IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 January 2026, accepted 12 February 2026, date of publication 23 February 2026, date of current version 26 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3667063

SURVEY

Generative Adversarial Networks in Speech Enhancement: A Survey

JUSTINA RAMONAITĖ, (Graduate Student Member, IEEE), GRAŽINA KORVEL, (Member, IEEE), AND GINTAUTAS TAMULEVIČIUS, (Senior Member, IEEE)

Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, 08412 Vilnius, Lithuania

Corresponding author: Justina Ramonaitė (justina.ramonaite@mif.vu.lt)

This work was supported by the Research Council of Lithuania (LMTLT) under Agreement S-MIP-24-118.

ABSTRACT Generative adversarial networks are a powerful type of model in deep learning. They have been successfully applied within different domains. This review focuses on the usage of generative adversarial networks for speech enhancement. In total, 87 studies are analyzed and summarized, with their publication period ranging from the earliest attempts to papers released in November 2025. This survey aims to provide the necessary background information for researchers planning to use or already applying generative adversarial networks to enhance speech signals. It examines generative adversarial network-based models by analyzing signal representations at the input and output, network architectures, and most importantly, loss function formulations. Temporal trends in these design decisions are analyzed to illustrate the evolution of the models over time. The surveyed models are further compared based on their reported performance on standard benchmark datasets. This evaluation helps identify which models achieve the best performance for specific speech enhancement tasks addressed in the literature. The limitations and future research directions reported in the surveyed studies are summarized, along with additional insights derived from model analysis.

INDEX TERMS Generative adversarial networks, speech enhancement, survey.

I. INTRODUCTION

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.

Generative adversarial networks (GANs) were introduced in 2014 [1]. The main novelty of this work was the proposal of adversarial training. In general, traditional GANs consist of two networks: the generator G and the discriminator D. The generator G takes random noise as input and generates new data intended to be similar to the target data. The discriminator D then evaluates if the provided input, which is either the real data or data generated by G, is real or fake. Typically, D is a classifier that outputs either a binary label or a probability. The goal of G is to generate data so that D recognizes it as real, whereas the goal of D is to distinguish real data from fake data accurately. The adversarial objectives of G and D correspond to a two-player minimax game. The conditional GAN (cGAN) [2] was proposed in the same year. The main difference between cGAN and the traditional GAN is the input of the model. Additional information is provided to cGAN alongside random noise in order to condition the

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

networks. GAN and cGAN were both introduced in the image domain and then successfully applied to the speech signal domain.

The first attempts to use GANs for SE were published in 2017 [3], [4]. SEGAN [3], often cited as the pioneering work on GANs in SE, adapted cGAN with the least-squares generative adversarial network (LSGAN) loss for raw waveform denoising. In contrast, a different approach was taken in [4], which used spectrograms as input to the Pix2Pix framework, which requires no random input noise. Since then, GANs have remained popular in SE research. Many extensions have been proposed, often focusing on the choice of input representation, improving generator and discriminator architectures, stabilizing the adversarial training process, or introducing task-specific objectives, such as perceptual, metric-based, or intelligibility-oriented losses.

This survey provides an overview of GAN-based SE methods, organizing them by their core design components. Firstly, we review adversarial loss formulations for SE, as the adversarial objective distinguishes GAN-based approaches. To highlight their differences and practical implications, we present the mathematical formulations of commonly used loss functions and provide a quantitative summary of their adoption in existing studies.

Secondly, the representations of the speech signal as the model input and output are discussed. Speech signals can be represented in multiple domains, including time-domain waveforms, spectral magnitudes, and complex-valued time-frequency representations. These representations allow to formulate the SE task with different training targets, such as direct signal mapping or mask estimation, which directly determines what the generator is trained to estimate and reconstruct. Motivated by the choice of representations and target formulations, we introduce a model taxonomy organized by signal representation and training target.

While analyzing model architectures, we distinguish between GAN architectures, which define the organization of the adversarial framework, and network-level architectures, which describe the neural designs of the generator and discriminator. Furthermore, to evaluate the existing models, we summarize their performance across widely used benchmark datasets and evaluation metrics.

The survey is organized as follows: Section II summarizes previous surveys that relate to SE or GANs within this area of research, Section III outlines the approach that was used for collecting papers, Section IV provides the mathematical background of different adversarial loss functions that were employed in the selected publications. Section V contains all the main aspects that are considered in this survey, starting with the formulation of inputs and outputs of the model and the corresponding taxonomy in Subsection V-A, continuing with the presentation of trends in adversarial losses in Subsection V-B, the evolution of architectures together with the taxonomy of generator architectures in Subsection V-C and lastly the design and training targets of the model in Subsection V-D. Section VI is dedicated to the comparative analysis which presents the tasks, dataset, metrics and contains the summary of reported results. The survey is finalized with the discussion of limitations and future directions in Section VII and conclusions in Section VIII.

II. RELATED WORKS

This section describes previous reviews on SE published between 2020 and 2025. The papers from this period covered various perspectives, ranging from more general works that discussed a variety of methods to those that focused on specific architectures or other narrower topics.

A detailed examination of the evolution of frequency domain monaural SE over a period of sixty years, comparing traditional methods and deep learning methods, was given in publication [5]. This paper provided an in-depth analysis of technological advancements in the field. The review is notable for providing metric comparisons using datasets adapted for both normal-hearing and hearing-impaired listeners. Another survey, which contained all types of SE algorithms, including statistical, deep learning, and hybrid models, was provided in [6]. This systematic review covered 47 studies from 2015 to 2024. The main questions that were answered were related to the most common approaches, the transforms (e.g. STFT, DCT) that were employed, the number of channels of the signal. Other questions included the models used to identify speech and noise (e.g. Gaussian), what datasets were used and what evaluation metrics were chosen. Lastly, the review concluded with a discussion of challenges and limitations.

A similar analytical depth can be found in review [7], which focused on deep neural network (DNN) techniques for SE and separation. This work focused more on the theoretical and practical aspects of applying DNN models without a detailed exploration of specific objective metrics. In contrast, review [8] explored the general trends in deep learning techniques, from early DNNs to sophisticated models such as CNN and GAN, providing an overview of these developments without detailed analysis. Similarly, publications from 1993-2022 that focused on applications of deep neural networks for supervised SE were reviewed in the article [9]. This work provided a detailed guide on DNNs, ranging from multi-layer perceptrons to GANs, including descriptions of the main types of models, training algorithms and targets, acoustic features, databases and activation functions. The authors also highlighted the strengths and weaknesses of DNNs for SE and listed possible future directions and remaining challenges. Survey [10] covered a wide range of methods as well, it added to the technical discussions found in the other reviews provided and enriched the dialogue by providing critical perspectives on perceptual assessment and computational efficiency.

Other reviews were narrower and focused on a particular area of SE. Multi-objective methods were surveyed in the paper [11]. The literature review was complemented by a comparative analysis of selected representative methods.

Review [12] was concentrated on the cocktail-party problem, which includes speech separation. Both of these papers covered conventional as well as deep learning methods. Survey [13] focused on multi-channel SE using machine learning and deep learning algorithms. This survey included works from a variety of journals, challenges, and books coming from different countries. Paper [14] focused on latency, which is crucial for real-life applications. The author presented different layers that are used and different methodologies that are applied in order to reduce the amount of resources required by SE models.

Some surveys were dedicated to specific advanced architectures. In the paper [15], the use of transformer models for SE, among other applications such as speech recognition or synthesis, was discussed. Meanwhile, survey [16] looked at the application of the U-Net architecture to audio enhancement tasks only. It focused on how this architecture is applied to different types of audio, including speech, and demonstrated the effectiveness of U-Net in improving audio quality.

Overviews dedicated to generative networks were [17], [18], [19]. Two of them, [17], [18], focused only on GANs. The survey [17] covered multiple topics in which GANs are used in speech, including speech synthesis, SE in a broader sense, and speech augmentation. The SE part included GANs designed for denoising. They were discussed in chronological order, from their initial use in 2017 to 2021. Specific models were named, and their inputs, architectures, and loss functions were described. Their advantages and disadvantages were listed, and they were compared based on their relation, for example, loss functions. Different variants or modifications of previous models were also presented visually. While denoising is the main focus of our review, it is worth noting that survey [17] covers many more topics that fall under SE. GANs are used for the improvement and enhancement of synthesized speech, to make it sound more natural, voice conversion, for example, transferring voice to another language, or emotion conversion, correction of impaired speech to make it intelligible. The authors introduced used datasets and evaluation metrics, however, they were presented in one pool without any separation.

