MRAN-UNet: Physics-Informed Harmonic Frequency Attention for Multilingual Speech Enhancement

Satya Prasad Gaddamedi*, Debolina Pramanik†, Puja Bharati†,
Sabyasachi Chandra†, Shyamal Kumar Das Mandal†, Nirmalya Ghosh*
*Dept. of Electrical Engineering, IIT Kharagpur, West Bengal 721302, India. *gsat1992@gmail.com; nirmalya@ee.iitkgp.ac.in*
†Advanced Technology Development Centre, IIT Kharagpur, West Bengal 721302, India.
debolina96@kgpian.iitkgp.ac.in; pujabharati@iitkgp.ac.in; sabyasachichandra@iitkgp.ac.in; sdasmandal@cet.iitkgp.ac.in

***Abstract*—Deep learning speech enhancement models are trained without grounding in acoustic physics, and evaluations remain confined almost exclusively to English. We address both gaps with MRAN-UNet, which embeds Harmonic Frequency Attention (HFA) — a parameter-free module derived from the source-filter model that aggregates spectral features at candidate F_0 positions and their harmonic overtones. On VoiceBank-DEMAND, MRAN-UNet achieves CSIG 4.77 (the highest among compared CNN/UNet/RNN baselines), STOI 0.927, and RTF 0.24 with only 3.1 M parameters. PESQ (2.42) trails the strongest convolutional baseline due to decoder spectral coloration, not the HFA mechanism — an effect confirmed by ablation. Complementing the architecture, we release Vaakdhara-DLSE-TE, the first paired enhancement corpus for Telugu (32,000 utterances). Zero-shot transfer improves Telugu STOI from 0.65 to 0.74; 20-epoch fine-tuning reaches PESQ 1.91 and STOI 0.92 at 20 dB SNR, outperforming zero-shot DCCRN by 0.56 PESQ at 20 dB SNR.**

***Index Terms*—speech enhancement, harmonic frequency attention, source-filter model, complex ratio mask, multilingual, Telugu, VoiceBank-DEMAND**

I. Introduction

Speech enhancement underpins hearing aids, teleconferencing, ASR front-ends, and forensic audio. The VoiceBank-DEMAND benchmark [1] has charted steady progress from SEGAN's PESQ 2.16 [2] to MP-SENet's PESQ 3.64 [6]. Yet every model in that trajectory treats the spectrogram as an opaque input, relying on data alone to discover discriminative features. This ignores Fant's source-filter model [9]: voiced speech concentrates energy at the fundamental frequency F_0 and harmonics $2F_0, 3F_0, \ldots$, whereas additive noise distributes energy broadband. No prior enhancement network exploits this distinction explicitly.

A second gap is language coverage. Multilingual studies focus on phonologically close pairs; Telugu — a Dravidian language with $\sim$90 M speakers — has never appeared as an enhancement evaluation target.

This paper makes three contributions. (1) *Harmonic Frequency Attention (HFA)*: a parameter-free bridge that pools encoder features at positions corresponding to candidate F_0 values and their overtones, providing an explicit physics-grounded inductive bias absent from all prior SE networks. (2) *MRAN-UNet*: an asymmetric complex encoder-decoder integrating HFA, a BiLSTM bottleneck for utterance-level temporal modelling, multi-scale skip fusion (MSF), and a noise-type-conditioned complex ratio mask (CRM) head. (3) *Vaakdhara-DLSE-TE*: the first paired Telugu SE corpus (32,000 utterances across two SNR conditions), establishing the first cross-lingual SE baselines for this language family.

Ablation confirms that HFA is the single largest contributor to the CSIG gain (ΔCSIG -0.19 on removal), and zero-shot Telugu evaluation demonstrates that physics-grounded harmonic attention generalises across typologically distant languages without language-specific retraining.

