



MSFEYNet: A Hybrid Convolution Transformer Network with Spectral Feature Extraction & MSFEB Modelling for Single-Channel Speech Enhancement

Silpa Peethala, Sunny Dayal Vanambathina

Abstract: This paper presents MSFEYNet, a novel dual-decoder U-Net architecture for single-channel speech enhancement that integrates convolutional and Transformer blocks to exploit local spectral information and long-range contextual dependencies jointly. The encoder extracts hierarchical multi-scale spectral representations, while two dedicated decoders simultaneously reconstruct the clean speech magnitude and phase spectra, enabling accurate speech recovery. To enhance feature representation, the proposed architecture incorporates three key components: the Triple Path Fused Attention (TPFA) module for efficient long-range axial spectral modelling, the Large Group Large Kernel Attention (LGLKA) module for capturing fine-grained local axial structures, and an Enhanced Feed Forward Network (EFFN) with dynamic axial-aware feature reweighting to improve contextual feature learning. Furthermore, a Multi-Scale Feature Extraction Block (MSFEB) extracts both fine-grained local features and broad contextual information using convolutional filters with varying receptive fields, thereby enhancing multi-scale spectral-temporal feature representation. Extensive experiments on two datasets demonstrate that MSFEYNet achieves robust, well-balanced performance across a comprehensive set of objective evaluation metrics through the synergistic integration of the dual-decoder framework. The transformer-based attention mechanism and Multi-scale Feature Extraction Block (MSFEB) effectively capture local and global contextual dependencies, leading to significant improvements in speech intelligibility and perceptual quality. Experimental results further demonstrate that MSFEYNet consistently outperforms contemporary state-of-the-art speech enhancement methods, achieving substantial gains in STOI (Short-Time Objective Intelligibility) and PESQ (Perceptual Evaluation of Speech Quality) while maintaining excellent generalisation across diverse acoustic conditions and varying noise levels.

Keywords: Y-Net, MSFEB, STOI, PESQ.

Nomenclature:

ACK: Acknowledgement
QC: Quality Check
AI: Artificial Intelligence
TF: Time-Frequency

LSTM: Long Short-Term Memory
SE: Speech Enhancement
LGLKA: Large Group Large Kernel Attention
TPFA: Triple Path Fused Attention
EFFN: Enhanced Feed Forward Network
MSFEB: Multi-Scale Feature Extraction Block
CNNs: Convolutional Neural Networks
TPFA: Triple-Path Fused Attention
IN: Instance Normalization
FFN: Feed-Forward Network
EFFN: Enhanced Feed Forward Network
SNR: Signal-To-Noise Ratio
STOI: Short-Time Objective Intelligibility
PESQ: Perceptual Evaluation of Speech Quality

I. INTRODUCTION

Various types of environmental interference can significantly degrade the performance of telecommunication systems, automatic speech recognition, and hearing aids in real-world scenarios. Monoaural speech enhancement (SE) has therefore become increasingly important, aiming to recover clean speech from an acoustically contaminated mixture to improve both quality and intelligibility [1]. Existing SE network architectures fall into two main types. The first is the U-Net-based multi-scale network, where each layer typically incorporates convolutional neural networks (CNNs), long short-term memory (LSTM), self-attention mechanisms, or their combinations. U-Net emphasises leveraging multi-scale information to help separate speech from background noise, thereby improving SE performance [2] [3]. The second type follows the classic encoder-enhancement-decoder backbone, which generally operates on single-scale features. Here, the encoder and decoder are built with CNNs, while the enhancement module may be implemented with xLSTMs [4], CNNs [5] [6], Transformers [7], or Mamba blocks [8], [9]. The underlying rationale is that the CNNs in the encoder and decoder leverage their inherent inductive bias toward local patterns and efficient weight sharing to extract fine-grained spectral-temporal features, such as the harmonic structure of target speech and narrowband noise signatures, which are crucial for isolating speech signals from complex acoustic interference. Meanwhile, the enhancement module models long-range dependencies across extended speech segments, compensating for the limited receptive field of pure CNNs, which often struggle to capture contextual information such as phoneme coherence and prosodic continuity.

Manuscript received on 06 August 2026 | First Revised Manuscript received on 14 August 2026 | Second Revised Manuscript received on 22 August 2026 | Manuscript Accepted on 15 September 2026 | Manuscript published on 30 September 2026.

*Correspondence Author(s)

Silpa Peethala*, School of Electronics Engineering,,VIT-AP University, Amaravati (Andhra Pradesh), India. Email ID: sipla.23phd7181@vitap.ac.in. ORCID ID: 0009-0002-9823-6013

Dr Sunny Dayal Vanambathina, School of Electronics Engineering,,VIT-AP University, Amaravati (Andhra Pradesh), India. Email ID: sunny.dayal@vitap.ac.in.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>





The dual-path model has also been extended to the time-frequency (TF) domain, demonstrating superior efficiency and accuracy in SE tasks relative to its time-domain counterpart. TF-domain implementations typically comprise two components: a sub-band network that aggregates local information within individual frequency sub-bands, and a full-band network that captures global contextual dependencies across the entire frequency range. These networks can be instantiated using diverse structures, including Transformers and their variants [10], [11], [12], [13], [14], or Mamba blocks [15], [19], [20], [21].

Diverging from the aforementioned network architectures and modules, we propose MSFEYNet, a novel dual-decoder U-Net architecture featuring a hybrid convolution-Transformer design, named for its distinctive Y-shaped topology. MSFEYNet's encoder extracts general spectral representations, while one decoder reconstructs the clean magnitude spectrum and the other recovers the clean phase spectrum. Each MSFEYNet layer integrates multiple convolutional and Transformer blocks, enabling joint capture of local details and long-range contextual information. This hybrid design synergises the complementary strengths of CNNs and Transformers, overcoming the limitations of using either in isolation for speech modelling. Notably, both convolutional and Transformer blocks in MSFEYNet follow the Meta-Former structure [16], which generalises the standard Transformer. Meta-Former comprises two core components: a token mixer and a channel mixer, corresponding to the self-attention mechanism and the feed-forward network in the standard Transformer, respectively.

The TF-domain dual-path model consists of two sub-networks, each independently capturing contextual information along one axis of the feature map. This structure, however, supports only indirect cross-axis information propagation across the entire feature map, complicating efficient computation of globally integrated attention. A Triple-Path Fused Attention (TPFA) mechanism integrates sub-band self-attention, full-band self-attention, and local pattern modelling. By explicitly leveraging the complementary information provided by these three feature types, a dedicated feature-fusion block integrates them. Serving as the token mixer of the Meta-Former [16], TPFA captures long-range axial dependencies in speech spectra while retaining sensitivity to local details. To complete the Transformer block, we further propose an Enhanced Feed-Forward Network (EFFN) as the channel mixer, augmenting the standard feed-forward network with a dynamic weight-adjustment module operating across time, frequency, and channel dimensions. As a result, the proposed Transformer block enables more effective spectral feature modelling, improving SE performance. Inspired by the design philosophy of the TF-domain dual-path model, we extend a similar axial modelling strategy to CNNs and propose a Large Group Large-Kernel Attention (LGLKA) module. This module is built on three key considerations. First, a large kernel expands the convolution's receptive field and enhances its modelling capability. Second, the large-kernel 2D convolution is decomposed into two separate 1D convolutions operating independently along the time and frequency axes. This decomposition significantly reduces computational complexity while effectively extracting axial-

wise local features. Third, we employ grouped 1D convolutions with varying kernel sizes to capture multi-scale representations. LGLKA serves as the token mixer in the MSFEYNet convolutional block, while its channel mixer remains identical to that in the Transformer block. Building on the above designs, we integrate convolutional and Transformer blocks within a dual-decoder U-Net architecture to propose MSFEYNet. With a topology and components that explicitly account for the distinctive properties of speech spectra, MSFEYNet delivers an effective and practical solution for SE under diverse background noises and challenging acoustic conditions.