The other survey [18] focused only on the application of GANs for SE. This overview referenced papers from 2017 to 2022. A lot of attention was paid to architectures, their development over the years, as well as their advantages and disadvantages. Commonly used datasets were indicated and evaluation metrics were compared. Although this paper concisely presented the main aspects of the application of GANs in SE, it was quite short and did not go into great details or provide a comparison based on other criteria than architecture.

To our knowledge, the most recent review of generative networks used for SE is the paper [19], which was published in 2024. The scope of this survey covered a broader spectrum of generative models, including autoencoders, diffusion models, and GANs. This work provided an overview of each type of model, including descriptions, highlighted notable advancements, and addressed advantages and limitations. A comparison of these generative models in terms of quality, diversity, and speed was also provided. The authors touched on the technical aspects as well, they introduced the main tools that are employed when working with generative SE models, as well as signals in general. Despite providing a broad overview of generative models for SE, the survey [19] omits several essential aspects. Firstly, it organizes the literature by model family (autoencoders, GANs, and diffusion models) and focuses on high-level methodological trends. However, it does not provide a detailed analysis of GAN models. Secondly, the survey does not discuss the architectural components in detail, thus, it remains unclear how specific architectural decisions shape the overall model. Thirdly, it does not involve a metric-based comparison of models for specific SE tasks.

Our survey focuses on low-level methodological trends and provides a detailed introduction to the use of GANs for SE. It takes the position that the design of GAN-based SE systems is influenced by interactions among model architecture, signal representation, and training objectives. Based on this position, our survey analyzes architectural and methodological design choices and categorizes existing approaches into coherent taxonomies that trace their evolution over time. In addition, the survey presents a comparative analysis of evaluation metrics and reported performance across existing studies.

The main questions addressed in this survey are:

- What has been already done and what are the current trends when it comes to signal representations that are used as input and output of the models, the architectures and adversarial losses?
- What datasets are used for training and evaluation of the models and what results have been achieved with the benchmark dataset?
- What limitations and open challenges remain in current GAN-based approaches?

This review, to our knowledge, is the first to provide such an in-depth analysis of GANs in SE. Other surveys either provided a more general overview that covered multiple types of models or focused on a particular topic. Those who did discuss GANs provided their reviews in a very concise manner, introducing GANs for enhancement in other areas, such as speech synthesis, or in combination with other generative approaches.

III. METHODOLOGY

All publications were collected using the *Publish or Perish* tool with *Google Scholar* as the data source. The initial search was performed by filtering words that occur in the article title. These combinations were formed from 3 key components: *speech* **AND** *enhancement* OR *denoising* OR *separation* OR *dereverberation* **AND** *GAN* OR *GANS* OR *generative adversarial*.

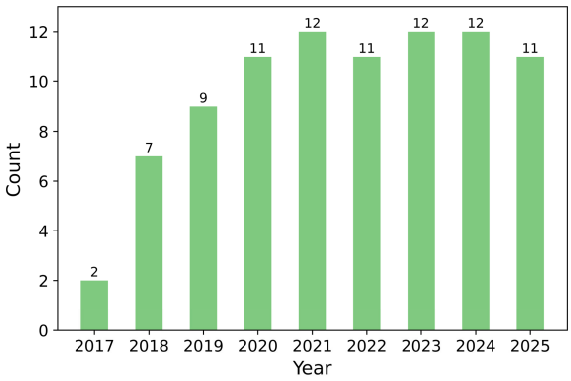

FIGURE 1. Number of included publications per year from 2017 to 2025 November.

The initial search was conducted in January 2025 and repeated in November 2025 in order to refresh and update the collected references. The list of published works, including journal articles, conference proceedings, and preprints, was obtained, and the number of citations for each work was fixed. This final list contained 189 items after removal of duplicates.

In order to ensure inclusion of only peer-reviewed material, the publication screening process began with the elimination of works unrelated to GAN-based SE and preprints. Note that some papers appear as preprints in *Google Scholar*, however they have been published as journal articles or conference proceedings, for example [3]. Such papers were not eliminated in this step. The citation rate per year (CPY) was then calculated for the remaining publications. Studies published up to 2023 were included if their $CPY \geq 3$. The more recent publications were included based on relevance and novelty, as their citation period was too short for *CPY* evaluation, given the time-dependent nature of citation-based metrics. During the screening process, 104 papers were eliminated. Lastly, 2 papers that did not match the main search criteria, namely [20] and [21], were added to the list due to their relevance to the topic.

Fig. 1 shows the number of selected publications per year.

IV. GAN DEFINITION

GANs are a type of generative model consisting of two parts: the generator G, which performs the enhancement, and the discriminator D, which classifies the input as real or fake. This helps G improve and generate realistic output. 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 discarded when testing is performed.

A. ADVERSARIAL OBJECTIVES

The unique aspect of GANs is adversarial training. The authors of the original paper [1] represented it as a minimax game with value function $V(D, G)$

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

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

cGAN differs from (1) by the additional input, thus, the formula is modified to

$$\min_G \max_D V(D, G) = \mathbb{E}_{(x,y)\sim p_{data}}[\log D(x, y)] + \mathbb{E}_{y\sim p_{data}, z\sim p_z}[\log(1 - D(G(z, y), y))], \quad (2)$$

where y is the additional information. In the case of SE, it would usually be the noisy signal y.

The loss used to train GANs has evolved significantly since then. Usually, the adversarial loss is extended by reconstruction loss and other components if they are required. New functions that define the adversarial loss have been introduced as well. This section describes 5 most common adversarial losses used with GANs for SE: least squares loss [22], Wasserstein loss [23], [24], metric loss [25], relativistic loss [26] and hinge loss [27]. Least squares, relativistic and hinge losses mainly aim to stabilize training and improve the quality of the output. Wasserstein loss stabilizes training and provides interpretable learning curves. Metric loss focuses on optimizing generator with respect to selected evaluation metric.

1) LEAST SQUARES LOSS

Least squares generative adversarial network (LSGAN) was introduced in 2016 (the paper was updated in 2017) [22]. 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. The final loss function for LSGAN was defined as

$$\min_D V_{LSGAN}(D) = \mathbb{E}_{x\sim p_{data}(x)}[(D(x) - b)^2] + \mathbb{E}_{z\sim p_z(z)}[(D(G(z)) - a)^2]$$
$$\min_G V_{LSGAN}(G) = \mathbb{E}_{z\sim p_z(z)}[(D(G(z)) - c)^2], \quad (3)$$

where a is the label of fake data, b is the label of real data and c is the label that G should have for fake data if it is successful. As identified by the authors, there are two ways to choose the values for a, b and c:

- they need to satisfy $b-c = 1$ and $b-a = 2$ OR
- $b = c$, so that G generates such samples that D considers them as real.

The conditional version of LSGAN would look similar to (3), except there would be additional input as in (2).

2) WASSERSTEIN LOSS

Wasserstein GAN [23] 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.

The Earth Mover's distance is defined as

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \left[\|x-y\|\right], \tag{4}$$

where, as described in [23], $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$.

This distance can be approximated via parameterized family of functions $\{f_w\}_{w\in\mathcal{W}}$ that are all K-Lipschitz for some K

$$\max_{w\in W} \mathbb{E}_{x\sim p_{data}(x)}[f_w(x)] - \mathbb{E}_{z\sim p_z(z)}[f_w(G(z))]. \tag{5}$$

The critic f_w corresponds to the discriminator, so the final objective can be written as

$$\min_G \max_D V_{WGAN}(D, G) = \mathbb{E}_{x\sim p_{data}}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z))] \tag{6}$$

Note that, in order to satisfy all the requirements, additional actions such as weight clipping are performed when applying this in practice. The authors of [24] identified undesirable behavior associated with weight clipping and proposed WGAN-GP, which replaces weight clipping with a gradient penalty.

