq 82.77\%$), and PESQ (≥ 2.16) in street noise, whereas the best SDR (≥ 4.3 dB) is achieved in-car noise. In the case of babble noise at 0dB SNR under unmatched conditions, the STOI improves from 71.84% with noisy speech to 87.91% with GRN+IRM, resulting in a 16.07% improvement in STOI. Also, for the factory noisy case (unmatched condition) at 0dB SNR, the PESQ improves from 1.61 with UnP to 2.41 with the proposed GRN+IRM, representing a 0.80 (33.19%) improvement. Furthermore, in car noise at 5dB, the PESQ improves from 2.44 with DNN+IRM to 2.73 with the proposed GRN+IRM, representing a 0.29 (11.74%) improvement over the DNN+IRM baseline. In the unmatched condition of street noise at 5dB, the SDR value increases from 5.12dB with UnP to 9.71dB with GRN+IRM, resulting in an improvement of 4.59 dB. On average, at low SNR (-5 dB) under unmatched conditions, the proposed GRN+IRM significantly improves the STOI, PESQ, and SDR over noisy unprocessed speech. Table 4 provides average scores encompassing all background noises for matched conditions (GRN+IRM-Matched), unmatched conditions (GRN+IRM-Unmatched), and the average of both conditions.

Table 5 shows the performance of the causal local attention (CLA) for which the values of W are varied between (5–15). The results (PESQ and STOI) indicate that the value of W greater than 15 shows no competitive results, and the best SE results are obtained with $Z = 5$. Therefore, $Z = 5$ is fixed for the proposed SE. It was observed that the causal local attention outperformed the causal dynamic attention. These findings support the assumption that substantial preceding information is not necessary for effective speech enhancement, as noisy conditions can change rapidly over time. These observations apply to the attention networks, as the attention GRU performed better than the baseline GRU.

Z	STOI (in%)			PESQ		
SNR	−5dB	0dB	5dB	−5dB	0dB	5dB
5	86.21	89.56	94.54	2.18	2.47	2.77
15	82.45	84.74	89.83	2.15	2.42	2.69
25	79.56	82.74	88.12	2.08	2.34	2.66

Table 5. Causal local attention with different weight values.

Model	Metric	Error
GRN+MSE	PESQ: 2.32, STOI: 83.87%	3.23×10^{-4}
GRN+WMSE	PESQ: 2.45, STOI: 86.21%	3.61×10^{-4}
Improvement	PESQi: 4.91%, STOIi: 2.34% & 10.52%	10.52%

Table 6. Dynamically-Weighted vs. non-dynamically-weighted loss.

Model	Para#	MACs	Param size
GRN+IRM	2.138 M	0.245 G/s	2.71 MB
LSTM ²⁰	4.672 M	0.412 G/s	5.43 MB
RLSTM ⁵⁷	10.0 M	1.347 G/s	13.55 MB

Table 7. Computational efficiency.

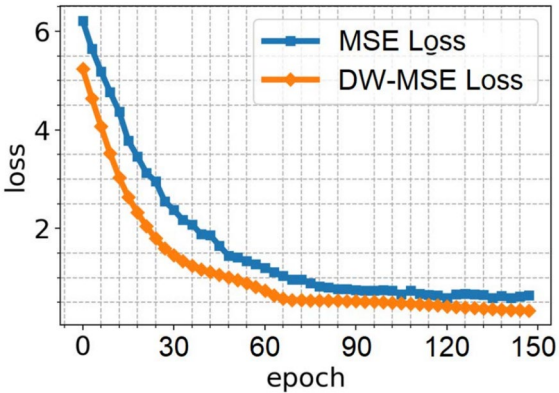


Fig. 5. Learning error: DW-MSE and without DW-MSE loss.

Table 6 shows a comparison of errors and predicted results (STOI and PESQ) between GRN+IRM with weighted and without weighted loss functions. The results indicate that the weighted loss function improved the PESQ and STOI after incorporating the proposed GRN+IRM. The use of weighted mean square error (WMSE) reduces the errors to 3.23×10^{-4} as compared to 3.61×10^{-4} with non-weighted MSE. The learning error is reduced by 10.52% with the weighted loss function. Due to limitations in computational resources in practical applications, it is crucial to establish an optimal balance between the model's performance improvement and parameter efficiency. Table 7 illustrates the efficiency of parameters in the proposed speech enhancement model. The parameter efficiency of these SE models reveals that the integration of the attention process into GRU does not significantly affect the parameter count (2.138M) and parameter size (2.71 MB) compared to LSTM (4.672M, 5.43 MB) and residual LSTM (RLSTM) (10M)⁵⁷. To employ the proposed GRN+IRM on embedded systems, it is essential to minimize hardware memory usage. Consequently, we present a summary of multiply-accumulate operations (MACs). The proposed GRN+IRM model achieves 0.245 G/s MACs with an attention process, ensuring efficiency without compromising SE performance. The integration of GRU has significantly reduced parameter numbers, parameter size, and MACs. We further analyzed the convergence of the proposed model after incorporating weighted MSE, as shown in Fig. 5. It can be observed that the weighted MSE converges faster than the traditional MSE.