The main contributions of this paper are summarised as follows:

- We propose MSFEYNet, a dual-decoder U-Net that jointly enhances magnitude and phase spectra while leveraging multi-scale feature extraction.
- We introduce TPFA, LGLKA, and EFFN within a unified MetaFormer-style design to model long-range axial dependencies, local axial structures, and adaptive cross-axis feature interactions.
- Extensive experiments on four benchmark datasets show that YNet achieves strong and well-balanced performance across denoising and dereverberation tasks.

The remainder of this paper is organised as follows. Section 2 formulates the speech enhancement task. Section 3 details the proposed MSFEYNet architecture. Section 4 presents the experimental setup, followed by results and analysis in Section 5. Finally, Section 6 provides concluding remarks.

II. PROBLEM FORMULATION

A. Problem Formulation

We focus on the problem formulation for speech denoising, as it is the most representative task in monaural speech enhancement. Under the additive noise assumption, the relationship between noisy speech $x[n]$, clean speech $s[n]$, and noise $d[n]$ is expressed as:

$$x[n] = s[n] + d[n], \quad (1)$$

where n denotes the discrete-time index; applying the short-time Fourier transform (STFT) to Eq. (1) yields its time-frequency (TF) domain representation:

$$X_{tf} = S_{tf} + D_{tf}, \quad (2)$$

Where $X_{t,f}$, $S_{t,f}$ and $D_{t,f}$ represent the TF spectra of noisy speech, clean speech and noise, respectively, at frame index t and frequency index f . Omitting the subscript $\{t,f\}$ in Eq. (2), we obtain the corresponding matrix formulation.

$$X_m e^{jX_p} = S_m e^{jS_p} + D_m e^{jD_p} \quad (3)$$

where j is the imaginary unit; X_m and X_p denote the magnitude and phase of the noisy spectrum X ; S_m , S_p and D_m , D_p are the magnitude and phase matrices of clean speech and noise, respectively, defined analogously. All six matrices are of size $T \times F$, with T representing the total number of frames and F the total number of frequency bins.

The goal of speech denoising in the TF domain is to recover the clean speech S from the noisy speech X . In practice, given the noisy magnitude X_m and noisy phase X_p as inputs,

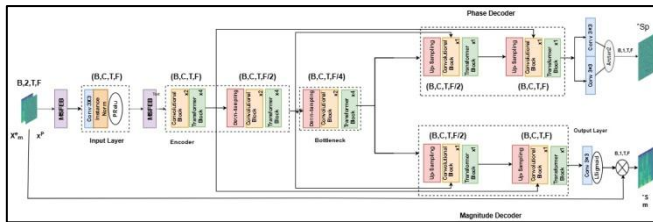


the objective is to estimate the clean magnitude $\mathbf{S}_m$ and clean phase $\mathbf{S}_p$. However, rather than feeding the direct concatenation $[\mathbf{X}_m, \mathbf{X}_p]$ into the proposed Y-Net, we instead concatenate the compressed magnitude $\mathbf{X}_{cm}$ with $\mathbf{X}_p$ as the model input. This corresponds to a power-law compression of the magnitude spectrum; a technique widely used in speech enhancement to better align with human auditory perception. Following the same configuration as in Section II, we set the compression factor $c = 0.3$. In other speech enhancement tasks, such as dereverberation, the goal is to reconstruct target speech from differently degraded versions.

III. MSFE-YNet

Fig. 1 illustrates the overall architecture of the proposed MSFE-YNet. The network adopts a dual-decoder U-Net framework that jointly estimates the enhanced magnitude and phase spectra from the noisy speech input. To improve the quality of the learned spectral representations, the magnitude input is first processed by a Multi-Scale Feature Extraction Block (MSFEB), which captures fine-grained local details and broad contextual information through multiple convolutional kernels with different receptive fields. This enables the network to learn rich spectral representations while preserving important speech structures.

After that, the information is processed by an input layer that comprises a standard convolution with a 3×3 kernel (denoted as Conv 3×3 in Fig. 1), followed by Instance Normalisation (IN) and a PReLU activation function. This layer expands the input feature dimension to C channels. The resulting spectral embedding is subsequently passed through a dual-decoder U-Net. The shared encoder and bottleneck first extract general spectral features, which are then passed to two separate decoders: one dedicated to capturing magnitude features and the other to reconstructing phase features. Each decoder is followed by a dedicated output layer that generates the enhanced magnitude and phase spectra, respectively. Specifically, the phase decoder's output layer consists of two parallel 3×3 convolutions and an Arctan2 [3] activation, whereas the magnitude decoder's output layer employs a single 3×3 convolution followed by an LSigmoid activation.



[Fig.1: MSFE-YNet Architecture [17]]

The LSigmoid function is defined as

$$\text{LSigmoid}(k) = \frac{\beta}{1 + e^{-\alpha k}} \quad (4)$$

where k is the input, $\beta = 2.0$, and α is a trainable frequency-dependent parameter that enables the activation shape to adapt across frequency bands. The magnitude decoder's output layer generates a time-frequency (TF) mask. $\hat{\mathbf{M}} \in \mathbb{R}^{T \times F}$ (the batch dimension is omitted for simplicity). The enhanced magnitude spectrum $\hat{\mathbf{S}}_m \in \mathbb{R}^{T \times F}$ is subsequently obtained via the following masking and decompression operation.

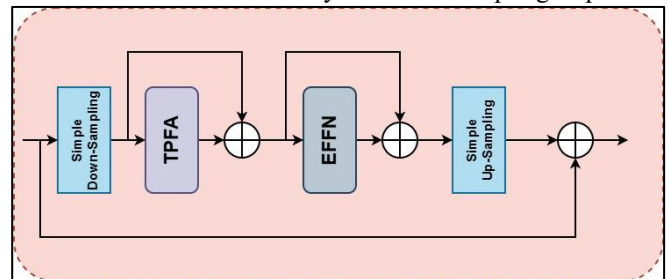
$$\hat{\mathbf{S}}_m = (\mathbf{X}_{cm} \otimes \hat{\mathbf{M}}^c) \backslash c \quad (5)$$

where $\otimes$ denotes element-wise multiplication, $\mathbf{X}_{cm}$ represents the compressed magnitude of the noisy spectrum, and $c = 0.3$ is the compression factor. A down-sampling layer is inserted before the second encoder layer and the bottleneck layer. This layer comprises a convolution with a 3×3 kernel and a stride of (1, 2), followed sequentially by instance normalisation and a PReLU activation. It halves the spatial resolution of the feature map exclusively along the frequency axis. Symmetrically, an up-sampling layer is placed before each decoder layer. This up-sampling layer is built around a sub-pixel convolution layer with a 1×3 kernel, again followed by instance normalisation and a PReLU activation function.

Overall, the proposed YNet combines an MSFEB-based multi-scale feature extraction block, a shared encoder-bottleneck, and dual magnitude-phase decoders within a unified framework. By jointly modelling magnitude and phase information while integrating convolutional and Transformer-based representations, the proposed architecture effectively suppresses environmental noise. It reconstructs high-quality speech with improved perceptual quality and intelligibility.

A. Transformer Block

The Transformer block, as illustrated in Fig. 2, consists of the following components in sequential order: a simple down-sampling module, the TPFA module with a residual connection, the EFFN module with a residual connection, and a simple up-sampling module. Furthermore, a global residual skip connection directly links the block's input to its output. Both the simple down-sampling and up-sampling operations are performed exclusively along the time axis. The down-sampling operation computes a weighted average of every two consecutive frames using two learnable scalar weights normalised by a softmax function, halving the temporal length to reduce the computational cost of subsequent modules. The up-sampling operation then duplicates each frame to recover the original temporal resolution. The global skip connection thereby compensates for any potential information loss introduced by the down-sampling step.



