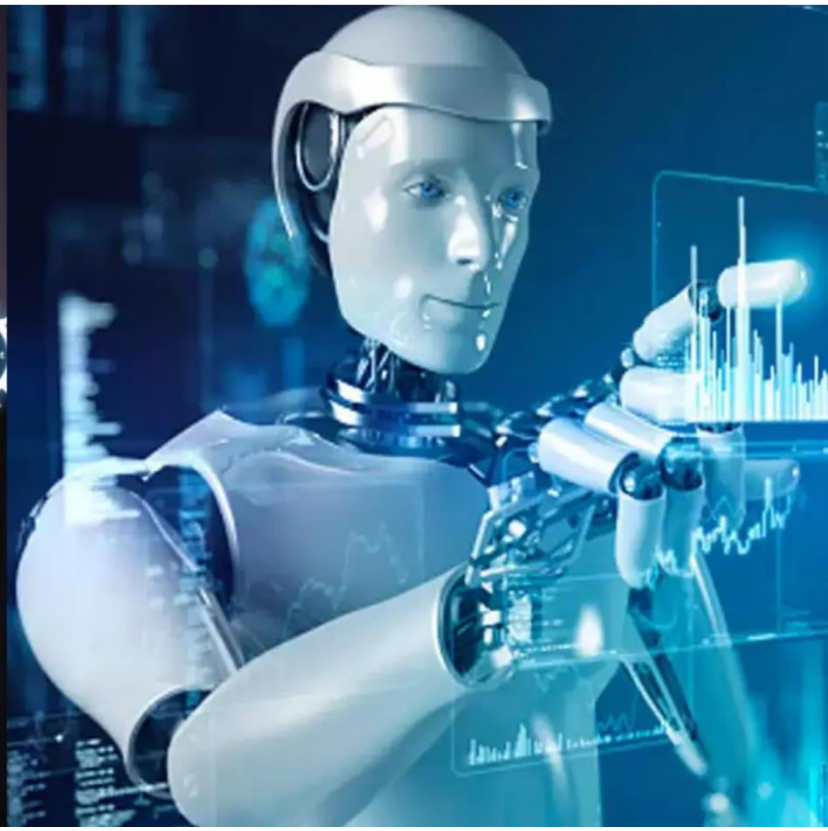
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 5, May 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

Review of Speech Enhancement using Artificial Intelligence

Sanket .N. Wawale

Department of Information Technology, Amrutvahini Polytechnic, Sangamner, Maharastra, India

ABSTRACT: The degradation of audio quality in legacy recordings poses a significant challenge in preserving cultural, historical, and archival content. With the advancement of Artificial Intelligence (AI), particularly deep learning techniques, it is now possible to restore, enhance, and regenerate lost or damaged audio data. This paper explores the methodologies, models, and applications of AI in audio restoration. It discusses signal degradation issues, modern AI-based reconstruction approaches, and evaluates their effectiveness compared to traditional signal processing techniques.

KEYWORDS: Artificial Intelligence, Audio Restoration, Deep Learning, Signal Processing, Speech Enhancement, Generative Adversarial Networks

I. INTRODUCTION

Audio recordings degrade over time due to physical wear, environmental factors, and technological limitations of earlier recording systems. Common issues include noise, distortion, missing segments, and reduced dynamic range. Traditional restoration techniques rely heavily on manual filtering and signal processing, which often fail to recover lost data effectively.

Artificial Intelligence introduces data-driven approaches capable of learning complex patterns in audio signals. By leveraging neural networks, AI can reconstruct missing audio components and enhance overall sound quality, enabling more accurate restoration of old recordings.

Speech enhancement (SE) is a field of audio signal processing that focuses on improving the quality and intelligibility of speech signals. It covers a wide range of tasks, including noise or reverberation suppression, source separation, and removal of other distortions. These improvements are critical in applications ranging from telecommunications to assistive hearing technologies, where clear communication is essential. Depending on the adopted taxonomy, speech separation is sometimes considered a distinct research area. This survey, however, categorizes it as a subset of SE. A variety of methods have been employed to improve signal quality in the presence of distortions. Although the results have significantly improved, challenges remain, including further performance enhancement as well as practical deployment aspects, such as latency, robustness across different noise types and generalization to unseen acoustic conditions.