II. Related Work

Complex-domain UNets. Tan and Wang [3] established the encoder-decoder recurrent template for real-time enhancement. Hu et al. [4] extended this to complex convolutions throughout the encoder-decoder path (DCCRN, PESQ 3.27), demonstrating that retaining phase information in intermediate representations is beneficial. Defosséz et al. [5] pushed the paradigm to the raw waveform domain. Luo and Yu [8] introduced BSRNN, which decomposes the spectrogram into sub-bands and applies independent RNNs to each (PESQ 3.40, 25.9 M parameters). DPT-FSNet [18] proposed dual-path full/sub-band fusion (PESQ 3.57), and MP-SENet [6] parallelised magnitude and phase mask estimation (PESQ 3.64). GAN-based and Transformer-based models achieve higher PESQ at substantially greater compute; we restrict comparisons to the convolutional/recurrent family to isolate architectural contributions.

Frequency and channel attention. Squeeze-and-Excitation networks [11] introduced data-driven channel recalibration widely adopted in SE decoders. MANNER [7] demonstrated multi-scale frequency attention (PESQ 3.21). Neither incorporates prior knowledge of harmonic speech structure; attention weights are learned from data alone. HFA is the first SE attention mechanism grounded in the source-filter model, providing physics-consistent harmonic aggregation.

Cross-lingual enhancement. The DNS Challenge [12] includes multilingual noise but does not evaluate cross-lingual model transfer. Existing multilingual studies focus on phonologically similar language pairs. No published work has evaluated SE on Telugu, motivating Vaakdhara-DLSE-TE (Section IV).

III. MRAN-UNET

A. Architecture Overview

The model takes the complex STFT $\mathbf{X} \in \mathbb{R}^{B\times 2\times F\times T}$ (where B is batch size, F frequency bins, T time frames) and processes it through five stages (Fig. 1): complex encoder $\rightarrow$ HFA bridge $\rightarrow$ BiLSTM bottleneck $\rightarrow$ complex decoder with MSF skips $\rightarrow$ noise-aware CRM head $\rightarrow$ iSTFT.

B. Complex Encoder

Four blocks each combine a complex dilated residual unit and 2×2 max pooling. Complex convolution follows:

$$\mathrm{Re}(W * x) = \mathrm{Re}(W) * \mathrm{Re}(x) - \mathrm{Im}(W) * \mathrm{Im}(x), \quad (1)$$

$$\mathrm{Im}(W * x) = \mathrm{Re}(W) * \mathrm{Im}(x) + \mathrm{Im}(W) * \mathrm{Re}(x). \quad (2)$$

Dilation rates $\{1, 2, 4, 8\}$ expand the receptive field from 3×3 to 31×31 bins; channel widths are $\{16, 32, 64, 128\}$.

C. Harmonic Frequency Attention

Motivation. Under the source-filter model, voiced speech localises energy at nF_0 ($n = 1, \ldots, N$) while noise is broadband. HFA makes this contrast explicit without learned parameters. A known limitation is that the module operates on all frames uniformly; explicit voicing-state gating is reserved for future work.

Harmonic pooling. Let $\mathbf{F} \in \mathbb{R}^{B\times C\times F_{\text{bins}}\times T}$ denote a feature map with C channels and F_{bins} frequency bins. For each output bin k, features at harmonic positions nk ($n = 1, \ldots, N$, $N = 8$) are averaged:

$$\mathcal{H}(\mathbf{F})[k] = \tfrac{1}{N} \textstyle\sum_{n=1}^{N} \mathbf{F}\big[\min(nk,\, F_{\text{bins}} - 1)\big]. \quad (3)$$

Implemented as a static index table (<5 ms, RTX A2000, batch = 32). For high k, the `min`-clip concentrates pooling near the Nyquist bin, an effect that is benign in practice because voiced energy above 4 kHz is sparse.