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, MetricGAN [25] 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. They are defined as

$$\begin{aligned}\min_D V_{MetricGAN}(D) &= \mathbb{E}_{(x,y)\sim p_{data}}[(D(y, y) - 1)^2 \\ &\quad + (D(G(x), y) - Q'(G(x), y))^2] \\ \min_G V_{MetricGAN}(G) &= \mathbb{E}_{(x,y)\sim p_{data}}[(D(G(x), y) - s)^2],\end{aligned} \tag{7}$$

where $0 \le Q'(G(x), y) \le 1$ is the function that represents the selected metric mapped to the interval [0, 1], s is the desired score of generated data. Note, this loss was proposed directly for speech signal tasks, thus, the form of the loss function is conditional. The authors also discarded the input noise vector z, so it is not represented here. The discriminator does not classify whether the data are real or fake but instead evaluates a score. As it is intended to mimic the chosen metric, the input also includes a different pair, rather than clean & noisy and enhanced & noisy, it uses clean & clean and clean & enhanced. The authors presented this work specifically for SE, thus, they considered PESQ and STOI evaluation metrics.

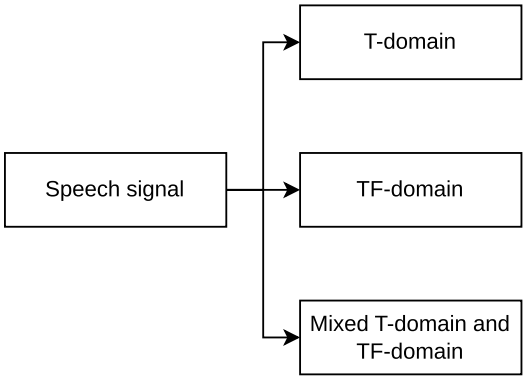

FIGURE 2. Types of signal representations.

4) RELATIVISTIC LOSS

Relativistic standard GAN (RSGAN), introduced in [26], 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.

The loss functions are defined as

$$\begin{aligned}\min_D V_{RSGAN}(D) &= \mathbb{E}_{x\sim p_{data}, z\sim p_z}[\log(\sigma(C(x) - C(G(z))))] \\ \min_G V_{RSGAN}(G) &= \mathbb{E}_{x\sim p_{data}, z\sim p_z}[\log(\sigma(C(G(z)) - C(x)))],\end{aligned} \tag{8}$$

here C represents the output of the last layer of the discriminator before the sigmoid function is applied, σ is the sigmoid function.

The same paper [26] introduced more versions of the relativistic loss functions, including relativistic average GAN (RaGAN) loss and relativistic average LSGAN loss (RaLSGAN), which were successfully applied to SE.

5) HINGE LOSS

Hinge-type loss was introduced in [27]. The authors showed that it improves the training stability and reduces the chance of mode collapse. The general hinge loss function is

$$\begin{aligned}\min_D V_{Hinge}(D) &= \mathbb{E}_{x\sim p_{data}}[\max(0, 1\text{-D}(x))] \\ &\quad + \mathbb{E}_{z\sim p_z}[\max(0, 1 + D(G(z)))] \\ \min_G V_{Hinge}(G) &= -\mathbb{E}_{z\sim p_z}[D(G(z))].\end{aligned} \tag{9}$$

V. GAN IN SPEECH ENHANCEMENT

GAN-based SE models are characterized by three key design components: input–output representation, loss formulation, and model architecture. Together, these components determine the performance of the model. In this section, we analyze the realization of these three components in existing GAN-based SE papers. The relationship between these components and their relevance to model training is then discussed.

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 the following subsections, we first discuss the signal representations that define the model input and output, then review the training targets enabled by these representations, and finally describe the training objectives used to optimize them.

1) SIGNAL REPRESENTATION

This survey considers 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, Fig. 2.

The T-domain approach uses generative models trained to estimate clean waveform fragments from noisy ones. Authors using this approach emphasize two advantages: first, it avoids the information loss introduced by feature engineering; second, it allows the phase to be embedded in the temporal domain. Among the 87 analyzed studies, 27 models operate in the time 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, such as [3] and [28], used fixed-length frames and local convolutional processing, which limited the temporal context to short segments. Later models, such as [29] and [30], [31], use multi-scale or attention mechanisms to capture long-range temporal dependencies, eliminating the need for windowing. These architectural distinctions are discussed in detail in Section V-C.

The majority of the analyzed studies (54 out of 87) operate in the TF-domain. These models process short-time Fourier transform (STFT) representations, and they differ primarily in how they handle the phase component of the STFT.

Earlier models operated either directly on STFT magnitude spectra (e.g. [4], [32]) or on derived spectral features obtained from them, such as log-Mel [4], [33] log-Gammatone [34], [35] or ERB-band energies [36]. The main difference between these features is their frequency resolution (linear or perceptual) and magnitude scaling (linear or logarithmic). However, all of them are derived from the STFT magnitude and reuse the noisy phase for waveform reconstruction. A number of later magnitude models retain this formulation while incorporating phase information only in the training target ([37], [38]). This extension is discussed in the following subsections.

The shift toward input representations that included phase began around 2021 with the introduction of CDGAN [39], which replaced magnitude features with the real and imaginary components of the STFT. The reviewed articles present two approaches to using the phase as input. The first approach involves providing the generator with magnitude-phase pairs and treating them as separate input channels (e.g., DisCoGAN and Phase-Aware MetricGAN). The second one uses the real and imaginary spectral components of the complex STFT as two separate input channels. Within this second approach, some models use only the real and imaginary representation of the complex spectrum [39], while others extend it by also including the magnitude as an additional (third) input channel [40], [41]. It is important to note that including the phase as input does not mean that the model will reconstruct the phase at the output. Many complex-domain models still rely on the noisy phase during waveform reconstruction. Whether or not the phase is reconstructed depends on the training target.

The remaining 6 models in this survey are classified as mixed T-domain and TF-domain models, because they integrate representations of both domains. Two of these models, [42], [43], are multi-input and use inputs from both domains. In [42], the waveform is improved by conditioning on MFCC features. In contrast, [43] improves the TF domain by conditioning on latent representations extracted from the noisy waveform. The other models only use the noisy waveform as input. However, they are considered mixed because they process two different representational domains (time and TF) inside the generator. In this case, the TF representation is generated within the generator rather than provided as input. These models differ in how they internally combine time and TF information. SCP-CMGAN [44] converts the waveform to a spectrogram and performs enhancement entirely in the TF domain. GAN-in-GAN [45] uses two separate GANs, one operating in the TF domain and the other in the time domain. HiFi++ [21], and M-DGAN [46] use fully integrated mixed generators.

It is important to note that not all models that compute spectrograms internally are considered mixed T-domain and TF-domain models in this survey. For instance, the time-domain GAN by [47] operates only on the waveform, and STFT layers are used to compute the mask loss [47]. Similarly, the original HiFi-GAN [48] is considered a waveform-to-waveform enhancement model. In this model, the Mel spectrogram discriminator and the spectrogram losses provide only adversarial feedback. Since the given models do not use spectral features as part of the generator's input, they are not considered mixed models.

The annual counts of the different signal representations used as model inputs in research from 2017 to 2025 are shown in Fig. 3.

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.

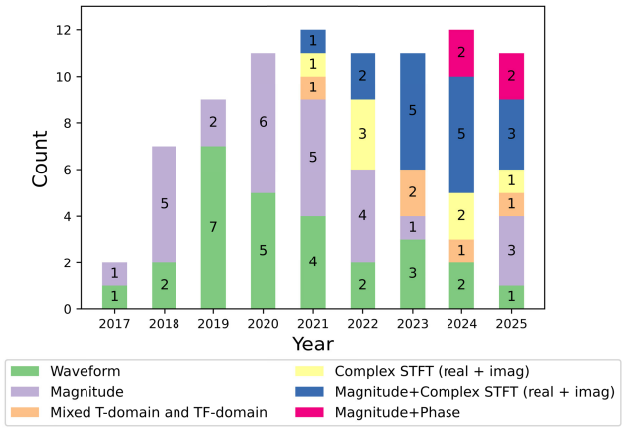

FIGURE 3. Stacked bar plot representing the usage of different model inputs by years.

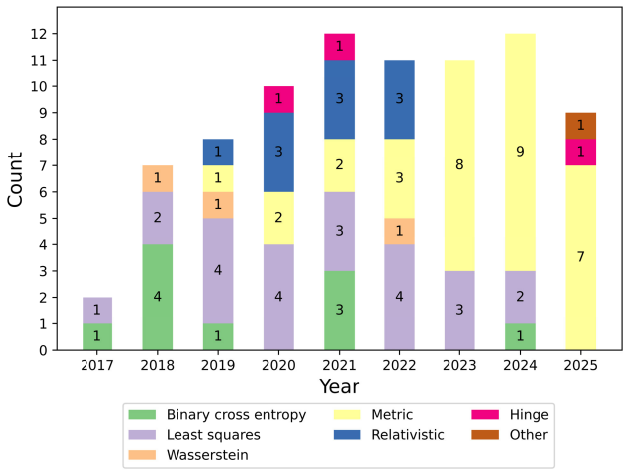

FIGURE 4. Stacked bar plot representing the usage of different adversarial losses by years.