II. SOURCES OF AUDIO DEGRADATION

- **Environmental noise:-** Environmental noise is an accumulation of noise pollution that occurs outside. This noise can be caused by transport, industrial, and recreational activities. Noise is frequently described as 'unwanted sound'. Within this context, environmental noise is generally present in some form in all areas of human, animal, or environmental activity. The adverse effects of noise exposure (i.e. noise pollution) could include: interference with speech or other 'desired' sounds,
- **Equipment limitations:-** In recent audio equipment, induction and inflow of non-audible high frequency (HF) electrical noise to audio equipment is increasing more and more due to implementation of personal computer / network connectivity and high-resolution capability in digital audio equipment. These results show the importance of consideration of HF characteristic and countermeasure to the HF noise also in circuit design and tuning of audio equipment.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

- **Storage damage:-** The slightest damage to a sound recording immediately results in a loss of information. Its physical vulnerability dictates the precautions which we need to take if we are to ensure the preservation of acoustic source material. The recorded signal and the material of which the record is made are also important factors, so that no matter how careful one might be the risk of damage can still be relatively high. the loss of quality through normal use is relatively high in the case of mechanical sound carrying media.
- **Analog-to-digital conversion errors:-** During the analog-to-digital conversion process, quantization inaccuracy causes quantization noise. It effectively raises the noise floor and reduces the dynamic range of the ADC. According to this equation, adding one bit to the ADC's bit depth increases the SNR by around 6 dB.

III. TYPES OF DEGRADATION

- **Additive Noise:** Background hiss and hum. Additive noise is a signal corruption model where unwanted random noise is added to a desired signal, represented as ($r(t) = s(t) + n(t)$). It is independent of the original signal and is often used to model natural processes like thermal noise in communication systems or grain in images
- **Non-linear Distortion:** Clipping and saturation. Nonlinear distortion is the alteration of a signal's waveform when passing through a system where the output is not directly proportional to the input, commonly occurring in overdriven amplifiers or communication channels. It introduces unwanted, new frequency components—harmonics and intermodulation products—that were not present in the original signal, often leading to distortion and saturation
- **Temporal Dropouts:** Missing signal segments. Temporal Dropout is a regularization technique designed to prevent models from over-relying on specific, easily recognizable frames or time steps.
- **Spectral Loss:** Reduced frequency bandwidth. Spectral loss functions, commonly used in audio synthesis and neural networks, evaluate data in the frequency domain rather than the time domain to better preserve high-frequency details. This approach typically uses Fourier transforms to compare the spectrograms of real and generated signals, focusing on energy distribution rather than direct numerical similarity.

IV. OVERVIEW OF AI IN AUDIO PROCESSING

AI models process audio in different representations:

- **Waveform-based processing** Waveform-based processing in AI involves feeding raw, time-domain signal data (e.g., audio, ECG, RF) directly into machine learning models, bypassing traditional feature engineering like spectrograms. This approach is popular in audio classification and 6G communications, enables models to learn fine-grained, high-dimensional patterns directly, improving accuracy and reducing latency in applications like sound generation, seismic analysis, and semantic communication[2].
- **Spectrogram-based processing (STFT)** :- Spectrogram-based processing in AI converts 1D audio signals into 2D time-frequency visual representations, enabling powerful image-processing techniques (like CNNs) to analyze sound. By mapping frequency energy over time (e.g., mel spectrograms), AI models achieve high accuracy in speech recognition, environmental sound classification, and medical diagnostics
- **Mel-frequency cepstral coefficients (MFCCs)** :- Mel-frequency cepstral coefficients (MFCCs) are a foundational feature extraction technique in AI for representing audio, transforming sound into a numerical, compressed format that models human auditory perception. Widely used in speech recognition (ASR), speaker identification, and music analysis, they map sound frequencies to the Mel-scale, emphasizing important lower frequencies just as human hearing does.

V. AI TECHNIQUES FOR AUDIO REGENERATION

A. Denoising Auto-encoders- Autoencoders learn to map noisy inputs to clean outputs. The encoder compresses the signal, while the decoder reconstructs it.

B. Generative Adversarial Networks(GANs)- GAN uses adversarial learning

- Generator reconstructs audio
- Discriminators evaluates realism

They are particularly effective for reconstructing missing audio segments.

C. Recurrent Neural Networks(RNNs):- RNNs capture temporal dependencies, making them suitable for speech signals.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

D. Transformer Models:- Transformers use attention mechanisms to model long-range dependencies and improve reconstruction quality.