Comparison with related studies

This study compares the performance of the GRN+IRM with several selected studies from the literature to showcase its superiority. The comparison is performed for three different SNR levels (−5dB, 0dB, and 5dB) and the results are presented in Table 8. The study finds that the GRN+IRM model, with the IRM training objective, performs highly competitively as compared to recent models, except PL-CRN⁶¹, which performs slightly better at less adverse SNR (5 dB). CRN-BLSTM⁶⁰ gained 0.48 (21.62%) PESQ over noisy mixture, which indicates 7.36% lower performance than the proposed GRN+IRM. Similarly, CNN-GRU⁶² gained 0.59 (25.32%) PESQ over a noisy mixture, which shows 3.64% lower performance than the proposed GRN+IRM. Furthermore, the gain in STOI for MCBNet⁵⁹ over noisy mixture is 8.31%, indicating 8.03% less STOI gain as compared to the GRN+IRM. Additionally, the STOI improves from 84.25% with DCCRN⁶⁴ to 86.75% with GRN+IRM. The proposed GRN+IRM outperforms related models by significant margins, such as a PESQ improvement of 0.31 (14.28%) and an STOI improvement of 11.55% over the state-of-the-art GRN⁶⁷ and AECNN⁶⁸ models at the -5dB SNR level.

Subjective evaluation

Furthermore, we conducted subjective listening tests to evaluate the quality of the enhanced speech. We randomly selected 300 sentences from different background noises at −5dB, 0dB, and 5dB to assess the performance of the DNN, LSTM, and proposed GRN+IRM. The participants were asked to rate the speech quality on a scale from 0 to 5. The subjective tests are performed in a soundproof room using high-quality headphones. Before the tests, training sessions were conducted to familiarize the listeners with the procedures. Figure 8 displays the results of the Mean Opinion Score (MOS), a numerical measure of the human-judged overall quality, also known as the subjective listening test, where the proposed GRN+IRM model demonstrated superior MOS performance. The average MOS score was higher than 2.80 (with MOS ≥ 2.86 at −5dB) for negative SNRs, indicating considerable SE performance. For SNR ≥ 0dB, the GRN+IRM model yielded a MOS score greater than 3.0 (MOS ≥ 3.0 at SNR ≥ 0dB). ANOVA statistical analysis for average MOS scores at -5dB, 0dB, and 5dB yielded [F(3,10) = 44.5, p < 0.0001], [F(3,10) = 35.8, p < 0.0001], and [F(3,10) = 27.2, p < 0.0001], indicating the statistical significance of the MOS scores achieved by the GRN+IRM model. FDNN and LSTM also demonstrated improved performance, as deep learning can produce better speech quality. Figure 6 shows the average MOS score of all listeners, where the y-axis shows the MOS score and the x-axis indicates the input SNRs.

Joint optimization and ASR

In conventional joint speech enhancement and ASR, a noisy magnitude spectrum $|Y|$ is used as the input feature. The conventional joint training method comprises two main components: speech enhancement and speech recognition. Initially, the model is trained using both noisy and clean parallel data to enhance speech quality. Subsequently, the improved speech output serves as the sole input feature for the speech recognition model^{71–73}. To optimize the entire system, a combined loss function for both enhancement and speech recognition is employed. This enables the simultaneous training of enhancement and ASR models. However, this approach completely depends on the enhanced features of the speech recognition model, which may still be affected to some extent by speech distortions. Therefore, this study follows the joint optimization approach shown in Fig. 7.