[Fig.2: Transformer Block [33]]

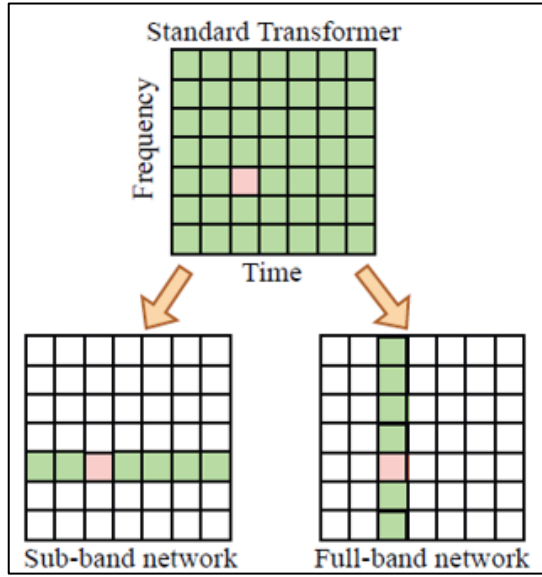
B. Self-Attention Mechanism

The TF-domain dual-path model introduces a lightweight self-attention mechanism for speech modelling that follows the same principle as axial attention [17], [32]. Fig. 3 illustrates the attention mechanisms in the standard Transformer and the TF-domain dual-path model. Unlike the standard Transformer, which requires each token to attend

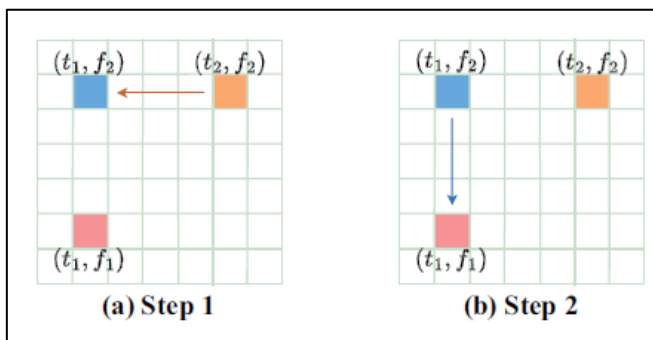


to all other tokens, the TF-domain dual-path model processes attention separately along each axis via two independent Transformer blocks: the sub-band network and the full-band network. Given a feature map of dimensions $B \times C' \times T' \times F'$, where B denotes the batch size, C' the channel dimension, and T' and F' the temporal and frequency dimensions, respectively, the sub-band network and the full-band network apply F' and T' independent self-attention modules along the time and frequency axes, respectively. Specifically,

When $B = 1$, the sub-band network consists of F' parallel standard Transformers, each processing T' tokens of length C' ; similarly, the full-band network consists of T' parallel standard Transformers, each processing F' tokens of length C' .



[Fig 3: Attention Mechanisms in the Standard Transformer (top) and the TF-Domain Dual-Path Model: The Sub-band Network (Bottom Left) and the full-band Network (Bottom Right). The Target Token (pink) Attends to other Tokens (Green) During Attention Computation [32]]



[Fig.4: Mechanism of Information Propagation in Axial Attention: Transmitting Information from the Token at (t_2, f_2) to the Token at (t_1, f_1) through (a) Step 1 and (b) Step 2 [32]]

While the axial attention mechanism can capture contextual information along both axes of the input feature map, it still lacks direct connections between tokens that are not aligned on the same row or column. To illustrate this, Fig. 4 shows how information from a token at position (t_2, f_2) (orange) can reach the target token at (t_1, f_1) (red) through an intermediary

token at (t_1, f_2) (blue) in two steps. These steps correspond to successive operations of the sub-band and full-band networks

- Step 1 (Sub-band network): The token at (t_2, f_2) passes its information to the intermediary token at (t_1, f_2) via self-attention along the time axis.
- Step 2 (Full-band network): The intermediary token at (t_1, f_2) propagates the aggregated information to the target token at (t_1, f_1) via self-attention along the frequency axis.

Through this two-step process, information from any token can ultimately reach any other token, enabling global information exchange across the entire feature map in principle, much like standard self-attention. However, because each axial attention step operates only along a single axis, information exchange is inherently indirect and slower than in standard self-attention. This limits the model's ability to capture dense local and global contextual relationships efficiently. Furthermore, axial attention alone does not incorporate local spatial patterns, which are crucial for modelling structured data like speech spectra. To address these limitations, we propose the TPFA module, depicted in Fig. 5. TPFA fuses the two axial attention steps into a single unified block and further incorporates local features extracted via convolutional operations. This design mitigates the inefficiency of multi-step information propagation and enriches the representation with local inductive biases. Moreover, it consolidates the two separate Transformer blocks of the dual-path model into one unified Transformer block in Fig. 2. As a result, TPFA achieves more efficient attention computation and more expressive feature fusion.

C. C.TPFA

Let $\mathbf{X} \in \mathbb{R}^{B \times C' \times T' \times F'}$ denote the input and output tensors of the TPFA module, respectively. As illustrated in Fig. 5, TPFA's forward pass is defined by the following steps.

$$\mathbf{X}_t, \mathbf{X}_f = \text{Split}(\text{LayerNorm}(\mathbf{X})\mathbf{W}_{QKV}) \quad (6)$$

$$\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t = \text{Split}(\mathbf{X}_t) \quad (7)$$

$$\mathbf{Q}_f, \mathbf{K}_f, \mathbf{V}_f = \text{Split}(\mathbf{X}_f) \quad (8)$$

$$\mathbf{I}_0 = \text{DWC}_{\text{onv}3 \times 3}(\text{Conv}(\mathbf{V}_t, \mathbf{V}_f)) \quad (9)$$

$$\mathbf{I}_t = \text{MHSA}_T(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) \quad (10)$$

$$\mathbf{I}_f = \text{MHSA}_F(\mathbf{Q}_f, \mathbf{K}_f, \mathbf{V}_f) \quad (11)$$

$$\mathbf{I}_{\text{attn}} = \text{Concat}(\mathbf{I}_t, \mathbf{I}_f) \quad (12)$$

$$\mathbf{I}'_{\text{attn}}, \mathbf{I}'_0 = \mathbf{F}(\mathbf{I}_{\text{attn}}, \mathbf{I}_0) \quad (13)$$

$$\mathbf{Y} = (\mathbf{I}'_{\text{attn}} + \mathbf{I}'_0) \mathbf{W}_{\text{proj}} \quad (14)$$

Where $\mathbf{W}_{QKV} \in \mathbb{R}^{C' \times 3C'}$ and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{C' \times C'}$ are linear projection matrices; "Split" and "Concat" denote channel splitting and concatenation; $\text{DWConv}_{3 \times 3}$ denotes depth-wise convolution with a 3×3 kernel; MHSA_T and MHSA_F are multi-head self-attention modules applied along the time and frequency axes, respectively; and $\mathbf{F}$ denotes the fusion block (described later). For brevity, tensor reshaping operations are omitted unless otherwise specified. For MHSA_T , we first reshape $\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t$ into $\mathbf{Q}'_t, \mathbf{K}'_t, \mathbf{V}'_t \in \mathbb{R}^{(BF') \times T' \times C'}$. The operation is then defined as



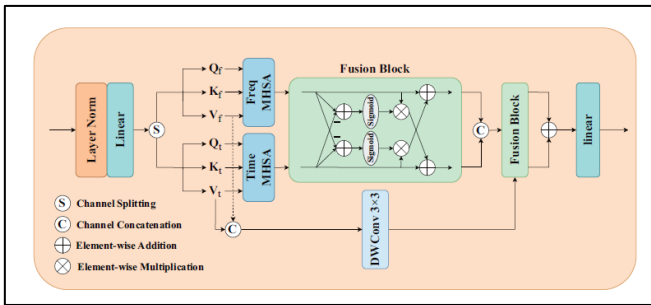
$$\text{head}_t = \text{softmax}\left(\frac{Q'_{ti} [K'_{ti}]^T}{\sqrt{d}}\right) V'_{ti}, i = 1, \dots, h \quad (15)$$

$$\text{MultiHead} = \text{Concat}(\text{head}_i, \text{head}_h) \quad (16)$$

where $Q'_{ti}, K'_{ti}, V'_{ti} \in \mathbb{R}^{(BF) \times T \times C'}$ represent the query, key and value of the i^{th} head, $d=C'/h$ is the per-head dimension, and h is the number of attention heads. In Eq. (15), $[\cdot]^T$ denotes transposition over the last two dimensions. The formulation of MHSA_F is analogous, except that Q_f, K_f, V_f are reshaped to $Q'_f, K'_f, V'_f \in \mathbb{R}^{(BT) \times F' \times C'}$. To set $h=2$ for both MHSA_T and MHSA_F. The proposed fusion block F is defined as