VI. OVERVIEW OF GENERATIVE ADVERSARIAL NETWORKS (GAN)

Generative Adversarial Networks are a type of generative model which has two parts: the generator G, which performs the enhancement, and the discriminator D, which classifies the input as real or fake. During training, these two networks engage in a process known as adversarial training. In this process, the generator attempts to produce realistic speech, while the discriminator continually improves its ability to distinguish clean, real speech from the generator's output. Generator G takes noisy speech y and outputs enhanced $\hat{x} = G(y)$. Discriminator D is deleted when testing is performed.

A. ADVERSARIAL OBJECTIVES

The different research [3] represented Adversarial as a minimax game with value function V (D, G)

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}}[\log D(x)] + E_{z\sim p_z}[\log(1 - D(G(z)))]$$

where z is random noise, and x is a real data sample.

Following are some common adversarial losses used with GANs for Speech enhancement

1) LEAST SQUARES LOSS:-

Least squares generative adversarial network (LSGAN) was introduced in 2016 [4]. The core idea of this work was a new loss function that would address the vanishing gradient problem and reduce the instability of training GANs. The authors used the least squares loss function rather than the sigmoid cross-entropy. The former penalizes the samples if they are far from the decision boundary, even if they are on the correct side.

2) WASSERSTEIN LOSS

Wasserstein GAN [5] is based on Earth Mover's distance. It improves the learning stability as well as avoids other common problems of GANs, such as mode collapse.

3) METRIC LOSS

As GANs started to be actively used for SE, it was noticed that sometimes they under-perform based on speech-specific metrics. To address this, Metric GAN [6] was proposed. The key difference between this approach and other loss functions is that in this case particular metrics are used to ensure that they are improved by the model.

4) RELATIVISTIC LOSS –

Relativistic standard GAN (RSGAN) [7], argues that the standard GAN, which uses binary cross-entropy loss, lacks one important quality - when the probability of fake samples being real increases, the probability of real samples being real should decrease.

5) HINGE LOSS

It improves the training stability and reduces the chance of mode collapse.

VII. GAN IN SPEECH ENHANCEMENT

GAN-based SE models are represented by three components: input–output representation, loss formulation, and model architecture. Together, these components determine the performance of the model. The relationship between these components and their relevance to model training is discussed in existing GAN-based Speech Enhancement.

A. MODEL INPUT AND OUTPUT

The choice of signal representation is the main aspect of speech-related tasks. Consequently, GAN-based SE models differ in their formulation of input and output signals. In this paper overview of the signal representations that define the model input and output, then review the training targets enabled by these representations, and model architecture are discussed.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

1) SIGNAL REPRESENTATION

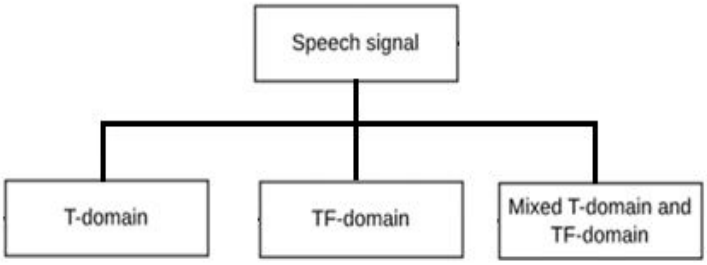


Fig.1. Types of signal representations

This paper focuses on two primary domains of signal representation: time (T) and time–frequency (TF). Accordingly, the analyzed GAN-based SE models can be grouped into three categories based on their signal representation: T-domain, TF-domain, and Mixed T-domain and TF-domain, The T-domain approach uses generative models trained to estimate clean waveform fragments from noisy ones. It has two advantages: first, it avoids the information loss introduced by feature engineering; second, it allows the phase to be embedded in the temporal domain. Although all T-domain models operate directly on raw waveforms, they differ in how waveforms are segmented and presented to the network. Early architectures used fixed-length frames and local convolutional processing which limited the temporal context to short segments. Later models use multi-scale or attention mechanisms to capture long-range temporal dependencies, eliminating the need for windowing. Earlier models operated either directly on STFT magnitude spectra [8] or on derived spectral features obtained from them, such as log-Mel [8] log-Gammatone [9],[10] or ERB-band energies [11].