Metric	PESQ					STOI				
SNR (in dB)	−5dB	0dB	5dB	Average	PESQ↑	−5dB	0dB	5dB	Average	STOI↑
Noisy unprocessed	1.46	1.74	2.01	1.74	–	60.25	69.76	81.21	70.41	–
DeepResGRU ³⁰	2.09	2.29	2.49	2.29	0.55	74.13	81.81	85.51	80.48	10.07
CFN-GCFU ⁵⁸	1.98	2.24	2.62	2.28	0.54	71.61	78.19	86.21	78.67	8.26
MCBNet ⁵⁹	2.01	2.32	2.52	2.28	0.54	72.81	79.15	84.15	78.71	8.30
CRN-BLSTM ⁶⁰	1.93	2.23	2.51	2.22	0.48	70.31	77.08	81.96	76.45	6.04
PL-CRN ⁶¹	2.06	2.51	2.85	2.47	0.73	73.16	84.42	90.15	82.57	12.16
CNN-GRU ⁶²	2.01	2.34	2.65	2.33	0.59	74.61	83.11	90.11	82.61	12.20
DTLN ⁶³	1.91	2.34	2.67	2.31	0.57	72.72	85.19	90.68	82.86	12.45
DCCRN ⁶⁴	1.85	2.34	2.78	2.32	0.58	74.51	85.87	92.38	84.25	13.84
DNN-TGSA ⁶⁵	2.01	2.31	2.58	2.30	0.56	74.41	81.21	84.12	79.91	9.50
DeepXi ⁶⁶	1.99	2.21	2.41	2.20	0.46	72.01	81.21	91.99	81.73	11.32
GRN ⁶⁷	1.86	2.16	2.42	2.15	0.41	69.76	76.89	81.42	76.02	5.61
AECNN ⁶⁸	1.92	2.19	2.45	2.19	0.45	72.01	77.78	82.51	77.43	7.02
CRN ⁶⁹	1.92	2.22	2.49	2.21	0.41	70.11	76.95	81.88	76.31	5.90
GAN ⁷⁰	1.72	2.15	2.44	2.11	0.37	65.01	75.71	82.61	74.44	4.03
LSTM ⁷⁰	1.82	2.15	2.44	2.14	0.40	68.78	75.81	81.54	75.37	4.96
GRN+IRM (Proposed)	2.17	2.48	2.75	2.45	0.71	83.29	85.86	92.08	86.75	16.34

Table 8. Comparison with related SE models, where the symbol “↑” indicates improvement over noisy speech. Significant values are in bold.

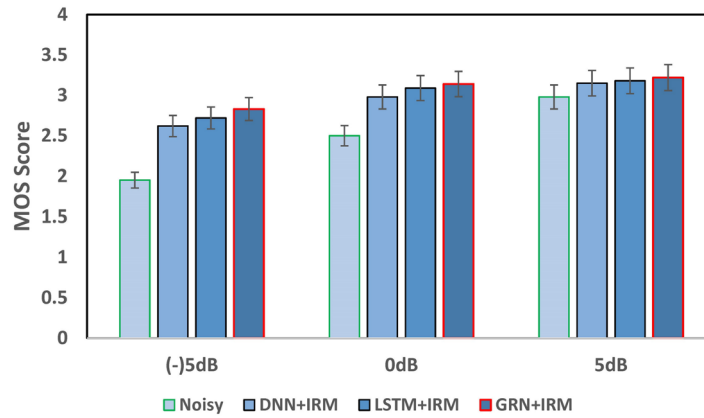


Fig. 6. Average MOS of all participants at SNRs.

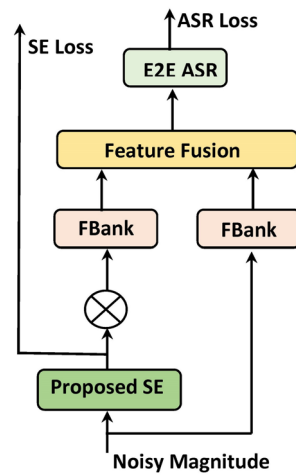


Fig. 7. Schematic of joint SE and ASR.

The spectrograms produced by the speech enhancement network can often display noticeable distortions in the resulting speech. This problem arises when noise dominates in specific time-frequency bands, overshadowing the intended speech signals. Consequently, the speech enhancement identifies these noisy time-frequency bands and removes a significant amount of information, resulting in distortions that lead to the loss of important speech elements like formats. Even though the speech enhancement effectively reduces background noise, these distortions persist undetected by the ASR system. Such distortions ultimately contribute to the decline in ASR performance. To tackle this challenge, we implement the fused GRU (F_{GRU}) approach to combine noisy and enhanced features, as illustrated in Fig. 8. This method aims to mitigate the impact of these distortions and enhance the overall performance of the ASR system.