$$\{\hat{X}\}_1 = \sigma(X_2 - X_1) \otimes X_2 + X_1 \quad (17)$$

$$\{\hat{X}\}_2 = \sigma(X_1 - X_2) \otimes X_1 + X_2 \quad (18)$$



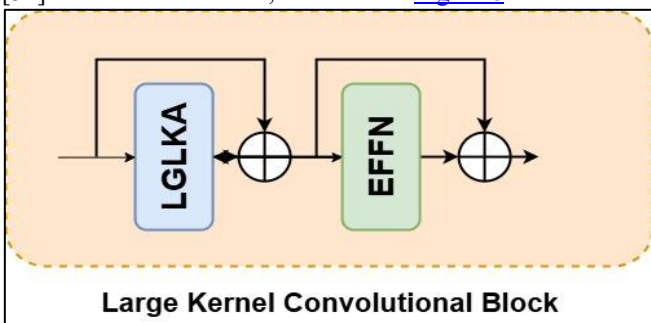
[Fig.5: The TPFA Module [33]]

where σ denotes the sigmoid activation function, X_1 and X_2 are the two inputs, and $\hat{X}_1$ and $\hat{X}_2$ are the corresponding outputs. This block first computes the residual between the two inputs, passes it through a sigmoid gate to selectively activate complementary features, and then integrates each input with the gated complementary information from the other.

In Fig. 5, two such fusion blocks are applied sequentially: the 1st integrates the two axial attention outputs (I_t and I_f), and the second fuses the resulting attended features with the local features I_0 extracted by the depth-wise convolution DWConv_{3x3}. Consequently, the TPFA module captures long-range axial structure across the entire feature map while retaining sensitivity to local details. In terms of model complexity, the fusion blocks involve only simple arithmetic operations and introduce no additional parameters, while the DWConv_{3x3} layers contribute only a small number of parameters.

D. LKCB

The large-kernel convolutional block has two sub-blocks [34]: EFFN and LGLKA, as shown in Figure 6.



[Fig.6: LKCB [34]]

E. EFFN

Where X denotes the input feature, and Y represent the corresponding output; the forward pass of the EFFN module is shown in Figure 7.

$$X_c = \text{SiLU}(\text{LayerNorm}(x)W_1) \quad (19)$$

$$X_{c1}, X_{c2} = \text{Split}(\text{DW Conv}_{3 \times 3}(X_c)) \quad (20)$$

$$X_{g1}, X_{g2} = \text{Split}(x_{c1} \otimes x_{c2}) \quad (21)$$

$$X'_{g1} = X_{g1} \otimes \text{Conv}_{1 \times 1}(\text{Pooling}_F(X_{g1})) \quad (22)$$

$$X'_{g2} = X_{g2} \otimes \text{Conv}_{1 \times 1}(\text{Pooling}_T(X_{g2})) \quad (23)$$

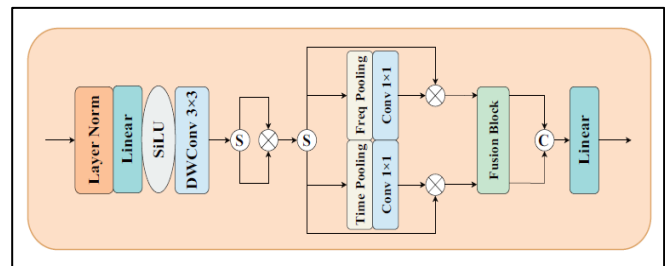
$$Y = \text{Concat}(F(X'_{g1}, X'_{g2}))W_2 \quad (24)$$

$$X_c = \text{SiLU}(\text{LayerNorm}(x)W_1) \quad (19)$$

$$X_{c1}, X_{c2} = \text{Split}(\text{DWConv}_{3 \times 3}(X_c)) \quad (20)$$

$$X_{g1}, X_{g2} = \text{Split}(X_{c1} \otimes X_{c2}) \quad (21)$$

$$X'_{g1} = X_{g1} \otimes \text{Conv}_{1 \times 1}(\text{Pooling}_F(X_{g1})) \quad (22)$$



[Fig.7: Enhanced Feed Forward Network (EFFN) Model [33]]

Here, W_1 and W_2 denote linear projection matrices that expand and compress the channel dimension by factors of 4 and 2, respectively. SiLU denotes the SiLU activation function, while Split and Concat denote channel splitting and concatenation, respectively. The symbol $\otimes$ represents element-wise multiplication. DWConv_{3x3} denotes a depthwise convolution with a 3x3 kernel, where conv_{1x1} denotes a pointwise convolution. Furthermore, Pooling_F and Pooling_T represent average pooling operations performed along the frequency and time dimensions, respectively, and F denotes the feature fusion block. Compared with the conventional Transformer [33] feed-forward network (FFN), which consists primarily of layer normalisation, a nonlinear activation function and two linear projection layers, the proposed Enhanced Feed Forward Network (EFFN) introduces four principal architectural enhancements. First, the DWConv 3 x 3 operation captures fine-grained local spatial representations while preserving computational efficiency; second, the element-wise multiplication $X_{c1} \otimes X_{c2}$ generates a highly nonlinear feature representation, thereby enriching the network's expressive capacity. Third, the axial average pooling operations, Pooling_F and Pooling_T, aggregate contextual information along the frequency and temporal dimensions, respectively. A Conv 1 x 1 layer subsequently processes the resulting features to model inter-channel dependencies, producing adaptive attention weight maps that dynamically recalibrate the feature representations X_{g1} and X_{g2} . Finally, the fusion block integrates the recalibrated features to generate a more discriminative representation for subsequent processing.



F. LGLKA

From Figure 8, given an input X and the corresponding output Y , the forward propagation of the Large Global Large Kernel Attention (LGLKA) [34] module within the convolutional block is formulated as follows. Here, SiLU denotes the Sigmoid Linear Unit activation function, whereas Split and Concat represent the channel splitting and channel concatenation operations, respectively. The attention map X_{attn} in equation (31) serves as a dynamic weighting tensor that adaptively recalibrates the feature representation X_s as defined in equation (25). Specifically, X_{attn} is generated by first applying a $\text{DWConv} 3 \times 3$ layer to capture fine-grained local spatial information, followed by four groups of one-dimensional depthwise convolutions with progressively increasing kernel sizes of 6, 10, 14, and 18. These multi-scale convolutions effectively capture long-range contextual dependencies and axial patterns along both the temporal and frequency dimensions. The $\text{Conv} 1 \times 1$ layers in equations (25) and (32) preserve the channel dimension while enabling efficient channel-wise feature interaction, consistent with the MetaFormer design paradigm. The proposed LGLKA module is based on the large-kernel attention mechanism, which combines the locality modelling capability of convolutions with the contextual representation strength of attention mechanisms. By incorporating multi-scale parallel axial convolutions with expanded receptive fields, the LGLKA module effectively models both local and global spectral-temporal dependencies, providing a computationally efficient complement to conventional axial attention for speech spectrum representation learning.

$$X_s = \text{SiLU}(\text{Conv}_{1 \times 1}(\text{LayerNorm}(X))) \quad (25)$$

$$X_{1:4} = \text{Split}(\text{DWConv}_{3 \times 3}(X_s)) \quad (26)$$

$$X'_1 = \text{DWConv}_{6 \times 1}(\text{DWConv}_{1 \times 6}(X_1)) \quad (27)$$