2) TRAINING TARGETS

Depending on the given speech signal representation, the generator either predicts enhanced features directly or produces a mask to be applied to the noisy input. The choice of input representation is closely aligned with the training target. Models that operate on raw waveforms use a waveform-to-waveform mapping method. In this case, the generator learns a nonlinear transformation that projects the noisy waveform onto a clean one. Mask-based targets are not applicable in the time domain. In the mask-estimation approach, models incorporate phase information using a phase-sensitive mask (PSM) . Although the PSM depends on the phase difference between the clean and noisy signals, the model itself receives only the STFT magnitude. The enhanced speech is then reconstructed using the noisy phase. Because PSM-based models introduce a phase at the training target, they are often referred to as ‘‘phase-aware’’ models. Mixed models that integrate T-domain and TF-domain representations use mixed training targets that combine processing from both domains. These methods are different in nature. One group consists of waveform-generating models conditioned on TF [12]. In models the generator operates in the TF domain. It uses log-magnitude and phase as inputs to produce a clean waveform. A separate waveform-domain discriminative model provides latent conditioning features through masked multi-head attention. Some methods fuses TF-domain enhancement and TF-domain residual noise estimation to predict the final complex spectrogram. Lastly, GAN-in-GAN uses an inner GAN to predict the final complex spectrogram and an outer GAN to reconstruct the enhanced waveform.

B. MODEL ARCHITECTURES

1) GAN FRAMEWORK

GAN-based models can be categorized based on the following criteria:

1) How the discriminator receives an input

2) Whether the model performs one-directional or bidirectional domain mapping

According to these two criteria, GAN-based SE methods employ the following types: Vanilla GAN, Conditional GAN, and CycleGAN.

In the Vanilla GAN models, the discriminator receives only one input at a time: clean speech for real examples and enhanced speech for fake examples. It has no access to the noisy signal; the discriminator evaluates whether each input sample matches the clean speech distribution.

A conditional generative adversarial network (cGAN) consists of a generator and a discriminator, both of which receive auxiliary conditioning inputs. The discriminator is typically provided with input in the form of pairs of (noisy, clean) and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

(noisy, enhanced) for real and generated data, respectively. Some studies use cGAN terminology because the generator is conditioned on auxiliary information [9], [10]. In this review, we overlook the original formulation of conditional GANs , which states that a model is a cGAN only when the generator and discriminator are both conditioned on the same side information. Models in which the discriminator remains unconditional and receives only clean versus enhanced signals without paired conditioning inputs are classified as Vanilla GANs.

The third uses a cycle-consistent framework that enables training with unpaired data through a bidirectional mapping between noisy and clean domains. One generator maps noisy speech to clean speech, and a second generator reconstructs the noisy domain from the enhanced output. cGANs are the most widely adopted GAN architecture for SE. This is because the task naturally involves conditional generation from noisy observations. It is important to note that although metric adversarial objectives modify the discriminator objective by replacing real-fake classification with metric prediction , the generator-discriminator structure remains unchanged. Therefore, these methods remain standard conditional GAN-based SE models.

Early approaches directly condition the generator on the noisy signal to learn a deterministic enhancement mapping [9], [10]. The μ-law SGAN modifies the discriminator with a trainable spectrum compression layer [86]. HiFi-GAN models, on the other hand, can be viewed as a neural vocoders [12]. In these models, noisy speech is first converted into acoustic features and then re-synthesized into a waveform, and the discriminator only assesses speech naturalness. In all Vanilla GAN models, the discriminator remains unconditional.

The CycleGAN framework, first introduced for unpaired image translation , has been adapted for SE to address the scarcity of parallel noisy-clean data. These models have been used to learn mappings between noisy and clean speech without requiring parallel recordings [15]. These models have also shown competitive results when parallel data was available. Studies have employed the CycleGAN framework under paired training conditions only, demonstrating its ability to preserve speech structure and minimize distortion[14]. Several architectural improvements have been proposed for each GAN framework.

2) GENERATOR ARCHITECTURES

The GAN model was first introduced to the SE task with SEGAN. SEGAN consists of a single-stage convolutional encoder-decoder generator with skip connections. Since this pioneering work, numerous architectural improvements have been proposed. In this subsection, subsequent studies are synthesized into categories that reflect architectural improvements introduced to neural blocks of the generator. Later, baseline encoder-decoder models improved upon the fundamental feed-forward architecture. These improvements included multi-scale convolutions , residual blocks within encoder-decoder stages and gated linear units. Despite these refinements, the models retain a single-pass, feed-forward encoder-decoder structure without explicit temporal modeling or an attention module.