Regarding the feature fusion network, our approach involves employing two GRUs simultaneously, denoted by $G(\cdot)$, demonstrated in Fig. 9. The goal is to derive deep representations for enhanced ($\xi_{enhanced}$) and noisy (ξ_{noisy}) features. In the initial stage of fusing noisy features ξ_{noisy} with enhanced features $\xi_{enhanced}$ at $p = 1$, the hidden state h_0 is initialized randomly. At the reset gate of GRU for step p , the hidden state h_p and noisy input features ξ_{noisy} decide the status of the reset gate. The status of the update gate is also decided by h_p and ξ_{noisy} , given as:

$$r = \sigma(W_r, (\xi_{noisy}, h_p)) \quad (24)$$

$$z = \sigma(W_z, (\xi_{noisy}, h_p)) \quad (25)$$

where W_r and W_z are weights of reset and update gates. The reset gate r determines the memorization of past information by using element-wise product \odot , given as:

$$h'_p = r \odot h_p \quad (26)$$

$$h_c^p = \tanh(W_h(\xi_{noisy}, h'_p)) \quad (27)$$

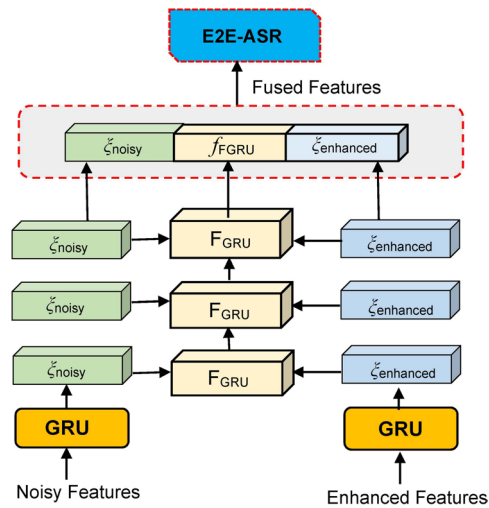


Fig. 8. Noisy and enhanced features fusion with GRU.

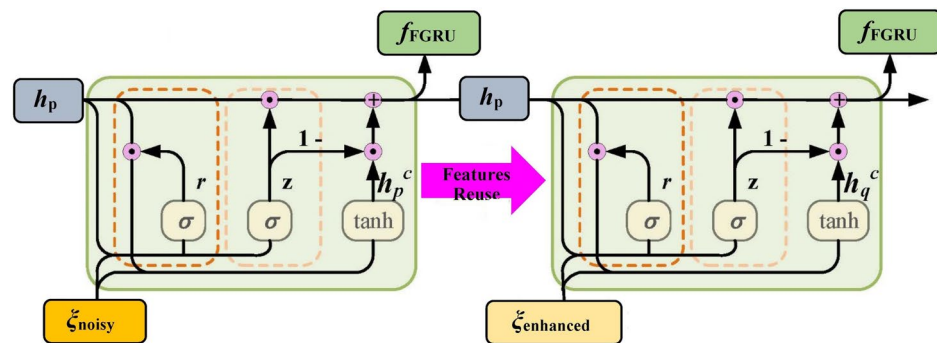


Fig. 9. Features fusion block with GRUs.

The h'_p helps in remembering long-term information. The selective fusion of features combines ξ_{noisy} and h_p at step p , given as:

$$h_q = z \odot h_p + (1 - z) \odot h_c^p \quad (28)$$

The above Eq. (28) connects the input gate $(1 - z)$ and the forget gate z . Finally, after three stages of F_{GRU} , the features are concatenated to obtain the final features F_{feat} , given as:

$$F_{feat} = \text{Concat}(\xi_{noisy}, f_{FGRU}, \xi_{enhanced}) \quad (29)$$

The fused features F_{feat} are used as input to the Transformer-based ASR system. To jointly train the ASR and the proposed SE, the loss function is given as:

$$Loss = Loss_{ASR} + \gamma Loss_{ENH} \quad (30)$$

The parameter γ controls the enhancement loss \mathcal{L}_{ENH} . In the SE module, the parameter γ is not fixed but is adaptively optimized during training to dynamically balance multiple objectives within the composite loss function. Instead of manually setting a static value, γ is learned alongside model parameters, allowing the model to adjust its focus based on the training dynamics and data characteristics. By learning γ adaptively, the model can prioritize noise suppression or speech fidelity at different stages of training, ultimately converging to an optimal balance that enhances overall performance. This approach leads to a more flexible and efficient enhancement process, tailoring the loss weights to suit the complexities of the input data and task requirements.