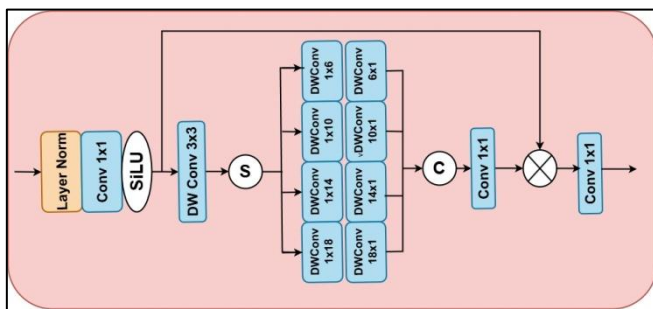
$$X'_2 = \text{DWConv}_{10 \times 1}(\text{DWConv}_{1 \times 10}(X_2)) \quad (28)$$

$$X'_3 = \text{DWConv}_{14 \times 1}(\text{DWConv}_{1 \times 14}(X_3)) \quad (29)$$

$$X'_4 = \text{DWConv}_{18 \times 1}(\text{DWConv}_{1 \times 18}(X_4)) \quad (30)$$

$$X_{\text{attn}} = \text{Conv}_{1 \times 1}(\text{Concat}(X'_{1:4})) \quad (31)$$

$$Y = \text{Conv}_{1 \times 1}((X_s \otimes (X_{\text{attn}}))) \quad (32)$$



[Fig.8: The LGLKA Module [34]]

G. Multi-scale Feature Extraction Block

Inspired by [18], the MSFEB gathers contextual information at different scales and is built on the widely used ResNet block.

IV. ABLATION STUDY OF THE MSFEYNET MODEL

To investigate the contribution of each architectural component, we constructed five network variants by

progressively incorporating the proposed modules. We compare each variant's performance to demonstrate the effectiveness of MSFEB, the Transformer blocks, and the dual-decoder framework.

MSFEB: This variant employs only the Multi-Scale Feature Extraction Block (MSFEB) for extracting rich local spectral features from the input speech. It does not include the Transformer blocks or the dual-decoder architecture. This model serves as the baseline to evaluate the contribution of the proposed feature extraction module.

Transformer as U-Net: In this configuration, the conventional U-Net blocks are replaced with Transformer blocks in the encoder, bottleneck, and decoder. This model evaluates Transformer-based feature learning without incorporating MSFEB or the proposed dual-decoder strategy.

MSFEB + Central Layers (Transformer + D3Net): This variant combines MSFEB with Transformer and D3Net modules only in the bottleneck (central layers) of the network. The encoder and decoder remain unchanged. This experiment investigates the effectiveness of enhancing the latent representation through global context modelling and dense feature aggregation while keeping the rest of the architecture relatively simple.

Transformer U-Net + Central Layers (Transformer+U-Net): This model extends the Transformer-based U-Net by integrating additional Transformer and U-Net models in the bottleneck. It evaluates whether strengthening the central representation improves speech performance over the standard Transformer U-Net.

Proposed Model MSFEYnet (MSFEB + Transformer + Dual Decoder): The proposed MSFEY-Net integrates all architectural components, including the MSFEB, Transformer blocks throughout the network, and a dual-decoder architecture with separate magnitude and phase decoders. The MSFEB effectively captures rich local multi-scale features, while the Transformer blocks model long-range contextual dependencies in the time-frequency domain. In addition, the dual decoder architecture jointly estimates the speech magnitude and phase components, enabling more accurate speech reconstruction. The synergistic integration of these components significantly enhances speech intelligibility and perceptual quality while maintaining computational efficiency, making MSFEYNet well-suited for real-time speech enhancement applications.

A. Datasets

We evaluated the proposed MSFEYNet using the English subset of the Mozilla Common Voice corpus [31], a large-scale, publicly available speech dataset collected from volunteer contributors worldwide. The Common Voice corpus contains approximately 1.65 million utterances from more than 84,000 speakers, covering diverse accents, speaking styles, and recording environments. For this study, we selected the English corpus to construct the training and validation datasets. Specifically, the training set consists of 50,000 clean speech utterances, while the validation set contains 4,000 utterances, both randomly sampled from the English Common Voice corpus.

These clean speech recordings were subsequently mixed with environmental noise at



different signal-to-noise ratio (SNR) levels to generate the noisy speech samples used for model training and validation.

Short-Time Objective Intelligibility (STOI) [29]: STOI is a widely used objective metric for evaluating speech intelligibility. It measures the correlation between the enhanced speech signal and its corresponding clean reference over short time segments, estimating the intelligibility of processed speech. STOI is particularly effective for assessing the impact of noise reduction and speech enhancement algorithms, as its scores strongly correlate with subjective listening tests.

Perceptual Evaluation of Speech Quality (PESQ) [30]: PESQ is a standardised, widely adopted objective metric for assessing the Perceptual quality of enhanced speech. It estimates speech quality by comparing the processed speech signal with its corresponding clean reference while accounting for perceptually relevant distortions introduced during the enhancement process. PESQ considers factors such as speech distortion, background noise, and signal degradation, providing an objective score that strongly correlates with subjective human listening evaluations.

B. Objective Comparison of Baseline Models Under Seen Noises

From [Tables 2](#) and [3](#), in the seen-noise evaluation, WAVUNET [22] achieves an average of 85.45% STOI [29] and 2.58 PESQ [30], demonstrating that waveform-domain learning effectively preserves speech intelligibility. However, its limited ability to capture long-range contextual information restricts perceptual speech quality under complex acoustic conditions. ICCRN [23] improves the average performance to 86.27% STOI and 2.64 PESQ by exploiting complex-valued spectral representations that jointly model speech magnitude and phase information. IUNNB [24] further increases the average STOI and PESQ to 87.25% and 2.74, respectively, indicating that enhanced feature interaction contributes to more accurate speech reconstruction. MSP [25] achieves 88.42% STOI and 2.85 PESQ by effectively learning multi-scale spectral representations. SMWETCW [26] further improves the average performance to 89.33% STOI and 2.94 PESQ, and 3.06 PESQ, demonstrating that transformer-based contextual modelling substantially improves speech intelligibility and perceptual quality. LMAT [7] achieves 92.21% STOI and 3.17 PESQ through efficient long-range dependency modelling with reduced computational complexity. PrimK-Net [6] further improves the performance to 93.25% STOI and 3.36 PESQ, owing to its enhanced multi-scale feature aggregation strategy.

The proposed MSFEYNet consistently achieves the highest performance, recording an average of 94.23% STOI and 3.45 PESQ across all evaluated SNR levels. The superior performance is attributed to the effective integration of the Multi-Scale Feature Extraction Block (MSFEB), Transformer blocks, and the dual-decoder architecture, which jointly exploit local spectral features, long-range contextual dependencies, and simultaneous magnitude-phase estimation. Consequently, the proposed network produces more natural speech with higher intelligibility while effectively suppressing background noise.

C. Objective Comparison of Baseline Models Under Unseen Noises

[Tables 4](#) and [5](#) further demonstrate the robustness and generalisation capability of the proposed model in unseen-noise experiments. WAVUNET [22] achieves average scores of 84.31% STOI and 2.48 PESQ, indicating reduced performance under previously unseen noise conditions. ICCRN [23] improves the results to 85.28% STOI and 2.55 PESQ by leveraging complex domain speech enhancement. IUNNB [24] obtains 86.31% STOI and 2.65 PESQ. While MSP [25] achieves 87.49% STOI and 2.77 PESQ, demonstrating gradual improvement through enhanced feature representation. SMWETCW [26] records 88.36% STOI and 2.83 PESQ, followed by TDSTN [27] which reaches 89.43% STOI and 2.94 PESQ through effective transformer-based contextual modelling. LMAT [7] further improves the performance to 90.42% STOI and 3.06 PESQ, whereas PrimK-Net [6] achieves 91.30% STOI and 3.25 PESQ, reflecting strong robustness against unseen acoustic environments.

The proposed MSFEYNet outperforms all competing approaches with an average 93.13% STOI and 3.36 PESQ. Across all SNR levels, particularly under challenging -5dB and -2dB conditions, the proposed model consistently preserves speech intelligibility and perceptual quality more effectively than the baseline methods. These improvements show that combining MSFEB Transformer-based contextual learning with the dual-decoder framework enables robust multi-scale feature extraction and accurate reconstruction of both speech magnitude and phase, providing excellent generalisation to previously unseen noisy environments. The consistent gains in both STOI and PESQ validate the effectiveness of the proposed MSFEYNet as a robust solution for single-channel speech enhancement.