Encoder–Decoder with Recurrent layers was first introduced in CRGAN to have a convolutional recurrent network incorporated into its encoder-decoder architecture. These layers are added to encoder-decoder models to capture long-term temporal dependencies. The latest model in this category, MOS-GAN , uses a dual-path block between the encoder and the decoder, with two RNN layers to capture frequency features and temporal dependencies. Encoder–Decoder models with Attention module have a SEGAN-like encoder-decoder architecture with skip connections. Convolution remains the primary operator, while the attention module modifies the flow of features. However, the models with the attention module differ in terms of where the network focuses its attention. The focus can be on the decoder side only on the encoder-decoder side or distributed across layers [14],[15]. Some research focus on using attention to capture global time-frequency dependencies during non-parallel training[14], [15]. Others use it to restore the formant structure distorted by reverberation. Additionally, some approaches use attention as a lightweight mechanism for reweighting features. Encoder-decoder models with sequence modeling incorporate either transformers or conformers, or SSMs, while maintaining a convolutional encoder-decoder. Sequence modeling modules are introduced at the bottleneck. Among the different models, some research employ a transformer architecture. It was a standard transformer architecture. Subsequent studies adopted speech-specific sequence modeling architectures, such as Dual-Path Transformers and Conformers, that account for the time-frequency structure and local temporal patterns.

Dual-path transformers stop treating speech as one long sequence and instead split the modeling process into two axes: time and frequency. A dominant pattern is that transformers are placed in the bottleneck between convolutional encoders and decoders [13]. The CRG-MGAN model is described as an improved transformer structure that integrates convolutional layers, recurrent networks, and spatial convolutional gating units to capture temporal-frequency

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

dependencies. It is important to note that two-path transformer models are presented as a way to balance performance and efficiency. According to the authors, structural decomposition controls model size and computation cost [13]. Unlike standard transformers, which rely solely on self attention, Conformer architectures incorporate convolutional modules that capture local temporal structures. Some studies extend the baseline Conformer architecture by using cascaded or multi-scale . Others retain the original CMGAN generator and focus exclusively on loss design or training target . MRGAN is a lightweight hybrid design that combines time-domain Conformer blocks with frequency-domain Transformer blocks. One sequence modeling model has begun integrating Mamba, a computationally efficient selective state-space model (SSM) that builds on structured state-space architectures. Unlike Transformers, which have quadratic complexity, Mamba scales linearly with sequence length. In MambaGAN ,linearity is achieved through the incorporation of Dual-Path Mamba Former (DPM) modules into the central processing stage of the generator. These DPM blocks are connected with an encoder and two separate decoders for magnitude-mask and phase estimation. Many generative adversarial network (GAN)-based SE systems use encoder-decoder generators to reconstruct enhanced signals. However, in some studies, the generator is implemented as a sequence model with a linear decoder that directly maps temporal features to enhancement parameters. These models are classified as models with non-encoder decoder generators. This use is aligned with metric-driven frameworks, such as MetricGAN [11]. Some research that do not follow an encoder-decoder architecture use synthesis-based generators that directly generate wave forms [12]. Despite their descriptions using encoder-decoder terminology, some earlier works are treated as non-encoder-decoder models in this review because their generators implement direct mappings rather than explicit encoder-decoder structures [9].

DUAL-STREAM ENCODER–DECODER GENERATORS - A dual-stream generator, also known as a forked generator, is an architectural topology in which a single generator contains multiple parallel processing branches. The terms dual-stream and forked are used interchangeably. In dual-stream generators, the decoding stage is split into multiple parallel branches, each operating on the exact encoded representations. In some models, the magnitude branch directly predicts the enhanced magnitude spectrum, whereas some estimate a magnitude mask. Although several GAN-based SE models are described as complex-valued, their generators usually have a standard encoder-decoder structure. For this reason, this survey treats complex-valued operations as a representational choice rather than a distinct architectural category. Additionally, it is worth noting that dual-stream generators should not be confused with dual-path sequence-modeling architectures. The first ones are architectures featuring parallel decoder branches, while the second ones process features sequentially along different dimensions without branching.