When using the spectrogram magnitude as input, models employ either magnitude mapping or magnitude mask estimation. In the mask-estimation approach, models incorporate phase information using a phase-sensitive mask (PSM) [37], [38]. 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. In addition to these two categories, some papers [34], [35] describe implicit mask learning models. These models use a generator with a sigmoid activation function in the output layer. This produces a data-driven TF representation that yields a mask-like structure.

Models employing complex-valued STFT inputs build upon earlier magnitude-only approaches by extending them to the complex domain. These models allow the generator to process both the real and imaginary components simultaneously. They differ in their training target and can either learn a complex-to-complex mapping [39], [49] or predict a complex ratio mask (CRM) [50], [51]. The paper of [52] appears as an exception within this set. Although this strategy benefits from complex-domain input, the model only predicts a magnitude mask.

The training target for models that incorporate the magnitude channel alongside the real and imaginary components is as follows: first, a magnitude mask is estimated, then, the complex residuals are predicted. The magnitude, real part, and imaginary part of the STFT are unified inputs. An exception to this category is the model of [95] that, despite also using magnitude, real, and imaginary components as inputs, predicts only an enhanced magnitude spectrum, using the complex components as auxiliary features. Almost all analyzed models within this category reuse the noisy phase in the reconstructed signal. Only two of the analyzed models [92], [93] produce a clean phase. This is because these models use a two-stage CycleGAN design that first estimates the magnitude and then maps directly to clean real and imaginary components. The architectural differences behind these behaviors will be discussed in the Section V-C.

Another group of studies uses magnitude and phase as inputs. These models predict phase directly, bypassing complex residual or CRM-based enhancement. While making phase predictions, these models estimate amplitude either by direct mapping [106], [107], [108] or by applying a mask to the magnitude component [109].

Mixed models that integrate T-domain and TF-domain representations use mixed training targets that combine processing from both domains. These strategies differ. One group consists of waveform-generating models conditioned on TF features ([21], [42]). In [43], 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. [46] fuses TF-domain enhancement and TF-domain residual noise estimation to predict the final complex spectrogram. Lastly, GAN-in-GAN ([45]) uses an inner GAN to predict the final complex spectrogram and an outer GAN to reconstruct the enhanced waveform.

Table 1 summarizes all studies included in our analysis and organizes them by their signal representation and training target. The table includes all the studies analyzed in our review. This allows for a direct comparison of how different approaches align with specific signal representations and training targets.

B. OBJECTIVE FUNCTIONS

The cost function is one of the key components of GAN, as an appropriate choice can improve performance and provide other benefits, such as stable training and faster convergence. This section provides an overview of the practical application of various adversarial losses, as well as complementary components that are included in the final objective.

TABLE 1. Taxonomy of signal representations and estimation targets in GAN-based SE.

Signal representation	Training target	Year	Models
Waveform	Waveform mapping	2017	SEGAN [3]
		2018	Adapted SEGAN [53], WCGAN-GP [28]
		2019	WGAN [54], SERGAN [55], GSEGAN [56], UNetGAN [57], Generalized SEGAN [58], SEGAN+ [59], Knowledge Distillation GAN [60]
		2020	HLGAN [61], CP-GAN [29], DSEGAN [62], HiFi-GAN [48], Forked GAN with mask-learning [47]
		2021	Progressive *G* multi-scale *D* GAN [63], Lightweight GAN [64], Sinc–SEGAN [65], SASEGAN [30]
		2022	PGGAN [66], SEGWGAN-HP [67]
		2023	DSEGAN-Self-Attention [68], MAMGAN [31], SEFGAN [69]
		2024	Convolutional multi-timescale GAN [70], SASEGAN-TCN [71]
		2025	CA-Res-SEGAN [72]
Magnitude	Magnitude mapping	2017	NG-Pix2Pix [4]
		2018	FSEGAN [33], LSTM+GAN [73], ACSE [20]
		2019	S-ForkGAN [74], MetricGAN [25]
		2020	iMetricGAN [75], MOCG [76], Multi-Resolution GAN [77], Transformer Metric-GAN [78]
		2021	Multi-Metric GAN [36], AeGAN [32], AIA-CycleGAN [79]
		2022	VSEGAN [80], CycleGAN-DAL [81], PSMGAN [38]
		2025	MOS-GAN [82], TFDense-GAN [83]
	Magnitude mask estimation	2020	PAGAN [84], M-CRGAN-MSE [37]
		2021	MetricGAN+ [85], μ-law SGAN [86]
		2022	MetricGAN-U [87]
		2023	MetricGAN-OKD [88]
		2025	MetricGAN+KAN [89]
	Implicit mask estimation	2025	MMSE-GAN [34], CNN-GAN [35]
Complex STFT (real + imag)	Complex-domain mapping	2021	CDGAN [39], DPTGAN [90]
		2024	Multi-CMGAN+/+ [49], N2N2N [91]
	Complex mask estimation	2022	DCCRGAN [50], SkipConvGAN [51]
	Magnitude mask estimation	2025	Deformable convolution GAN [52]
Magnitude + Complex STFT	Magnitude mask estimation and complex residuals prediction	2021	CycleGAN-DCD [92]
		2022	CMGAN [40], CinCGAN [93]
		2023	TPTGAN [41], CGA-MGAN [94], HA-MGAN [95], PAMGAN+/- [96], Convolutional recurrent MetricGAN [97]
		2024	UPB-CMGAN [98], CMGAN [99], TSMGAN-II [100], MSCTGAN [101], Revised CMGAN [102]
		2025	CRG-MGAN [103], MRGAN [104], CorrGAN [105]
Magnitude + Phase	Direct amplitude mapping and direct phase mapping	2024	MambaGAN+PCS [106], PHASEGAN [107]
		2025	PASEGAN [108], Phase-aware MetricGAN [109]
Mixed T-domain and TF-domain	Combined T-domain and TF-domain prediction	2021	HiFi-GAN-2 [42]
		2023	SCP-CMGAN [44], HiFi++ [21], GAN-in-GAN [45]
		2024	M-DGAN [46]
		2025	DisCoGAN [43]

1) ADVERSARIAL LOSS USAGE

Adversarial loss is the core element that ties *G* and *D* together as they engage in the minimax game. The mathematical formulations of base versions of the losses that have been used in shortlisted works has been presented in Section IV-A. Here, the main focus is on how their usage evolved over the years. Note that the loss functions were grouped for this analysis, for example, both (1) and (2), and other modifications of binary cross entropy are represented in one group.

Table 2 contains a list of models that were assigned to different adversarial loss functions, split by year. The same data is visualized in Fig. 4, which shows what types of losses were used throughout the years.

As can be seen from Fig. 4, the metric-type loss has gained popularity since it has been proposed and remains one of the most used losses. However, it appears that in the last year, researchers have started to explore new ideas. The authors of the paper [105] proposed a new loss, which uses the metric scores, however, the discriminator does not need to mimic

TABLE 2. Adversarial loss function groups split by year. Note, [37] and [45] were not included in this list as multiple loss functions were used in these works, [60], [72], and [108] were not included as the used loss was not identified.