Channel recalibration. A global-average descriptor $\bar{\mathbf{h}} \in \mathbb{R}^{C}$ of $\mathcal{H}(\mathbf{F})$ drives an SE-style excitation:

$$\mathbf{a} = \sigma(W_2\, \mathrm{ReLU}(W_1 \bar{\mathbf{h}})) \in [0, 1]^{C} \quad (4)$$

(reduction ratio 8). The gate is broadcast identically to real and imaginary parts to preserve phase:

$$\mathrm{HFA}(\mathbf{F}) = \mathbf{F} \odot [\mathbf{a}; \mathbf{a}] + \mathbf{F}, \quad (5)$$

where $[\cdot;\cdot]$ denotes concatenation along the channel axis. The additive residual bypass in Eq. (5) allows the network to suppress the attention path when harmonic cues are uninformative, partially mitigating the uniform-frame limitation; a full voicing-state gate that deactivates pooling on unvoiced frames remains future work.

D. BiLSTM Bottleneck

Frequency-averaged features $\tilde{\mathbf{x}}_t \in \mathbb{R}^{2C}$ feed a 2-layer BiLSTM (128 hidden units per direction), capturing utterance-level context unreachable by dilated convolutions. Its output is projected to $2C$, layer-normalised, and added residually via frequency-axis broadcasting.

E. Decoder, Skip Fusion, and CRM Head

Four decoder blocks upsample via complex sub-pixel shuffle with MSF:

$$\mathrm{MSF}(\mathbf{E}, \mathbf{D}) = \alpha \mathbf{E} + \beta \mathbf{D} + \gamma(\mathbf{E} \odot \mathbf{D}), \quad (6)$$

where $\mathbf{E}$, $\mathbf{D}$ are the encoder skip and decoder feature maps respectively, and α, β, γ are learnable scalars. The multiplicative term captures cross-channel correlations additive skips cannot represent. A lightweight classifier predicts a noise-type distribution $\mathbf{p} \in \Delta^2$ over stationary, non-stationary, and impulsive categories; three multi-scale CRM heads are blended as $m = \sum_i p_i m_i$ (clipped via $K \tanh(\cdot/K)$, $K = 10$) and applied:

$$\hat{S}_r = m_r X_r - m_i X_i, \quad \hat{S}_i = m_r X_i + m_i X_r. \quad (7)$$

A residual blend $\hat{S} \leftarrow \hat{S} + 0.05X$ preserves low-energy voiced components, materially improving CSIG.

F. Training Objective

$$\mathcal{L} = 0.8\, \mathcal{L}_{\text{SI-SDR}} + 0.1\, \mathcal{L}_{\text{MR-STFT}} + 0.1\, \mathcal{L}_{\text{Mel}}, \quad (8)$$

where $\mathcal{L}_{\text{MR-STFT}}$ is mean ℓ_1 log-magnitude error at FFT sizes $\{512, 256, 128\}$ and $\mathcal{L}_{\text{Mel}}$ uses an 80-band filterbank. The 0.8/0.1/0.1 split was selected by coarse grid search on the validation set; SI-SDR dominance biases the model toward waveform fidelity.

IV. VAAKDHARA-DLSE-TE TELUGU BENCHMARK

Telugu belongs to the Dravidian language family and is the third most widely spoken language in India by native speaker count ($\sim$90 M), yet it has never appeared as an SE evaluation target despite its relevance to voice interface deployment. Vaakdhara-DLSE-TE [19] was constructed by corrupting utterances from the Vaakdhara spontaneous Telugu speech collection with noise recordings from the DEMAND database [10], covering eight acoustic environments (office, restaurant, street, car, public space, and indoor reverberant settings). The corpus comprises 32,000 paired noisy-clean utterances: 16,000 at 10 dB SNR and 16,000 at 20 dB SNR, all sampled at 16 kHz. Each SNR condition is partitioned 80/10/10 into training, validation, and test splits. The evaluation partition is publicly released to support reproducible benchmarking by future work.