3) DISCRIMINATOR ARCHITECTURES

Generators model the long-term temporal dependencies inherent in speech signals to generate natural-sounding output, while the discriminators primarily detect differences between generated and natural speech. Consequently, discriminator architectures tend to be simpler than generator architectures. The architecture of discriminators in SE remains convolutional. More recent studies have introduced architectural refinements such as deformable convolutions, multi-scale inputs, and auxiliary attention modules. Recent discriminators incorporate multi-scale evaluation. This can be achieved through multi-spectrogram discriminators, which assess several STFT representations independently via parallel CNN branches , or through a combination of global and local discriminators, which operate on full utterances and short segments, respectively. Some research [13]replaced the standard convolution layers in the discriminator with deformable convolution layers. This introduced kernel offset prediction, enabling adaptive receptive fields in the time-frequency domain. In this design, the sampling locations adjust dynamically based on the input reverberant speech. The discriminator retains a convolutional structure by replacing fixed-grid convolution with deformable convolution blocks. The attention mechanisms that appear in discriminators are always auxiliary. They serve as feature-reweighting modules embedded within convolutional architectures, rather than as attention-based discriminators. In addition to architectural diversity, computational complexity has become a more relevant consideration for deployment. Early convolutional encoder-decoder models, particularly waveform-based SEGAN variants, tend to be lightweight. However, recent transformer, conformer, and dual-path architectures are more robust, but they increase model size and computational cost due to attention mechanisms and multi-branch processing .

VIII. CONCLUSION

This paper overviews GANs and how they have been applied for Speech Enhancement task. Several research work in this area show that GANs retain their popularity over the years. The main goal of this review is to introduce GANs for Speech Enhancement and provide sufficient information to assist researchers with interested in employing GANs for their Speech Enhancement tasks. The researchers have moved on from conventional representations, such as waveform or magnitude, to more advanced ones that jointly model magnitude and phase. The analysis shows that the MetricGAN

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505159|

loss, which was intended for speech domain, and its variations are still widely used. The trend shows the switch from baseline convolutional encoder-decoder to encoder-decoder and sequence modeling. Although there are significant advantages, there is still room for improvement. Different directions can be taken, ranging from datasets, computational cost reduction, real life application, increasing robustness to different types of distortions, to utilizing phase information, researching capabilities of state-space models, improving discriminator architectures, and designing robust perceptually guided training strategies.

REFERENCES

1. Sound quality degradation of audio equipment by hum noise induced by non-audible high frequency electrical noise- D Kobayashi and T Yoshida
2. Raw Waveform-based Audio Classification Using Sample-level CNN Architectures- Jongpil Lee, Taejun Kim, Jiyoung Park, Juhan Nam
3. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘‘Generative adversarial nets,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–14.
4. X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, ‘‘Least squares generative adversarial networks,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2813–2821.
5. M. Arjovsky, S. Chintala, and L. Bottou, ‘‘Wasserstein genera tive adversarial networks,’’ in Proc. 34th Int. Conf. Mach. Learn., Aug. 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr. press/v70/arjovsky17a.html
6. S. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, ‘‘MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,’’ in Proc. 36th Int. Conf. Mach. Learn., Jun. 2019, pp. 2031–2041. [Online]. Available: https://proceedings.mlr.press/v97/ fu19b.html
7. A. Jolicoeur-Martineau, ‘‘The relativistic discriminator: A key element missing from standard GAN,’’ 2018, arXiv:1807.00734.
8. D. Michelsanti and Z.-H. Tan, ‘‘Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,’’ in Proc. Interspeech, Aug. 2017, pp. 2008–2012.
9. M. H. Soni, N. Shah, and H. A. Patil, ‘‘Time-frequency masking-based speech enhancement using generative adversarial network,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 5039–5043.
10. N. Shah, H. A. Patil, and M. H. Soni, ‘‘Time-frequency mask based speech enhancement using convolutional generative adversarial network,’’ in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Nov. 2018, pp. 1246–1251.
11. H. Li and J. Yamagishi, ‘‘Multi-metric optimization using genera tive adversarial networks for near-end speech intelligibility enhance ment,’’ IEEE/ACM Trans. Audio, Speech, Language Process., vol. 29, pp. 3000–3011, 2021.
12. P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, ‘‘HIFI++: A unified framework for bandwidth extension and speech enhancement,’’ in Proc. ICASSP- IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2023, pp. 1–5.
13. H.-T. Chiang and J. H. L. Hansen, ‘‘A deformable convolution GAN approach for speech dereverberation in cochlear implant users,’’ in Proc. Interspeech, Aug. 2025, pp. 833–837.
14. G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, ‘‘CycleGAN-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism,’’ in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Dec. 2021, pp. 523–529.
15. G.Yu,Y.Wang,H.Wang,Q.Zhang,andC.Zheng,‘‘Atwo-stagecomplex network using cycle-consistent generative adversarial networks for speech enhancement,’’ Speech Commun., vol. 134, pp. 42–54, Nov. 2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details