Adversarial loss	Year	Models
Hinge	2020	HiFi-GAN [48]
	2021	HiFi-GAN-2 [42]
	2025	DisCoGAN [43]
Least squares	2017	SEGAN [3]
	2018	Adapted SEGAN [53], LSTM+GAN [73]
	2019	GSEGAN [56], S-ForkGAN [74], SEGAN+ [59], Generalized SEGAN [58]
	2020	Forked GAN with mask-learning [47], DSEGAN [62], MOCG [76], HLGAN [61]
	2021	Lightweight GAN [64], Sinc–SEGAN [65], SASEGAN [30]
	2022	SkipConvGAN [51], CycleGAN-DAL [81], PSMGAN [38], VSEGAN [80]
	2023	HiFi++ [21], DSEGAN-Self-Attention [68], SEFGAN [69]
	2024	SASEGAN-TCN [71], Convolutional multi-timescale GAN [70]
Metric	2019	MetricGAN [25]
	2020	Transformer MetricGAN [78], iMetricGAN [75]
	2021	Multi-Metric GAN [36], MetricGAN+ [85]
	2022	CMGAN [40], MetricGAN-U [87], PGGAN [66]
	2023	CGA-MGAN [94], HA-MGAN [95], PAMGAN+/- [96], MAMGAN [31], Convolutional recurrent MetricGAN [97], TPT-GAN [41], MetricGAN-OKD [88], SCP-CMGAN [44]
	2024	CMGAN [99], PHASEGAN [107], Multi-CMGAN+/+ [49], Revised CMGAN [102], TSMGAN-II [100], M-DGAN [46], MambaGAN+PCS [106], MSCTGAN [101], UPB-CMGAN [98]
	2025	TFDense-GAN [83], Deformable convolution GAN [52], MOS-GAN [82], MetricGAN+KAN [89], MRGAN [104], Phase-aware MetricGAN [109], CRG-MGAN [103]
Other	2025	CorrGAN [105]
Relativistic	2019	SERGAN [55]
	2020	PAGAN [84], CP-GAN [29], Multi-Resolution GAN [77]
	2021	Progressive *G* multi-scale *D* GAN [63], CycleGAN-DCD [92], AIA-CycleGAN [79]
	2022	DCCRGAN [50], CinCGAN [93], DPTGAN [90]
Binary cross entropy	2017	NG-Pix2Pix [4]
	2018	FSEGAN [33], ACSE [20], CNN-GAN [35], MMSE-GAN [34]
	2019	UNetGAN [57]
	2021	AeGAN [32], μ-law SGAN [86], CDGAN [39]
	2024	N2N2N [91]
Wasserstein	2018	WCGAN-GP [28]
	2019	WGAN [54]
	2022	SEGWGAN-HP [67]

the metric but instead has a negative correlation with the metric.

Note that not all works explicitly indicate the loss function they have used. In such cases, if the work somehow indicated a relation to previous works, then the group was assigned based on the assumption that the loss is the same, however, it was not possible for 3 [60], [72], [108]. Additionally, two more works were not included, [37], where multiple losses of different groups were investigated, and [45], where inner and outer GAN had losses from different groups.

2) RECONSTRUCTION AND ADDITIONAL LOSSES

Using adversarial loss alone is usually not sufficient to achieve satisfactory results. To overcome this issue, it is paired with a reconstruction loss, which measures the distance, usually L_1 or L_2, between noisy and enhanced samples. In [28], both were used together as an elastic network, in [48], they were used to form loss on different domain representations. Another loss that was used in [48] and other papers is the feature-matching loss computed on the intermediate layers of the discriminator. The additional losses also depend on the architecture. If a cycle-based approach is used, usually cycle consistency and identity mapping losses are included in the overall loss. As the input trends changed and complex components gained popularity, corresponding loss functions have been incorporated into the overall loss, usually referred to as magnitude loss and real-imaginary (RI) loss or TF loss [92]. For example, MambaGAN [106] follows this trend. Although the model predicts magnitude and phase, it also includes a complex-domain RI loss by converting these predictions into real and imaginary components, along with the other losses used in the model. Other losses are also experimented on, some examples include [58], where authors used acoustic-based loss, margin-based, and spectral

subtraction loss was used in [74] and [95] included PMSQE (perceptual metric for speech quality evaluation) and PASE (problem-agnostic speech encoder) based losses, [98] used unrestricted phase bias-aware loss, NOMAD (non-matching audio distance) loss was tested in [102].

C. MODEL ARCHITECTURES

This section examines the architectures of GAN-based SE models. Specifically, we distinguish between the GAN framework, which defines how the generators and discriminators are organized, and the neural architectures of the generator and discriminator.

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 *(noisy, enhanced)* for real and generated data, respectively. Some studies use cGAN terminology because the generator is conditioned on auxiliary information [34], [35]. In this survey, we adopt the original formulation of conditional GANs by [2], 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 group of analyzed papers 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.

GAN-based SE studies grouped by framework, including standard GAN, cGAN, and CycleGAN, are listed in Table 3.

As shown in Table 3, cGANs are the most widely adopted GAN architecture for SE, accounting for 73 out of 87 applications. 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 (see [87], [89]), the generator-discriminator structure remains unchanged. Therefore, these methods remain standard conditional GAN-based SE models. Table 3 highlights metric-driven variants in bold to indicate their classification as specialized cGAN models.

Among the 87 analyzed studies, 6 are Vanilla GAN models. Early approaches directly condition the generator on the noisy signal to learn a deterministic enhancement mapping [34], [35], [73]. 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 [21], [48]. 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 [110], has been adapted by [20] for SE to address the scarcity of parallel noisy-clean data. 8 studies published between 2018 and 2022 adopt the CycleGAN architecture. These models have been used to learn mappings between noisy and clean speech without requiring parallel recordings [20], [76], [92]. These models have also shown competitive results when parallel data was available. Studies such as those by [77] and [79] have employed the CycleGAN framework under paired training conditions only, demonstrating its ability to preserve speech structure and minimize distortion.

Numerous architectural improvements have been proposed for each GAN framework. The following subsections examine these improvements.

2) GENERATOR ARCHITECTURES

The GAN model was first introduced to the SE task with SEGAN [3]. 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. The taxonomy of generator architectures is given in Fig. 5.

Later, *baseline encoder-decoder models* improved upon the fundamental feed-forward architecture. These improvements included multi-scale convolutions [57], [67], residual blocks within encoder-decoder stages [67], [74], and gated linear units [81]. 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 [37] 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 [50], [95], [97]. The latest model in this category, MOS-GAN [82], uses a dual-path block between the encoder and the decoder, with two

TABLE 3. GAN frameworks in SE, bold font indicates metric-driven variants.

Framework	Year	Models
Vanilla Gan	2018	CNN-GAN [35], MMSE-GAN [34], LSTM+GAN [73]
	2020	HiFi-GAN [48]
	2021	μ-law SGAN [86]
	2023	HiFi++ [21]
CycleGAN	2018	ACSE [20]
	2020	Multi-Resolution GAN [77], MOCG [76]
	2021	CDGAN [39], CycleGAN-DCD [92], AIA-CycleGAN [79]
	2022	CycleGAN-DAL [81], CinCGAN [93]
Conditional GAN	2017	NG-Pix2Pix [4], SEGAN [3]
	2018	FSEGAN [33], Adapted SEGAN [53], WCGAN-GP [28]
	2019	WGAN [54], SERGAN [55], GSEGAN [56], **MetricGAN** [25], UNetGAN [57], S-ForkGAN [74], SEGAN+ [59], Generalized SEGAN [58], Knowledge Distillation GAN [60]
	2020	**Transformer MetricGAN** [78], **iMetricGAN** [75], PAGAN [84], Forked GAN with mask-learning [47], CP-GAN [29], DSEGAN [62], HLGAN [61], M-CRGAN-MSE [37]
	2021	AeGAN [32], Progressive *G* multi-scale *D* GAN [63], **Multi-Metric GAN** [36], Lightweight GAN [64], Sinc–SEGAN [65], SASEGAN [30], HiFi-GAN-2 [42], **MetricGAN+** [85]
	2022	**CMGAN** [40], **MetricGAN-U** [87], DCCRGAN [50], SkipConvGAN [51], **PGGAN** [66], PSMGAN [38], SEGWGAN-HP [67], VSEGAN [80], DPTGAN [90]
	2023	DSEGAN-Self-Attention [68], **CGA-MGAN** [94], **HA-MGAN** [95], **PAMGAN+/-** [96], **GAN-in-GAN** [45], **MAMGAN** [31], **Convolutional recurrent** MetricGAN [97], TPTGAN [41], **MetricGAN-OKD** [88], SEFGAN [69], **SCP-CMGAN** [44]
	2024	**CMGAN** [99], PHASEGAN [107], **Multi-CMGAN+/+** [49], N2N2N [91], **Revised CMGAN** [102], **TSMGAN-II** [100], **M-DGAN** [46], **MambaGAN+PCS** [106], SASEGAN-TCN [71], **MSCTGAN** [101], Convolutional multi-timescale GAN [70], **UPB-CMGAN** [98]
	2025	**TFDense-GAN** [83], **Deformable convolution GAN** [52], PASEGAN [108], CA-Res-SEGAN [72], **MOS-GAN** [82], **MetricGAN+KAN** [89], **MRGAN** [104], DisCoGAN [43], **Phase-aware MetricGAN** [109], **CRG-MGAN** [103], **CorrGAN** [105]

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 [77], [91], on the encoder-decoder side [70], [72], or distributed across layers [79], [92]. Also, each paper introduces attention for a different purpose. Some focus on using attention to capture global time-frequency dependencies during non-parallel training [79], [92]. Others use it to restore the formant structure distorted by reverberation [51]. Additionally, some approaches use attention as a lightweight mechanism for reweighting features [61], [64].