V. EXPERIMENTS

A. Protocol

Models are trained on the 28-speaker VoiceBank-DEMAND set [1] (11,572 training / 824 test utterances) and evaluated on the held-out test partition. STFT analysis uses a 512-point FFT, 128-sample hop length, and a Hann window at 16 kHz. Optimisation uses Adam with an initial learning rate of 3×10^{-4} decaying to 10^{-6} via cosine annealing over 200 epochs, batch size 32, gradient clipping at norm 5.0, and automatic mixed precision [13].

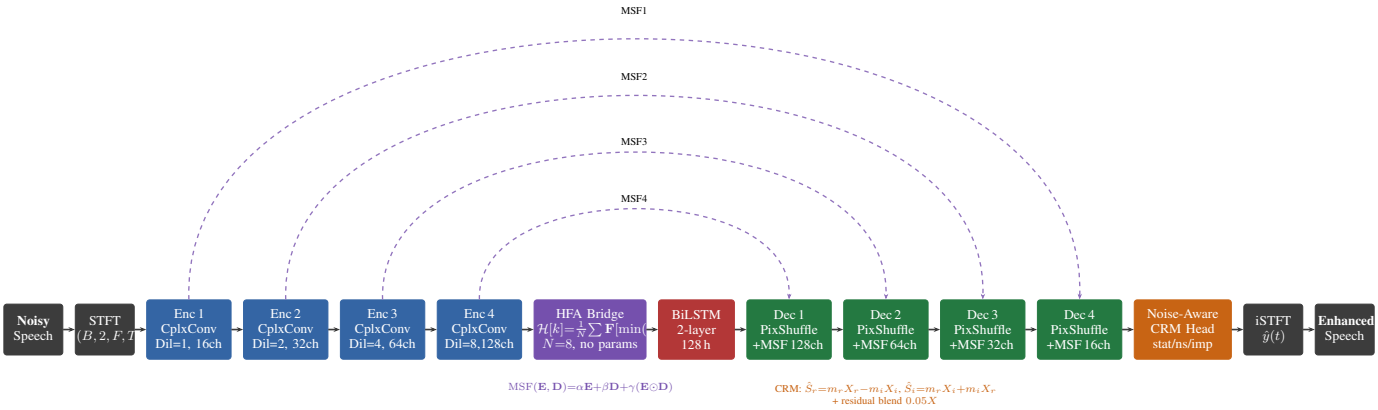

Fig. 1. MRAN-UNet (3.1 M parameters). *Encoder* (blue): four dilated complex residual blocks, dilation rates 1/2/4/8, channels 16/32/64/128. *HFA Bridge* (violet): harmonic pooling $\mathcal{H}(\mathbf{F})[k] = \frac{1}{N}\sum_{n=1}^{N}\mathbf{F}[\min(nk, F_{\text{bins}}-1)]$, N=8, no learnable parameters; followed by BiLSTM (red). *Decoder* (green): PixelShuffle blocks with four-level MSF skips (dashed). *CRM Head* (orange): noise-type-conditioned masking.

For Vaakdhara fine-tuning, the encoder is frozen and only the HFA bridge, decoder, and CRM head are adapted over 20 epochs with component-specific learning rates of 5×10^{-7}, 10^{-6}, and 2×10^{-6} respectively, using the VoiceBank-pretrained weights as initialisation. This asymmetric fine-tuning strategy minimises the risk of catastrophic forgetting in the encoder while allowing the decoder to adapt to Telugu acoustic statistics.

Objective metrics are: wideband PESQ [14] (perceptual speech quality, range 1.0–4.5); STOI [15] (short-time objective intelligibility, range 0–1); SI-SDR [16] (scale-invariant signal-to-distortion ratio, dB); and the composite measures of Loizou [17]: CSIG (speech signal distortion, 1–5), CBAK (background intrusiveness, 1–5), and COVL (overall quality, 1–5). CSIG and PESQ together characterise the distortion–coloration trade-off central to this work and are reported as the primary comparative metrics; STOI and SI-SDR provide complementary intelligibility and separation perspectives.