D. DATASET-II

We evaluated the proposed MSFEYNet on the publicly available Voice Bank + DEMAND dataset [28], a standard benchmark for single-channel speech enhancement. We resampled all speech signals to 16 kHz before training and evaluation. The dataset is derived from clean speech recordings in the Voice Bank corpus and environmental noise samples from the DEMAND database. The training set contains 11,572 utterances from 28 speakers (14 female and 14 male), while the test set comprises 824 utterances from unseen speakers (1 female and 1 male) under both seen and unseen noise conditions. This experimental protocol enables a comprehensive evaluation of the proposed model in terms of both enhancement accuracy and generalisation to previously unseen acoustic environments.

From [Tables 6](#) and [7](#), in the seen-noise evaluation, WAVUNET [22] achieves an average of 86.42% STOI [29] and 2.67 PESQ [30], demonstrating that waveform-domain learning effectively preserves speech intelligibility. However, its limited ability to capture long-range contextual information restricts perceptual speech quality under complex acoustic conditions. ICCRN

[23] improves the average performance to 87.45%





STOI and 2.78 PESQ by exploiting complex-valued spectral representations that jointly model speech magnitude and phase information. IUINNB [24] further increases the average STOI and PESQ to 88.63% and 2.89, respectively, indicating that enhanced feature interaction contributes to more accurate speech reconstruction. MSP [25] achieves 89.50% STOI and 2.99 PESQ by effectively learning multi-scale spectral representations. SMWETCW [26] further improves the average performance to 90.54% STOI and 3.17 PESQ, demonstrating that transformer-based contextual modelling substantially improves speech intelligibility and perceptual quality. TDSTN [27] reaches 91.52% STOI and 3.26 PESQ through effective transformer-based contextual modelling. LMAT [7] achieves 92.32% STOI and 3.38 PESQ through efficient long-range dependency modelling with reduced computational complexity. PrimK-Net [6] further improves the performance to 93.41% STOI and 3.48 PESQ, owing to its enhanced multi-scale feature aggregation strategy.

The proposed MSFEYNet consistently achieves the highest performance, recording an average of 94.44% STOI and 3.58 PESQ across all evaluated SNR levels. The superior performance is attributed to the effective integration of the Multi-Scale Feature Extraction Block (MSFEB), Transformer blocks, and the dual-decoder architecture, which jointly exploit local spectral features, long-range contextual dependencies, and simultaneous magnitude-phase estimation. Consequently, the proposed network produces more natural speech with higher intelligibility while effectively suppressing background noise.

Tables 8 and 9 further demonstrate the robustness and generalisation capability of the proposed model in unseen noise experiments. WAVUNet [22] achieves average scores of 86.35% STOI and 2.55 PESQ, indicating reduced performance when evaluated on previously unseen noise conditions. ICCRN [23] improves the results to 87.27% STOI and 2.65 PESQ by leveraging complex domain speech enhancement. IUINNB [24] obtains 88.58% STOI and 2.75 PESQ. While MSP [25] achieves 89.47% STOI and 2.86 PESQ, demonstrating gradual improvement through enhanced feature representation. SMWETCW [26] records 90.47% STOI and 3.12 PESQ, followed by TDSTN [27] which reaches 91.40% STOI and 3.23 PESQ through effective transformer-based contextual modelling. LMAT [7] further improves the performance to 92.25% STOI and 3.34 PESQ, whereas PrimK-Net [6] achieves 93.34% STOI and

3.45 PESQ, reflecting strong robustness against unseen acoustic environments.

The proposed MSFEYNet outperforms all competing approaches with an average 94.33% STOI and 3.55 PESQ. Across all SNR levels, particularly under challenging -5dB and -2dB conditions, the proposed model consistently preserves speech intelligibility and perceptual quality more effectively than the baseline methods. These improvements show that combining MSFEB Transformer-based contextual learning with the dual-decoder framework enables robust multi-scale feature extraction and accurate reconstruction of both speech magnitude and phase, providing excellent generalisation to previously unseen noisy environments. The consistent gains in both STOI and PESQ validate the effectiveness of the proposed MSFEYNet as a robust solution for single-channel speech enhancement.

V. CONCLUSION

This work introduces MSFEYNet, a novel dual-decoder architecture for single-channel speech enhancement that effectively addresses the challenges of speech reconstruction in adverse acoustic environments. By integrating the Multi-Scale Feature Extraction Block (MSFEB) with Transformer blocks in a YNet framework, the proposed model effectively captures both local spectral characteristics and long-range contextual dependencies. Furthermore, the dual decoder design jointly estimates speech magnitude and phase information, enabling more accurate speech reconstruction while preserving perceptual quality and intelligibility.

Extensive experimental evaluations on both seen and unseen noisy speech datasets demonstrate that MSFEYNet consistently outperforms state-of-the-art speech enhancement methods. The proposed model achieves superior performance across multiple objective evaluation metrics, including PESQ and STOI, while maintaining a computationally efficient architecture suitable for practical real-time speech enhancement applications through multi-scale feature extraction. Transformer-based contextual learning and dual-decoder optimisation enable robust noise suppression and improved speech quality under a wide range of noise conditions. In future work, the proposed MSFEYNet will be extended toward deployment on resource-constrained embedded platforms, such as Raspberry Pi and edge AI devices.

Table I: The Ablation Study of the MSFEYNet Model

S.No	MSFEB		Transformer as U-Net		MSFEB + Central Layers (Transformer + D3Net)		Transformer U-Net + Central Layers (Transformer + D3Net)		(Proposed method) MSFEYNet (MSFEB + Transformer + Dual Decoder)	
	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ
Seen	84.17	2.83	86.07	3.03	89.17	3.28	94.07	3.51	96.60	3.69
Unseen	86.69	2.95	87.19	3.02	90.19	3.19	93.21	3.30	95.02	3.59



Table II: STOI % Levels in the Presence of Seen Noise for the Proposed MSFEYNet

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	Prime K-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
STOI (%)	-5	83.80	84.11	85.26	86.86	87.27	88.32	89.66	90.21	91.25	92.14
	-2	84.21	85.37	86.48	87.51	88.62	89.76	90.89	91.07	92.16	93.27
	0	85.81	86.73	87.32	88.64	89.56	90.41	91.37	92.42	93.34	94.11
	2	86.11	87.01	88.05	89.07	90.09	91.10	92.20	93.25	94.27	95.01
	5	87.21	88.09	89.11	90.02	91.07	92.09	93.10	94.11	95.21	96.60
	Avg	85.43	86.27	87.25	88.42	89.33	90.34	91.44	92.21	93.25	94.23

Table III: PESQ Levels in the Presence of Seen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
PESQ	-5	2.36	2.42	2.50	2.62	2.73	2.85	2.99	3.03	3.19	3.21
	-2	2.49	2.51	2.60	2.71	2.82	2.95	3.07	3.15	3.23	3.31
	0	2.55	2.63	2.73	2.89	2.92	3.04	3.11	3.22	3.33	3.47
	2	2.68	2.76	2.87	2.98	3.07	3.18	3.29	3.32	3.43	3.54
	5	2.79	2.88	2.97	3.03	3.14	3.25	3.36	3.47	3.58	3.69
	Avg	2.58	2.64	2.74	2.85	2.94	3.06	3.17	3.24	3.36	3.45

Table IV: STOI % Levels in the Presence of Unseen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
STOI (%)	-5	82.71	83.01	84.28	85.76	86.37	87.42	88.56	89.31	90.35	91.24
	-2	83.11	84.27	85.38	86.41	87.52	88.66	89.79	90.17	91.16	92.17
	0	84.22	85.63	86.22	87.54	88.46	89.31	90.27	91.32	92.24	93.01
	2	85.14	86.11	87.25	88.67	89.19	90.18	91.31	92.51	93.17	94.21
	5	86.36	87.39	88.41	89.05	90.27	91.59	92.18	93.21	94.11	95.02
	Avg	84.31	85.28	86.31	87.49	88.36	89.43	90.42	91.30	92.21	93.13