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 surveyed models, the work [78] was the first to 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 [52], [83], [107]. The CRG-MGAN model [103] is described as an improved transformer structure that integrates convolutional layers, recurrent networks, and spatial convolutional gating units to capture temporal-frequency 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 [41], [52], [94].

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 [46] or multi-scale [101]. Others retain the original CMGAN generator and focus exclusively on loss design or training target [40], [99]. MRGAN [104] is a lightweight hybrid design that combines time-domain Conformer blocks with frequency-domain Transformer blocks.

One *sequence modeling model* [106] 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

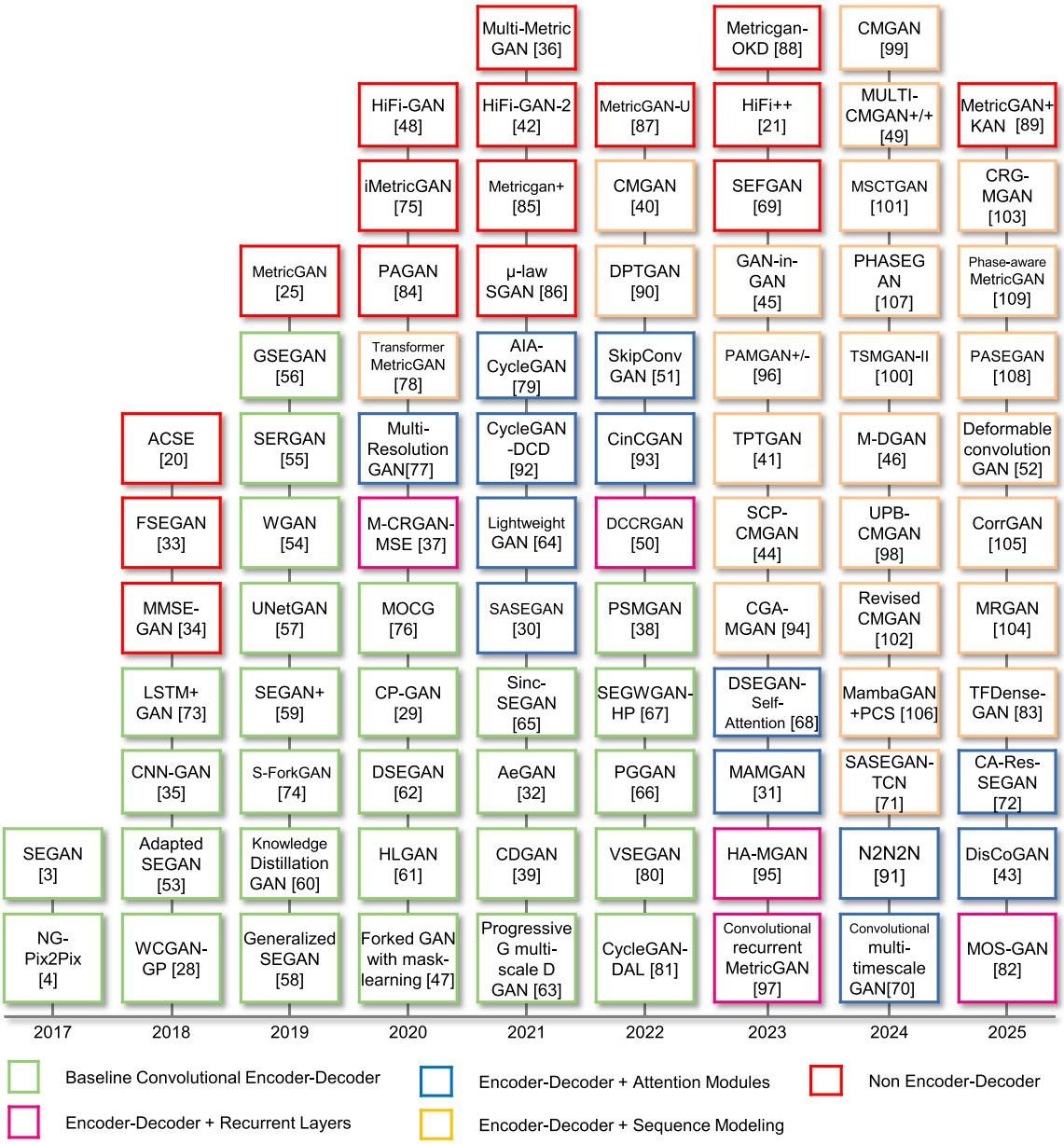

FIGURE 5. Taxonomy of GAN-based SE generator architectures.

MambaGAN [106], linearity is achieved through the incorporation of Dual-Path MambaFormer (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 [36], [87], [88], [89]. Other papers that do not follow an encoder-decoder architecture use synthesis-based generators that directly generate waveforms [21], [42], [48], [69]. Despite their descriptions using encoder-decoder terminology, some earlier works are treated as non-encoder-decoder models in this survey because their generators implement direct mappings rather than explicit encoder-decoder structures [33], [34], [84].

3) 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. Authors

TABLE 4. Dual-stream encoder-decoder generators.

Model	Year	Branches
S-ForkGAN [74]	2019	Speech decoder and noise decoder
Forked GAN with mask-learning [47]	2020	Speech decoder and noise decoder
MRGAN [104]	2024	Magnitude mask decoder and complex decoder
Phase-aware MetricGAN [109]	2025	Magnitude decoder and phase decoder
CRG-MGAN [103]	2025	Magnitude mask decoder and complex decoder

choose one term or the other based on whether they want to emphasize branching (forked), or parallel processing streams (dual-stream).

In dual-stream generators, the decoding stage is split into multiple parallel branches, each operating on the exact encoded representations. Models that utilize dual-stream generator outputs are given in Table 4. In model [109], the magnitude branch directly predicts the enhanced magnitude spectrum, whereas [103] and [104] 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.

4) 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 [83], or through a combination of global and local discriminators, which operate on full utterances and short segments, respectively [71].

The paper [52] 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.

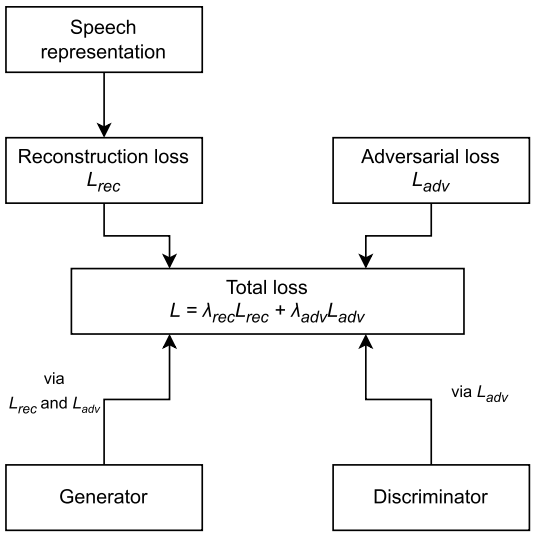

FIGURE 6. The relationships between speech signal representation, loss formulation, and GAN architecture in a standard SE system.

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 [62], [71].

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 [3]. 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 [41], [99].

D. MODEL DESIGN AND TRAINING

The initial step in the model design is typically the selection of a speech signal representation. The chosen signal representation does not determine the architecture of the generator or discriminator. Many models use the same architecture across different domains. After the architecture is fixed, appropriate loss functions are applied. The reconstruction loss depends directly on the speech signal representation, whereas the adversarial loss depends primarily on the GAN type and discriminator design. Both losses jointly contribute to the total training objective. The generator is trained with both reconstruction and adversarial losses, whereas the discriminator is optimized only with respect to the adversarial loss. The graphical representation of the relation between components is given in Fig. 6

Although adversarial objectives, such as those in LSGAN and WGAN, and their associated regularization techniques are essential for stability, they do not affect the underlying network design. In contrast, metric adversarial objectives modify the discriminator objective by replacing real-fake classification with metric prediction.