ASR results

For a noisy training dataset, speech sentences are selected from the LibriSpeech training set. These sentences are mixed with different noises, with randomly selected SNRs ranging from 0dB to 20dB. The inference set contains noises mixed with speech sentences from LibriSpeech at SNRs of 0dB, 5dB, 10dB, 15dB, and 20dB. Tables 9 and 10 show the results of the joint speech enhancement and transformer ASR. The joint training

Model	0dB	5dB	10dB	15dB	20dB	Average	Clean
Noisy	51.44	39.78	28.15	23.11	22.47	32.99	–
ASR-enhanced	24.35	16.69	13.29	12.01	11.11	15.49	9.32
ASR-enhanced-GRU	22.02	15.75	12.37	11.65	10.84	14.52	9.32
ASR-enhanced-LSTM	21.92	15.65	11.97	11.22	10.74	14.30	9.32
ASR-enhanced-fused	20.02	14.75	10.37	10.05	9.84	13.01	9.32

Table 9. CER results for Joint Speech Enhancement and ASR on Testing Set.

Model	0dB	5dB	10dB	15dB	20dB	Average	Clean
Noisy	51.44	39.78	28.15	23.11	22.47	32.99	–
ASR-Enhanced	22.18	14.54	11.35	10.09	9.72	11.42	8.21
ASR-Enhanced-GRU	20.13	13.98	11.45	10.02	9.65	11.27	8.21
ASR-Enhanced-LSTM	20.01	13.77	11.02	10.78	9.94	11.37	8.21
ASR-Enhanced-Fused	18.14	12.24	10.01	9.74	8.94	10.23	8.21

Table 10. CER results for joint speech enhancement and ASR on development set.

approach has the potential to improve the efficiency of end-to-end ASR, illustrating the efficacy of the joint training technique. We present the character error rate (CER) for ASR-Enhanced (indicate the concatenation of the noisy features ξ_{noisy} and features enhanced by the proposed SE $\xi_{enhanced}$) and ASR-Enhanced-Fused (indicate the concatenation of the f_{FGRU} , noisy features ξ_{noisy} , and features enhanced by the proposed SE $\xi_{enhanced}$). In addition, we provide results for ASR-Enhanced-LSTM (indicate the concatenation of the noisy features ξ_{noisy} and features enhanced by LSTM-based SE $\xi_{enhanced}$) and ASR-Enhanced-GRU (indicate the concatenation of the noisy features ξ_{noisy} and features enhanced by GRU-based SE $\xi_{enhanced}$). Table 9 shows the CERs for the testing dataset, whereas Table 10 provides results for the development set, respectively. With the proposed speech enhancement and joint ASR, the CERs are improved significantly. Since the proposed SE shows less speech distortion (obtained better SDR (7.09 dB) as compared to LSTM and GRU), the average CERs are improved from 14.30% (with ASR-Enhanced-LSTM) to 13.01% with the proposed ASR-Enhanced-Fused.

Summary and conclusion

This paper proposes a model that optimizes both speech enhancement and automatic speech recognition simultaneously. The objective is to seamlessly enhance speech quality while also refining representations to better suit the recognition task. While the integration of joint speech enhancement and automatic speech recognition techniques has displayed potential in achieving robust end-to-end ASR systems, conventional approaches typically rely on utilizing enhanced features as inputs for ASR systems. To overcome this limitation, our study adopted a dynamic fusion methodology. This approach combines both the enhanced features and the raw noisy features, to eliminate noise signals from the enhanced target speech while simultaneously capturing fine details from the noisy signals. By employing this fusion strategy, we alleviate speech distortions, thereby enhancing the overall performance of the ASR system. Our proposed model comprises an attentional codec equipped with a causal attention mechanism for SE, a fusion network based on Gated Recurrent Units (GRUs), and an ASR system. In the SE network, we utilize a modified GRU architecture where the traditional hyperbolic tangent (*tanh*) activation function is replaced with an attention-based rectified linear unit (ARELU).