Table V: PESQ Levels in the Presence of Unseen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
PESQ	-5	2.25	2.31	2.42	2.53	2.64	2.75	2.88	2.91	3.03	3.14
	-2	2.38	2.42	2.53	2.64	2.73	2.85	2.96	3.02	3.15	3.23
	0	2.46	2.54	2.64	2.78	2.83	2.94	3.01	3.13	3.24	3.38
	2	2.59	2.67	2.77	2.89	2.92	3.03	3.19	3.22	3.33	3.44
	5	2.69	2.77	2.88	2.99	3.02	3.13	3.26	3.37	3.48	3.59
	Avg	2.48	2.55	2.65	2.77	2.83	2.94	3.06	3.13	3.25	3.36

Table VI: STOI% Levels in the Presence of Seen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
STOI (%)	-5	83.89	84.21	85.46	86.96	87.37	88.42	89.68	90.29	91.35	92.24
	-2	84.31	85.47	86.58	87.61	88.72	89.86	90.91	91.18	92.26	93.37
	0	85.92	86.82	87.39	88.74	89.61	90.49	91.57	92.62	93.74	94.81
	2	86.21	87.33	88.42	89.36	90.28	91.15	92.26	93.28	94.38	95.11
	5	87.11	88.26	89.39	90.47	91.53	92.78	93.19	94.22	95.31	96.68
	Avg	85.49	86.42	87.45	88.63	89.50	90.54	91.52	92.32	93.41	94.44

Table VII: PESQ Levels in the Presence of Seen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
PESQ	-5	2.43	2.51	2.59	2.70	2.82	2.93	3.05	3.27	3.38	3.43
	-2	2.51	2.60	2.68	2.72	2.89	3.01	3.13	3.22	3.33	3.52
	0	2.61	2.69	2.80	2.91	2.93	3.11	3.21	3.30	3.44	3.59
	2	2.72	2.81	2.90	3.03	3.11	3.19	3.29	3.40	3.54	3.64
	5	2.64	2.78	2.93	3.11	3.20	3.33	3.44	3.51	3.69	3.71
	Avg	2.59	2.67	2.78	2.89	2.99	3.12	3.23	3.34	3.48	3.58



Table VIII: STOI% Levels in the Presence of Unseen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	Proposed MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
STOI (%)	-5	83.73	84.19	85.43	86.93	87.31	88.40	89.66	90.26	91.32	92.21
	-2	84.22	85.33	86.43	87.59	88.72	89.83	90.45	91.13	92.21	93.32
	0	85.32	86.71	87.28	88.70	89.61	90.40	91.51	92.53	93.65	94.44
	2	86.19	87.28	88.39	89.32	90.18	91.05	92.16	93.20	94.29	95.10
	5	87.10	88.22	89.33	90.38	91.51	92.68	93.20	94.12	95.21	96.58
	Avg	85.32	86.35	87.37	88.58	89.47	90.47	91.40	92.25	93.34	94.33

Table IX: PESQ Levels in the Presence of Unseen Noise

Metrics	SNR (dB)	Noise mixture	WAVUNET [22]	ICCRN [23]	IUINNB [24]	MSP [25]	SMWETCW [26]	TDSTN [27]	LMAT [7]	PrimeK-Net [6]	MSFEYNet
Year	-	-	2023	2023	2023	2023	2024	2024	2024	2025	2026
PESQ	-5	2.47	2.52	2.60	2.72	2.83	2.95	3.10	3.29	3.36	3.41
	-2	2.58	2.63	2.72	2.81	2.92	3.07	3.17	3.25	3.31	3.50
	0	2.64	2.74	2.84	2.98	2.09	3.15	3.23	3.32	3.41	3.52
	2	2.79	2.81	2.98	2.09	3.18	3.29	3.31	3.43	3.52	3.61
	5	2.13	2.02	2.09	3.14	3.25	3.36	3.47	3.58	3.63	3.69
	Avg	2.53	2.55	2.65	2.75	2.86	3.17	3.26	3.38	3.45	3.55

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organisations or agencies. This independence ensures the research is conducted objectively, without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate, with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

- Ochieng, P. (2023). Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis. *Artificial Intelligence Review*, 56, 3651–3703. DOI: [10.1007/s10462-023-10612-2](https://doi.org/10.1007/s10462-023-10612-2)
- Li, Y., Sun, Y., Wang, W., & Naqvi, S. M. (2023). U-Shaped Transformer with Frequency-Band Aware Attention for Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 1511–1521. DOI: [10.1109/TASLP.2023.3265839](https://doi.org/10.1109/TASLP.2023.3265839)
- Lu, Y.-X., Ai, Y., & Ling, Z.-H. (2023). MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra. In *Proceedings of Interspeech 2023* (pp. 3834–3838). DOI: [10.21437/Interspeech2023-1441](https://doi.org/10.21437/Interspeech2023-1441)
- Kühne, N. L., Østergaard, J., Jensen, J., & Tan, Z.-H. (2025). xLSTM-SENet: xLSTM for Single-Channel Speech Enhancement. In *Proceedings of Interspeech 2025* (pp. 5148–5152). DOI: [10.21437/Interspeech2025-108](https://doi.org/10.21437/Interspeech2025-108)
- Zhou, H., Zhou, Y., Cheng, Z., Zhao, Y., & Liu, Y. (2025). Improved Encoder-Decoder Architecture with Human-Like Perception Attention for Monaural Speech Enhancement. *IEEE Signal Processing Letters*, 32, 1670–1674. DOI: [10.1109/LSP.2025.3558690](https://doi.org/10.1109/LSP.2025.3558690)
- Lin, Z., Wang, J., Li, R., Shen, F., Xuan, X., 2025. PrimeK-Net: Multi-scale spectral learning via group prime-kernel convolutional neural networks for single-channel speech enhancement, in: Proc. ICASSP, pp. 1–5. DOI: [10.1109/ICASSP49660.2025.10890034](https://doi.org/10.1109/ICASSP49660.2025.10890034)
- Liang, X., Zhang, Z., Wang, M., & Xu, R. (2024, April). Lightweight multi-axial transformer with frequency prompt for single-channel speech enhancement. In *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 10511–10515). IEEE. DOI: [10.1109/ICASSP48485.2024.10446787](https://doi.org/10.1109/ICASSP48485.2024.10446787)
- Chao, R., et al., 2024. An investigation of incorporating Mamba for speech enhancement, in: Proc. IEEE Spoken Lang. Technol. Workshop (SLT), pp. 302–308. DOI: [10.1109/SLT61566.2024.10832332](https://doi.org/10.1109/SLT61566.2024.10832332)
- Luo, T., Zhou, F., Bai, Z., 2024. MambaGAN: Mamba-based metric GAN for monaural speech enhancement, in: Proc. Int. Conf. Asian Lang. Process. (IALP), pp. 411–416. DOI: [10.1109/IALP63756.2024.10661187](https://doi.org/10.1109/IALP63756.2024.10661187)
- Abdulatif, S., Cao, R., Yang, B., 2024. CMGAN: Conformer-based Metric-GAN for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* 32, 2477–2493. DOI: [10.1109/TASLP.2024.3393718](https://doi.org/10.1109/TASLP.2024.3393718)
- Hu, Y., Yang, Q., Wei, W., Lin, L., He, L., Ou, Z., Yang, W., 2025. MNNet: Speech enhancement network via modelling the noise. *IEEE Trans. Audio Speech Lang. Process.* 33, 1208–1219. DOI: [10.1109/TASLP.2025.3546819](https://doi.org/10.1109/TASLP.2025.3546819)
- Lu, Y.-X., Ai, Y., Ling, Z., 2025. Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement. *Neural Networks* 189, 107562. DOI: [10.1016/j.neunet.2025.107562](https://doi.org/10.1016/j.neunet.2025.107562)
- Chen, H., Zhang, J., Fu, Y., Zhou, X., Wang, R., Xu, Y., & Ke, D. (2025). TFDense-GAN: A Generative Adversarial Network for Single-Channel Speech Enhancement. *EURASIP Journal on Advances in Signal Processing*, 2025, Article 10. DOI: [10.1186/s13634-025-01210-1](https://doi.org/10.1186/s13634-025-01210-1)
- Wang, H., Tian, B., 2025. ZipEnhancer: Dual-path down-up sampling-based Zipformer for monaural speech enhancement, in: Proc. ICASSP, pp. 1–5. DOI: [10.1109/ICASSP49660.2025.10888703](https://doi.org/10.1109/ICASSP49660.2025.10888703)
- Wang, J., Lin, Z., Wang, T., Ge, M., Wang, L., Dang, J., 2025. Mamba-SEUNet: Mamba UNet for monaural speech enhancement, in: Proc. ICASSP, pp. 1–5. DOI: [10.1109/ICASSP49660.2025.10889525](https://doi.org/10.1109/ICASSP49660.2025.10889525)
- Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., Wang, X., 2024. MetaFormer baselines for vision. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 896–912. DOI: [10.1109/TPAMI.2023.3329173](https://doi.org/10.1109/TPAMI.2023.3329173)
- Dang, F., Chen, H., & Zhang, P. (2022, May). DPT-FSNet: Dual-path transformer-based full-band and sub-band fusion network for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6857–6861). IEEE. DOI: [10.1109/ICASSP43922.2022.9746171](https://doi.org/10.1109/ICASSP43922.2022.9746171)
- Zhang, C., Wang, L., Qiao, Y., & Gao, W. A multi-scale feature extraction fusion model for human activity recognition. *Scientific Reports*, 12(1), 20620 (2022). DOI: [10.1038/s41598-022-24909-2](https://doi.org/10.1038/s41598-022-24909-2)
- Chen, M., Zhang, Q., Wang, M., Zhang, X., Liu, H., Ambikairah, E., & Chen, D. (2024). Selective state space model for monaural speech enhancement. arXiv preprint arXiv:2411.06217.