CycleGAN-based models differ from the standard dependency structure between representation, architecture, and loss given in Fig. 6. They differ in that they introduce cycle-consistency. This requires two generators and two discriminators. In this case, the loss formulation determines the required number of generators and discriminators.

1) ADVERSARIAL CONFIGURATIONS

This section addresses the adversarial configuration between the generator and discriminator. These configurations do not define GAN frameworks themselves. Rather, they describe the connection between the generator and discriminator within a GAN framework.

In SE, the term *multi-stage* refers to refining the signal step-by-step. The authors of [92] and [93] propose a two-stage denoising system that combines magnitude estimation and phase recovery. In [68], the authors constructed two models, ISEGAN and DSEGAN. ISEGAN performs iterative generation across multiple stages that share a mapping. In contrast, DSEGAN performs deep generation across multiple disconnected stages that learn separate mappings between noisy and clean speech. The multi-stage models use different generators at each stage, and each stage operates on the output of the previous stage.

Some GAN models have one generator and multiple discriminators. These models are called *Multi-discriminator models*. All discriminators are trained simultaneously. However, they differ in what each discriminator is responsible for. In [47], two discriminators are used to distinguish fake and real speech and noise, respectively. In [29], the global discriminator takes the generated audio and the original audio as input, and their sampling segments are then fed into the local discriminator. These two discriminators work together to assess the quality of the enhanced speech hierarchically. References [42], [48], and [63] use a set of discriminators on the waveform sampled at different rates. Based on this idea, [21] use the multi-discriminator adversarial training framework; however, their discriminators operate at the same resolution and have a smaller number of weights. The paper [69] uses an ensemble of discriminators working on different time scales and periods. The paper [43] employs a multi-scale STFT-based discriminator, consisting of multiple discriminators that operate at different spectral resolutions, whereas the paper [82] examined the effectiveness of MOS metrics. The authors of the aforementioned papers report that using multiple discriminators improves training stability, perceptual speech quality, and robustness to diverse noise conditions compared to single-discriminator setups.

The nested adversarial formulation, which differs from multi-stage and multi-discriminator GANs, was proposed by [45]. The authors propose a GAN-in-GAN framework *GAN-in-GAN framework*, in which two GANs are nested and optimized simultaneously. One GAN operates in the spectrogram domain, while the other operates in the waveform domain.

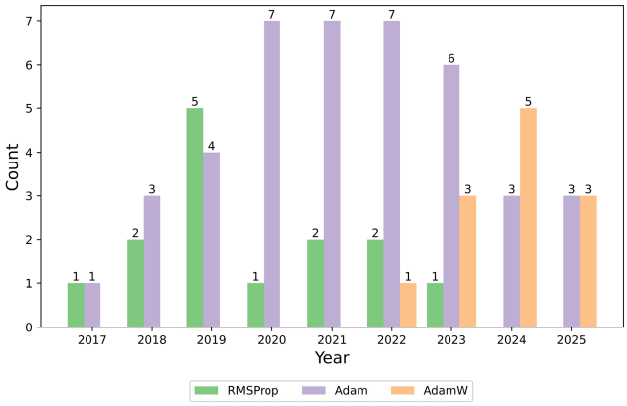

FIGURE 7. Bar plot representing the usage of different optimizers by years.

2) TRAINING AND OPTIMIZERS

GANs are known for their unstable and challenging training, making it difficult to choose the correct parameters.

Researchers use various approaches to tackle this issue. Some authors employ different learning rates for the generator and discriminator [54] or adjust the rates during learning by gradually changing it at some point of learning [79], [101], using warm-up [41], [86] or including schedulers [31], [104]. Others chose to address this issue via the update frequencies of the generator and the discriminator. It can go both ways, for example, in [4], G is updated twice for one D update, whereas in [38] and [48], D is updated twice. Another approach is to adjust the loss during training. Some examples of this include [58], where the model was initially trained adversarially and then acoustic losses were added. In [77] and [93], the identity mapping loss is used only for part of the training. Depending on the case, training can also be performed in stages, as in [48], [51], and [92], where some parts of the proposed framework are pretrained and then integrated with other parts. Lastly, some models used data from previous epochs to maintain the model's memory of what had already been learned [85], [89], [96].

Another one of the choices the researchers face is the optimizer. Among all selected papers that stated the optimizer they used, the 3 most popular are RMSProp, Adam, and AdamW. Their distribution over the years is presented in Fig. 7.

As can be seen from Fig. 7, the most popular optimizer has been Adam. RMSProp was used since the beginning of SE GANs, however, AdamW is now applied instead.

Of the 87 shortlisted papers, 18 did not indicate which optimizer they used. There were also other cases, for example, in [70], both RMSProp and Adam were used; the former was used to optimize the discriminator, and the latter was used to optimize the generator. The stochastic gradient descent algorithm with mean square error was used in [76].

Learning rates, epochs, and batch sizes heavily depend on the particular case, thus, these parameters are not compared in this study.

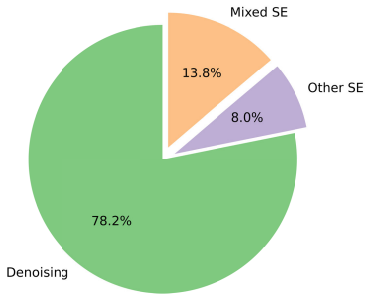

FIGURE 8. Pie chart representing the tasks that were researched in selected papers.

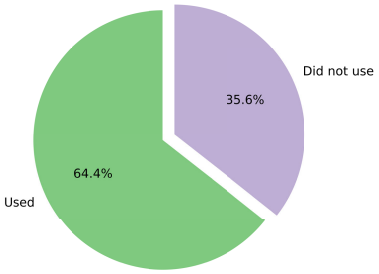

FIGURE 9. Pie chart representing the percentage of papers where *Noisy speech database for training speech enhancement algorithms and TTS models* [111] dataset was used.

VI. COMPARATIVE ANALYSIS

This section contains a comparison of the shortlisted papers. The factors taken into consideration are the tasks that were the primary focus of the article, the selection of datasets, the metrics used for evaluation of obtained results, and their reported values.

A. TASKS

Let $x(t)$ denote the clean speech signal, $n(t)$ the noise signal and $h(t)$ room impulse response filter. Signal $y(t)$ with both noise and reverberation is modeled as

$$y(t) = x(t) * h(t) + n(t), \tag{10}$$

here $*$ denotes the convolution operation. Either $h(t)$ or $n(t)$ may be omitted in case only one type of distortion is needed. Both denoising and dereverberation are subtasks of a broader SE task, which also includes other types of corruption removal.

Fig. 8 represents the tasks that the selected papers focused on. As illustrated in Fig. 8, the majority of shortlisted papers, 68, focused only on denoising task. This also aligns with the summary of datasets, which is presented in Section VI-B - the dataset that is used most often, [111], was made for denoising task.

The remaining articles were split into 2 groups and represented in Fig. 8 as other SE and mixed SE. Other SE means that the main focus of the study was another SE task. There were 7 such papers. 3 of them, namely [51], [52], [73] researched dereverberation. Note, that noise at 20 dB SNR was added in [52], however, the main focus was on dereverberation, thus, it was assigned to other SE. Other types of distortions, such as missing chunks, whispered speech, clipped amplitudes, and reduced bandwidth were investigated in [58]. Papers [36], [75] focused on near-end listening/speech intelligibility enhancement. SE for bone-conducted speech was performed in [81].

The remaining 12 papers included more than one type of speech signal corruption. Usually, an additional distortion was added to the background noise. The most common combination of tasks was denoising with dereverberation. This pair was treated as one joint task in [33]. Similarly, authors of [42] and [48] also performed SE on a dataset that had both of these distortions in order to obtain studio-quality outputs, however, they separately tested the models on denoising only. Tests on separate datasets with and without reverberation were conducted in [78]. Denoising and dereverberation were investigated in [87] as well, however, two separate experiments were dedicated for each type of corruption. The authors of [49] tested on noisy and reverberant sets, however, degradation was observed with the latter one. Reverberation was included within the training set used in [43], thus, this work was classified as mixed. However, based on the provided description, the evaluation only contained denoising task. Two papers added one more task to their work on denoising and dereverberation. The authors of [39] included separation within their proposed approach for enhancement, which addressed reverberation and noise. Denoising, dereverberation, and super-resolution problems were investigated in [99] and different datasets were used for three separate experiments. The remaining 3 publications focused on other types of distortions. Paper [59] included denoising and dewhispering as separate problems. Two experiments were performed in [88], one for denoising and one for near-end SE (listening enhancement). Lastly, the authors of [21] experimented with denoising and super-resolution tasks separately.