The proposed speech enhancement (GRN+IRM) consistently outperforms baselines across noisy testing scenarios. Specifically, under low SNR (−5dB) conditions, our SE network achieves superior STOI ($\geq 86.25\%$), PESQ (≥ 2.19), and SDR (≥ 4.56 dB) scores in matched conditions. Similarly, our model achieves the highest STOI ($\geq 82.77\%$), PESQ (≥ 2.16), and SDR (≥ 4.3 dB) values at low SNRs. Notably, causal local attention outperforms causal dynamic attention, concluding that extensive preceding information might not be necessary for effective speech enhancement, given the rapid changes in noisy conditions. Minimizing hardware memory usage is crucial to ensure the feasibility of deploying the proposed GRN+IRM on embedded systems. Therefore, we examined multiply-accumulate operations (MACs). The proposed model concludes 0.245 G/s MACs with an attention process, ensuring efficiency without compromising SE performance. Our study concludes that the GRN+IRM model, trained with the IRM objective, stands competitively against recent models. With our proposed speech enhancement and joint ASR, significant improvements are observed in character error rates (CERs). Due to reduced speech distortion (achieved a better SDR of 7.09dB compared to LSTM and GRU), the average CERs are enhanced from 14.30% (with ASR-Enhanced-LSTM) to 13.01% with our proposed ASR-Enhanced-Fused model.

The limitations of this include the performance may degrade for highly non-stationary noises (e.g., sudden bursts, overlapping speakers) due to the fixed attention window ($Z = 5$). Future work will explore adaptive window sizing or hybrid attention mechanisms. Further, the model is trained on LibriSpeech (English), and its generalizability to low-resource languages with different phonetic structures is untested. To address this in the future, transfer learning with limited labelled data could be investigated.

Data availability

The datasets generated used and analysed during the current study are available in the LibriSpeech and AURO-RA repository available at <https://www.openslr.org/12> and <http://aurora.hsnr.de/aurora-2.html>. The raw codes for attention-GRU are available at <https://github.com/NasirSaleem/Speech-Enhancement-ASR>.