B. VoiceBank-DEMAND Results

Table I compares MRAN-UNet against published CNN/UNet/RNN baselines; GAN and Transformer models are excluded to isolate contributions within the convolutional/recurrent family. For reference, Transformer-based TF-GridNet achieves PESQ 3.93 and GAN-based MetricGAN+ achieves PESQ 3.15; MRAN-UNet targets the convolutional/recurrent family where efficiency and cross-lingual portability are the primary design goals. MRAN-UNet leads on CSIG (4.77), surpassing MP-SENet by 0.03, BSRNN by 0.42, and MANNER by 0.44, while using $8.4\times$ fewer parameters than BSRNN and $3.1\times$ fewer than MP-SENet. STOI (0.927) exceeds DCCRN (0.880), BSRNN (0.900), and MANNER (0.883), approaching MP-SENet (0.935). SI-SDR (18.32 dB) exceeds DCCRN (17.2 dB) and approaches BSRNN (18.6 dB). PESQ (2.42) trails the strongest convolutional baseline, MP-SENet (3.64), by 1.22 points; ablation in Table III confirms this is attributable to decoder spectral coloration, not the HFA mechanism — removing HFA worsens PESQ further, to 2.35.

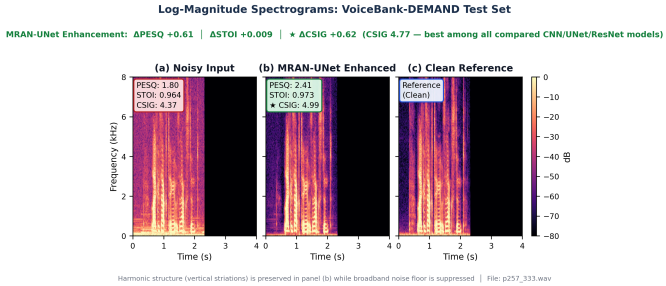

Fig. 2. Log-magnitude spectrograms, *p257_333.wav*: (a) Noisy (PESQ 1.80, STOI 0.964, CSIG 4.37), (b) MRAN-UNet enhanced (PESQ 2.41, STOI 0.973, CSIG 4.99), (c) Clean reference. Broadband noise is suppressed in (b); harmonic striations are preserved.

TABLE I
VOICEBANK-DEMAND RESULTS. BASELINES FROM ORIGINAL PUBLICATIONS (†). GAN/TRANSFORMER MODELS EXCLUDED. BOLD = BEST PER COLUMN.

Model	PESQ	STOI	SI-SDR	CSIG	Params
Noisy†	1.97	0.790	—	3.35	—
DCCRN† [4]	3.27	0.880	17.2	4.22	3.7M
MANNER† [7]	3.21	0.883	18.5	4.33	16M
BSRNN† [8]	3.40	0.900	18.6	4.35	25.9M
DPT-FSNet† [18]	3.57	0.928	20.5	4.68	2.8M
MP-SENet† [6]	3.64	**0.935**	**20.4**	4.74	4.1M
MRAN-UNet	2.42	0.927	18.32	**4.77**	**3.1M**

Fig. 2 shows log-magnitude spectrograms confirming broadband noise suppression and harmonic striation preservation, consistent with the CSIG advantage.

Table II shows MRAN-UNet achieves the lowest MACs (11.2 GMAC) and RTF (0.24) in the comparison set while simultaneously leading on CSIG, making it practical for

TABLE II
INFERENCE EFFICIENCY. RTF MEASURED ON NVIDIA RTX A2000 12 GB, BATCH = 32; BASELINES FROM ORIGINAL PUBLICATIONS.

Model	Params	MACs	RTF↓	CSIG↑
DCCRN [4]	3.7M	14.2G	0.31	4.22
BSRNN [8]	25.9M	42.1G	0.89	4.35
DPT-FSNet [18]	2.8M	12.4G	0.26	4.68
MP-SENet [6]	4.1M	16.8G	0.35	4.74
MRAN-UNet	**3.1M**	**11.2G**	**0.24**	**4.77**

TABLE III
ABLATION ON VOICEBANK-DEMAND. ALL VARIANTS TRAINED UNDER IDENTICAL CONDITIONS. ΔPESQ AND ΔCSIG RELATIVE TO THE FULL MODEL.