- DOI: <https://arxiv.org/abs/2411.06217>.
20. Chao, R., Nasretidinov, R., Wang, Y. C. F., Jukić, A., Fu, S. W., & Tsao, Y. (2025). Universal speech enhancement with regression and generative Mamba. arXiv preprint arXiv:2505.21198. DOI: [10.48550/arXiv.2505.21198](https://doi.org/10.48550/arXiv.2505.21198)
 21. Wang, J., Lin, Z., Wang, T., Ge, M., Wang, L., & Dang, J. (2025). *Mamba-SEUNet: Mamba UNet for Monaural Speech Enhancement*. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025) (pp. 1–5). DOI: [10.1109/ICASSP49660.2025.10889525](https://doi.org/10.1109/ICASSP49660.2025.10889525)
 22. Mathieu, F., Courtat, T., Richard, G., & Peeters, G. Learning interpretable filters in Wav-U-Net for speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023). DOI: [10.1109/ICASSP.49357.2023.10097153](https://doi.org/10.1109/ICASSP.49357.2023.10097153)
 23. J., & Zhang, X. ICCRN: In-place cepstral convolutional recurrent neural network for monaural speech enhancement. In International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE (2023). DOI: [10.1109/ICASSP.49357.2023.10096918](https://doi.org/10.1109/ICASSP.49357.2023.10096918)
 24. Fang, H., Becker, D., Wermter, S., & Gerkmann, T. Integrating uncertainty into neural network-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 1587–1600. (2023). DOI: [10.1109/TASLP.2023.3274628](https://doi.org/10.1109/TASLP.2023.3274628)
 25. Zmolikova, K., Pedersen, M. S., & Jensen, J. Masked spectrogram prediction for unsupervised domain adaptation in speech enhancement. *IEEE Open Journal of Signal Processing*, 5, 274–283 (2023). DOI: [10.1109/OJSP.2023.3305948](https://doi.org/10.1109/OJSP.2023.3305948)
 26. Fiorio, L. V., Karanov, B., Defraene, B., David, J., Widdershoven, F., van Houtum, W., & Aarts, R. M. Spectral masking with explicit time-context windowing for neural network-based monaural speech enhancement. *IEEE Access*, 12, 154843–154852 (2024). DOI: [10.1109/ACCESS.2024.3483443](https://doi.org/10.1109/ACCESS.2024.3483443)
 27. Alohal, M. A., Saleem, N., Rhouma, D., Medani, M., Elmannai, H., & Bourouis, S. Temporally dynamic spiking transformer network for speech enhancement. *IEEE Access*, 12, 146513–146526 (2024). DOI: [10.1109/ACCESS.2024.3444596](https://doi.org/10.1109/ACCESS.2024.3444596)
 28. Valentini-Botinhao, C., Wang, X., Takaki, S., & Yamagishi, J. (2016). *Investigating RNN-based Speech Enhancement Methods for Noise-Robust Text-to-Speech*. In Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW 9) (pp. 159–165). DOI: [10.21437/SSW.2016-19](https://doi.org/10.21437/SSW.2016-19)
 29. Kadambi, M., Sharma, C. M., Mandal, A., & Ghosh, P. K. (2025, December). Bottleneck Transformer-Based Approach for Improved Automatic STOI Score Prediction. In 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-7). IEEE DOI: [10.1109/ASRU65441.2025.11434770](https://doi.org/10.1109/ASRU65441.2025.11434770)
 30. Xu, Z., Zhao, Z., & Fingscheidt, T. (2023). Coded speech quality measurement by a non-intrusive PESQ-DNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 3404–3417. DOI: [10.1109/TASLP.2023.3317574](https://doi.org/10.1109/TASLP.2023.3317574)
 31. Kadir, A. D. I. A., Al-Haiqi, A., & Din, N. M. (2021, December). A dataset and TinyML model for coarse age classification based on voice commands. In 2021 IEEE 15th Malaysia International Conference on Communication (MICC) (pp. 75-80). IEEE. DOI: [10.1109/MICC53484.2021.9642091](https://doi.org/10.1109/MICC53484.2021.9642091)
 32. Yu, G., Li, A., Zheng, C., Guo, Y., Wang, Y., & Wang, H. (2022, May). Dual-branch attention-in-attention transformer for single-channel speech enhancement. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7847-7851). IEEE. DOI: [10.1109/ICASSP43922.2022.9746273](https://doi.org/10.1109/ICASSP43922.2022.9746273)
 33. Ma, X., Dai, X., Bai, Y., Wang, Y., & Fu, Y. (2024). *Rewrite the Stars*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5694–5703). IEEE. DOI: [10.1109/CVPR52733.2024.00544](https://doi.org/10.1109/CVPR52733.2024.00544)
 34. Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., & Hu, S.-M. (2023). *Visual attention network*. *Computational Visual Media*, 9(4), 733–752. DOI: [10.1007/s41095-023-0364-2](https://doi.org/10.1007/s41095-023-0364-2)



Dr Sunnydayal Vanambathina received his B. Tech in Electronics and Communication Engineering and M. Tech in Signal Processing from the National Institute of Technology (NIT) Calicut. He completed his PhD at the National Institute of Technology Warangal, India. He also conducted part of his research at the University of Seville, Spain. He subsequently pursued postdoctoral research at the University of Crete, Greece. Under his amiable guidance, 2 scholars have been awarded PhD degrees, and 4 more are currently pursuing their PhD under his supervision. To date, he has published 35 papers in SCI/Scopus-indexed journals, presented at 20 international conferences, filed 4 patents, and contributed 3 book chapters. His current research interests include speech enhancement, statistical signal processing, blind source separation, and machine learning techniques.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Publisher/Journal and/or the editor(s). The Publisher/Journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

AUTHOR'S PROFILE



Silpa Peethala received the B.Tech. degree in Electronic and Communication Engineering from JNTU Kakinada, India, in 2009, and an M.Tech. Degree in System and Signal Processing from JNTU Kakinada, India, in 2012. She is currently pursuing research in speech enhancement and deep learning-based audio signal processing. Her current research interests include single-channel speech enhancement, transformer-based architectures, and objective speech quality evaluation.