Received: 28 June 2024; Accepted: 30 May 2025

Published online: 02 July 2025

References

1. Reza, S., Ferreira, M. C., Machado, J. J. M. & Tavares, J. M. R. S. A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model. *Expert Syst. Appl.* **215**, 119293 (2023).
2. El-Shafai, W. et al. Optical ciphering scheme for cancellable speaker identification system. *Comput. Syst. Sci. Eng.* **45**(1), 563–578 (2023).
3. Passos, L. A., Papa, J. P., Hussain, A. & Adeel, A. Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids. *Neurocomputing* **527**, 196–203 (2023).
4. Windowing, F. F. T. Research article speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation. *Res. J. Appl. Sci. Eng. Technol.* **6**(6), 1081–1087 (2013).
5. Jannu, C., & Vanambathina, S.D. Weibull and nakagami speech priors based regularized nmf with adaptive wiener filter for speech enhancement. *Int. J. Speech Technol.* 1–13 (2023).
6. Ephraim, Y. & Malah, David. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **32**(6), 1109–1121 (1984).
7. Chen, Bin & Loizou, Philipos C. A laplacian-based mmse estimator for speech enhancement. *Speech Commun.* **49**(2), 134–143 (2007).
8. Michelsanti, D. et al. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1368–1396 (2021).
9. Yong, X., Jun, D., Dai, L.-R. & Lee, C.-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(1), 7–19 (2014).
10. Wang, Z.-Q., Wang, P. & Wang, D. Complex spectral mapping for single-and multi-channel speech enhancement and robust asr. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 1778–1787 (2020).
11. Li, A., Liu, W., Zheng, C., Fan, C. & Li, X. 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 (2021).
12. Abdullah, Salinna, Zamani, Majid & Demosthenous, Andreas. Towards more efficient dnn-based speech enhancement using quantized correlation mask. *IEEE Access* **9**, 24350–24362 (2021).
13. Saleem, N., Mustafa, E., Nawaz, A. & Khan, A. Ideal binary masking for reducing convolutive noise. *Int. J. Speech Technol.* **18**, 547–554 (2015).
14. Bao, Feng & Abdulla, Waleed H. A new ratio mask representation for casa-based speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(1), 7–19 (2018).
15. Saleem, N., Khattak, M. I., Al-Hasan, M. & Qazi, A. B. On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks. *IEEE Access* **8**, 160581–160595 (2020).
16. Fan, C. et al. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 198–209 (2020).
17. Saleem, N. & Khattak, M. I. Deep neural networks for speech enhancement in complex-noisy environments. *Int. J. Interactive Multimed. Artif. Intell.* **6**(1), 84–91 (2020).
18. Sun, L., Du, J., Dai, L.-R., & Lee, C.-H. Multiple-target deep learning for lstm-rnn based speech enhancement. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 136–140. IEEE, (2017).
19. Yechuri, S. & Vanambathina, S. A nested u-net with efficient channel attention and d3net for speech enhancement. *Circ. Syst. Signal Process.* 1–21 (2023).
20. Chen, Jitong, Wang, Yuxuan, Yoho, Sarah E., Wang, DeLiang & Healy, Eric W. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *J. Acoust. Soc. Am.* **139**(5), 2604–2612 (2016).
21. Huang, X., Chen, H. & Wei, L. A two-stage frequency-time dilated dense network for speech enhancement. *Appl. Acoust.* **201**, 109107 (2022).
22. Pandey, A. & Wang, D. L. Dense cnn with self-attention for time-domain speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1270–1279 (2021).
23. Saleem, N. et al. U-shaped low-complexity type-2 fuzzy lstm neural network for speech enhancement. *IEEE Access* **11**, 20814–20826 (2023).
24. Liang, Ruiyu, Kong, F., Xie, Y., Tang, G. & Cheng, J. Real-time speech enhancement algorithm based on attention lstm. *IEEE Access* **8**, 48464–48476 (2020).
25. Pandey, A., & Wang, D.L. Dual-path self-attention rnn for real-time speech enhancement. arXiv preprint [arXiv:2010.12713](https://arxiv.org/abs/2010.12713), (2020).
26. Yechuri, S., & Vanambathina, S. A nested u-net with efficient channel attention and d3net for speech enhancement. *Circ. Syst. Signal Process.* 1–21 (2023).
27. Xu, X., & Hao, J. U-former: Improving monaural speech enhancement with multi-head self and cross attention. in *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 663–369. IEEE (2022).
28. Chen, J., Wang, Z., Tuo, D., Wu, Z., Kang, S., & Meng, H.. Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement. in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7857–7861. IEEE (2022).
29. Guochen, Y. et al. Dbt-net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 2629–2644 (2022).
30. Saleem, N. et al. Deepresgru: Residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition. *Knowl.-Based Syst.* **238**, 107914 (2022).
31. Wang, Y., Han, J., Zhang, T. & Qing, D. Speech enhancement from fused features based on deep neural network and gated recurrent unit network. *EURASIP J. Adv. Signal Process.* 1–19, 2021 (2021).
32. Yuan, W. Incorporating group update for speech enhancement based on convolutional gated recurrent network. *Speech Commun.* **132**, 32–39 (2021).
33. Seltzer, M. L. Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays. In *2008 Hands-Free Speech Communication and Microphone Arrays*, pp. 104–107. IEEE, (2008).
34. Wang, Z.-Q. & Wang, D. A joint training framework for robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(4), 796–806 (2016).
35. Han, K. et al. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(6), 982–992 (2015).