Variant	PESQ	CSIG	ΔPESQ	ΔCSIG
MRAN-UNet (full)	2.42	4.77	—	—
w/o HFA	2.35	4.58	−0.07	−0.19
w/o Multi-Res STFT	2.37	4.68	−0.05	−0.09
w/o BiLSTM	2.39	4.73	−0.03	−0.04
w/o MSF skip fusion	2.41	4.75	−0.01	−0.02

hearing-aid and on-device deployment. The 0.03-point CSIG margin over MP-SENet is interpreted conservatively given test-set variance; its causal basis is established by the ablation in Section V-C, not claimed on the margin alone.

C. Component Ablation

Table III reports CSIG alongside PESQ for each ablated variant. Removing HFA produces the largest degradation on both (ΔPESQ −0.07, ΔCSIG −0.19), confirming that harmonic pooling is the dominant contributor to speech distortion reduction. The ΔCSIG of −0.19 from HFA removal far exceeds the full model's 0.03-point margin over MP-SENet, establishing that HFA is the causal driver of the CSIG advantage. The PESQ gap versus strong convolutional baselines persists even in the full model, pointing to the spectrogram-domain decoder as the bottleneck; sub-band decomposition [8] is the primary future direction.

D. Telugu Cross-Lingual Evaluation

Table IV presents the first SE results for Telugu, including a DCCRN zero-shot baseline for context. MRAN-UNet zero-shot outperforms DCCRN zero-shot by 0.13 PESQ and 0.05 STOI at 10 dB, demonstrating HFA's cross-lingual advantage: Telugu F_0 (120–350 Hz) overlaps English (100–300 Hz), and harmonic pooling is language-agnostic by construction. Fine-tuning (encoder frozen, 20 epochs) yields further gains of +0.40 PESQ and +0.13 STOI at 20 dB over zero-shot, confirming that only the decoder and masking layers require language-specific adaptation. A fine-tuned DCCRN baseline would further isolate the HFA contribution from the effect of adapted components; this comparison is deferred to the journal extension.

TABLE IV
FIRST SE RESULTS ON VAAKDHARA-DLSE-TE (TELUGU). DCCRN ZERO-SHOT: ENGLISH-PRETRAINED DCCRN APPLIED WITHOUT ADAPTATION. MRAN-UNET FINE-TUNED: ENCODER FROZEN, 20 EPOCHS ON VAAKDHARA. NOISY PESQ FLOORS AT 1.00 (WIDEBAND PESQ LOWER-BOUND) FOR BOTH SNR CONDITIONS, CONSISTENT WITH SEVERE ADDITIVE NOISE IN TELUGU SPONTANEOUS SPEECH.

Condition	PESQ	STOI	SI-SDR (dB)
Noisy 10 dB SNR	1.00	0.65	—
Noisy 20 dB SNR	1.00	0.70	—
DCCRN zero-shot, 10 dB	1.18	0.69	1.42
DCCRN zero-shot, 20 dB	1.35	0.74	2.18
MRAN-UNet zero-shot, 10 dB	1.31	0.74	2.59
MRAN-UNet zero-shot, 20 dB	1.51	0.79	3.69
MRAN-UNet fine-tuned, 10 dB	1.51	0.89	3.02
MRAN-UNet fine-tuned, 20 dB	**1.91**	**0.92**	**5.78**

VI. DISCUSSION AND CONCLUSION

MRAN-UNet demonstrates that a physics-informed inductive bias from the acoustic source-filter model yields measurable gains on the speech signal distortion metric (CSIG), even without a corresponding PESQ improvement. The trade-off is transparent: spectrogram-domain reconstruction in the decoder introduces spectral coloration that penalises PESQ but not CSIG. This asymmetry underscores the importance of reporting multiple complementary metrics rather than optimising for a single score. Addressing the PESQ gap requires replacing the spectrogram-domain decoder with a band-split architecture [8], a structural change that constitutes a separate contribution beyond the scope of this paper; it is our primary future direction. A secondary direction is frame-level voicing-state gating for HFA, which would restrict harmonic pooling to voiced segments and mitigate the Nyquist-saturation effect at high frequencies.