36. Li, F., Nidadavolu, P. S., & Hermansky, H. A long, deep and wide artificial neural net for robust speech recognition in unknown noise. in *Interspeech*, pp. 358–362. (2014).
37. M.L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp 7398–7402. IEEE (2013).
38. Liu, B., Nie, S., Zhang, Y., Ke, D., Liang, S., & Liu, W. Boosting noise robustness of acoustic model via deep adversarial training. in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5034–5038. IEEE (2018).
39. Chang, X., Zhang, W., Qian, Y., Le Roux, J., & Watanabe, S. End-to-end multi-speaker speech recognition with transformer. in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6134–6138. IEEE (2020).
40. Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pp. 1–5. IEEE (2017).
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. Attention is all you need. *Adv. Neural Inform. Process. Syst.* **30**, (2017).
42. Saleem, N., Gunawan, T. S., Dhahbi, S., & Bourouis, S. Time domain speech enhancement with cnn and time-attention transformer. *Digital Signal Process.* 104408 (2024).
43. Vanambathina, S. D., Nandyala, S., Jannu, C., Devi, J. S., Yechuri, S., & Parisae, V. Speech enhancement using u-net-based progressive learning with squeeze-tcn. In *International Conference on Advances in Distributed Computing and Machine Learning*, pp. 419–432. Springer (2024).
44. Parisae, V. & Bhavanam, S. N. Multi scale encoder-decoder network with time frequency attention and s-tcn for single channel speech enhancement. *J. Intell. Fuzzy Syst.* **46**(4), 10907–10907 (2024).
45. Nakadai, K., Hidai, K., Okuno, H. G., & Kitano, H. Real-time speaker localization and speech separation by audio-visual integration. in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, vol. 1, pp. 1043–1049. IEEE (2002).
46. Jannu, Chaitanya & Vanambathina, S. D. An overview of speech enhancement based on deep learning techniques. *Int. J. Image Graph.* **25**(01), 2550001 (2025).
47. Jannu, C. & Vanambathina, S. D. Multi-stage progressive learning-based speech enhancement using time-frequency attentive squeezed temporal convolutional networks. *Circ. Syst. Signal Process.* **42**(12), 7467–7493 (2023).
48. 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).
49. Jannu, C., & Vanambathina, S.D. Convolutional transformer based local and global feature learning for speech enhancement. *Int. J. Adv. Comput. Sci. Appl.* **14**(1), (2023).
50. Ullah, R. et al. End-to-end deep convolutional recurrent models for noise robust waveform speech enhancement. *Sensors* **22**(20), 7782 (2022).
51. Rajamani, S. T., Rajamani, K. T., Mallol-Ragolta, A., Liu, S., & Schuller, B. A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition. 6294–6298 (2021).
52. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. 5206–5210 (2015).
53. Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvett, D., Kelleher, H., Pearce, D., & S. Fabien. Evaluation of a noise-robust dsr front-end on aurora databases (2002).
54. Zheng, N. & Zhang, X.-L. Phase-aware speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(1), 63–76 (2018).
55. Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pp 4214–4217. IEEE (2010).
56. Beerends, J. G., Hekstra, A. P., Rix, A. W. & Hollier, M. P. Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part II: Psychoacoustic model. *J. Audio Eng. Soc.* **50**(10), 765–778 (2002).
57. Kim, J., El-Khamy, M. & Lee, J. Residual lstm: Design of a deep recurrent architecture for distant speech recognition. *Proc. Interspeech* **2017**, 1591–1595 (2017).
58. Xian, Y., Sun, Y., Wang, W. & Naqvi, S. M. Convolutional fusion network for monaural speech enhancement. *Neural Netw.* **143**, 97–107 (2021).
59. Lan, T. et al. Multi-scale informative perceptual network for monaural speech enhancement. *Appl. Acoustics* **195**, 108787 (2022).
60. Wang, Z., Zhang, T., Shao, Y. & Ding, B. Lstm-convolutional-blstm encoder-decoder network for minimum mean-square error approach to speech enhancement. *Appl. Acoustics* **172**, 107647 (2021).
61. Li, A., Yuan, M., Zheng, C. & Li, X. Speech enhancement using progressive learning-based convolutional recurrent neural network. *Appl. Acoustics* **166**, 107347 (2020).
62. Hasannezhad, M., Ouyang, Z., Zhu, W.-P., & Champagne, B. An integrated cnn-gru framework for complex ratio mask estimation in speech enhancement. pp. 764–768 (2020).
63. Westhausen, N. L., & Meyer, Bernd T. Dual-signal transformation lstm network for real-time noise suppression (2020).
64. Yanxin, Hu. et al. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *Proc. Interspeech* **2020**, 2472–2476 (2020).
65. Kim, J., El-Khamy, M., & Lee, J. T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. 6649–6653 (2020).
66. Zhang, Q., Nicolson, A., Wang, M., Paliwal, K. K. & Wang, C. Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 1404–1415 (2020).
67. Tan, K. & Wang, D. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. 6865–6869 (2019).
68. Pandey, A., & Wang, D.L. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. 6875–6879 (2019).
69. Tan, K. & Wang, D. A convolutional recurrent neural network for real-time speech enhancement. **2018**, 3229–3233 (2018).
70. Shah, N., Patil, H.A., & Soni, M. H. Time-frequency mask-based speech enhancement using convolutional generative adversarial network. 1246–1251 (2018).
71. Bhardwaj, V. et al. Automatic speech recognition (asr) systems for children: A systematic literature review. *Appl. Sci.* **12**(9), 4419 (2022).
72. Rahman, A. et al. *Advancement and Challenges* (IEEE Access, Arabic speech recognition, 2024).
73. Hadwan, M., Alsayadi, H. A., & Al-Haggee, S. An end-to-end transformer-based automatic speech recognition for qur'an reciters. *Comput. Mater. Continua.* **74**(2), (2023).

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/607/46. The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial sup-

port (QU-APC-2025). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R747), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Funding

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/607/46. The researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025). The Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R747), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.F.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025