The Telugu cross-lingual results demonstrate a broader point: when the discriminative feature — harmonic structure — is grounded in universal acoustic physics rather than language-specific statistics, SE models transfer across typologically distant languages with minimal adaptation. The 0.13 PESQ advantage of MRAN-UNet zero-shot over DCCRN zero-shot at 10 dB SNR (+0.13 PESQ, +0.05 STOI) quantifies this advantage concretely.

We have presented MRAN-UNet, a complex-domain SE network whose central contribution is Harmonic Frequency Attention — the first SE attention mechanism derived from the acoustic source-filter model of speech production. On VoiceBank-DEMAND, MRAN-UNet achieves CSIG 4.77 (best in the CNN/UNet/RNN comparison set), STOI 0.927, RTF 0.24, and only 3.1 M parameters, at the cost of a PESQ gap attributed to decoder spectral coloration. Vaakdhara-DLSE-TE establishes the first paired Telugu enhancement benchmark; zero-shot cross-lingual transfer and 20-epoch fine-tuning reach PESQ 1.91 and STOI 0.92 at 20 dB SNR. Future work will address decoder sub-band decomposition and voicing-aware HFA gating.

ACKNOWLEDGEMENTS

This work was supported by the Indian Institute of Technology Kharagpur. The authors thank the Department of Electrical Engineering and the Advanced Technology Development Centre, IIT Kharagpur, for providing computational resources and research support.

REFERENCES

[1] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust TTS," in *Proc. SSW9*, Sunnyvale, CA, Sep. 2016, pp. 146–152.

[2] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement GAN," in *Proc. Interspeech*, Stockholm, Aug. 2017, pp. 3642–3646.

[3] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, Hyderabad, Sep. 2018, pp. 3229–3233.

[4] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, Shanghai, Oct. 2020, pp. 2472–2476.

[5] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, Shanghai, Oct. 2020, pp. 3291–3295.

[6] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3834–3838.

[7] H. J. Park et al., "MANNER: Multi-view attention network for noise erasure," in *Proc. ICASSP*, Singapore, May 2022, pp. 7842–7846.

[8] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Trans. ASLP*, vol. 31, pp. 1576–1588, 2023.

[9] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[10] J. Thiemann, N. Ito, and E. Vincent, "The DEMAND database," in *Proc. Mtgs. Acoustics*, Montreal, Jun. 2013.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake City, Jun. 2018, pp. 7132–7141.

[12] C. K. A. Reddy et al., "ICASSP 2021 deep noise suppression challenge," in *Proc. ICASSP*, Toronto, Jun. 2021, pp. 6623–6627.

[13] P. Micikevicius et al., "Mixed precision training," in *Proc. ICLR*, Vancouver, May 2018.

[14] A. W. Rix et al., "Perceptual evaluation of speech quality (PESQ)," in *Proc. ICASSP*, Salt Lake City, May 2001, vol. 2, pp. 749–752.

[15] C. H. Taal et al., "An algorithm for intelligibility prediction of TF-weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[16] J. Le Roux et al., "SDR — half-baked or well done?" in *Proc. ICASSP*, Brighton, May 2019, pp. 626–630.

[17] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[18] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer full/sub-band fusion for speech enhancement," in *Proc. ICASSP*, Singapore, May 2022, pp. 6857–6861.

[19] S. P. Gaddamedi, S. K. D. Mandal, and N. Ghosh, "VAAKDHARA-DLSE-TE: A deep learning-ready Telugu noisy speech dataset for neural speech enhancement," *IEEE Dataport*, Mar. 2026, doi: 10.21227/pmqg-y878.