B. DATASETS

The selection and preprocessing of the dataset depends on the specific task being solved, as they must reflect the corresponding distortions. There are two choices for researchers: they can either use a prepared dataset or construct one of their own. This consequently leads to a wide variety of different datasets and their combinations. This section focuses on the most popular pre-made datasets and other commonly used sets for SE tasks.

Noisy speech database for training speech enhancement algorithms and TTS models [111] is the most widely used pre-made dataset for denoising. Fig. 9 represents how many of the selected papers have used it either as the main dataset or for supplementary experiments. As can be seen, the majority, 56 of 87, papers included *Noisy speech database for training speech enhancement algorithms and TTS models* in their research.

The latest version of this set was formed by using clean speech recordings from CSTR VCTK Corpus (Centre for Speech Technology Voice Cloning Toolkit) [112] and noise from DEMAND [113]. There are 2 training sets included. One contains 28 speakers, 14 females and 14 males, from the same accent region - England. The other contains 56 speakers, 28 females and 28 males, from different accent regions - the United States and Scotland. The 28-speaker dataset is more commonly used. Each speaker has around 400 recordings. The signals are sampled at 48 kHz. 10 types of noise were added to the recordings, 2 of them were generated, and the others were taken from a noise database. Each noise was added with SNR values of 0 dB, 5 dB, 10 dB, and 15 dB. The testing set was formed similarly, except only 4 real-world noises from DEMAND were used, and all SNR values were increased by 2.5 dB. The testing set contains only 2 speakers, a female and a male, both from England.

If the authors choose to make a dataset of their own, repeating picks for clean speech are *DARPA TIMIT acoustic-phonetic continuous speech corpus* [114], CSTR VCTK Corpus, and Librispeech [115]. All of them contain English read speech.

It is important to note that there are more datasets that are used, however, they did not occur in the selected papers as frequently. There have also been research performed with other languages, for example, a Mandarin corpus was used in [73], transfer learning with Catalan and Korean was investigated in [53], Korean and Chinese were used in [61], German and Spanish were included in [75], Japanese in [38], Tibetan in [67].

Common datasets for noise samples are NOISEX-92 [116], DEMAND [113] and *100 nonspeech environmental sounds*. All sets contain various environmental noise, overall it ranges from everyday sounds from surrounding areas like parks or home spaces from DEMAND, to sounds that are usually rare in real life, such as military vehicles or machine guns from NOISEX-92.

Datasets from challenges are also commonly used, for example, the Deep Noise Suppression Challenge dataset [117], and the Reverb Challenge corpus [118].

Among all papers that explicitly report their sampling rates, the data are usually (re)sampled to 16 kHz. There are a few exceptions with 8 kHz [39], [67], [76], 44.1 kHz [75] and 48 kHz [38], [97].

C. EVALUATION METRICS

The selection of objective evaluation metrics depends on the task that is being solved. In this case, most papers focus on denoising, so the metrics are repeated throughout. They help to compare the model's performance when the same dataset is used. Below are descriptions of the most common metrics used in SE.

- Perceptual Evaluation of Speech Quality (**PESQ**) [119] is a measure that allows to evaluate the quality of speech by comparing a clean and a degraded signal. This measure is recognized as an International Telecommunication Union (ITU-T) recommendation. The values range from -0.5 to 4.5. A higher value indicates better audio quality.
- Short Time Objective Intelligibility measure [120] (**STOI**) ranges from 0 to 1. Unlike PESQ, which focuses on quality, STOI measures intelligibility. Similar to PESQ, higher values are derisible.
- Composite measures are often reported together with PESQ and STOI, they are introduced in [121]. The 3 metrics are: composite measure for signal distortion (**CSIG**), composite measure for noise distortion (**CBAK**), and composite measure for overall speech quality (**COVL**). All 3 metrics range from 1 to 5, with higher values indicating a higher-quality signal.

Another important case to note is task-specific reporting. Some examples include equal error rate [4], [29], word error rate [29], [32], [33], [43], [69], character error rate [73], phone error rate [74], which are related to automatic speech or speaker recognition and verification.

Lastly, a portion of the selected papers included subjective listener tests, which help evaluate the quality of the proposed models in real-world settings. Examples include SEGAN [3], MetricGAN [25], HiFi-GAN and more.

D. SUMMARY OF REPORTED RESULTS

This section contains the summary of the results that have been provided in the selected papers. Some authors provided their results in different formats, such as graphs and, thus, were not included in the following tables. In order to keep the results concise, only one model or version of a model was chosen per paper, however, it is important to note that in many cases, the results were provided in a more extensive manner, including ablation studies or multiple versions of the model.

Table 5 represents the best results obtained when *Noisy speech database for training speech enhancement algorithms and TTS models* [111] is used for evaluation. In this case, the most commonly reported metrics, namely PESQ, STOI, and composites CSIG, CBAK, and COVL, were selected.

Out of 56 papers that experimented with this dataset, 54 were included in Table 5. Remaining 2 papers were excluded due the format of their reporting, [53] reported their results mostly in graphs, [99] continued the research on CMGAN with ablation study.

Gradual improvement of the performance of the models can be observed in Table 5. PESQ of almost 3 was already achieved in 2020 with HiFi-GAN [48] and M-CRGAN-MSE [37] models. The latest models usually reach around 3.5. The highest PESQ of 3.72 from the papers that have been selected for this review was obtained with MambaGAN+PCS [106] model. Other metrics have not been reported by al the papers, however, similar tendency can be observed. STOI score is reported to be 0.96 by

the later models, with the highest value of 0.97 achieved with TFDense-GAN [83]. The same model reached one of the highest CSIG of 4.8, whereas MambaGAN+PCS [106] scored 4.82, however there are many other models that have this score above 4.7, including GAN-in-GAN [45], SCP-CMGAN [44], UPB-CMGAN [98], CorrGAN [105] and Phase-aware MetricGAN [109]. CBAK scores are generally lower than CSIG. Only 2 models reached scores that are at least 4, namely CorrGAN [105] with 4 and M-DGAN [46]. Best COVL scores were similar, more models, including the majority from 2024-2025, achieved results above 4. The best score of 4.4 was reached with MambaGAN+PCS [106].

Table 6 and Table 7 represent PESQ and STOI scores obtained with other datasets. Table 6 contains results of the denoising task, whereas Table 7 is dedicated to other tasks. Note, that in total there are 51 such paper that used other dataset that is not [111], however, only 35 of them have been included in the tables. Other papers have not been added due to the results being reported in a different format that does not align with other papers. For example, only WER was reported in [33], CER was reported in [73], authors of [53] reported results in graphs, only the differences between noisy and enhanced values of metrics have been reported in [43]. Paper [49] was excluded because other metrics (CSIG, CBAK, COVL, SI-SDR) were provided instead of PESQ and STOI. The results from these tables can not be compared among each other, as the datasets differ, and even if some do match, the conditions, such as SNR or exact noise types, may not match. The reports may lack information such as the initial scores before enhancement, as can be seen from Table 6. However, similarly as with the denoising task with [111] dataset, a general tendency of improvement can be observed.

Although the surveyed studies primarily report performance on benchmark datasets, comparative analysis enables a practical interpretation of model suitability for different real-world scenarios. It can be seen from Table 5 and Table 6 that models such as MambaGAN+PCS [106], TFDense-GAN [83], TSMGAN-II [100], CorrGAN [105] perform well with denoising task based on PESQ scores. SkipConvGAN [51] provided better results with dereverberation. Models such as HiFi-GAN [48], HiFi-GAN-2 [42] and HiFi++ [21] can be used for studio-quality SE.

VII. LIMITATIONS AND FUTURE DIRECTIONS

The application of GANs in SE has been successfully realized and improved over the years since the first attempts in 2017. However, there are still areas for further research and improvement. Some authors indicated the limitations or possible future directions in their papers themselves. Naturally, many noted further experimentation and improvements of the model to increase the target scores, however there are other recurring themes. They are presented in this section together with general observations from the analysis of the selected papers.

A. GENERALIZATION

One of the main limitations is generalization of the model. The model should be capable to perform multiple tasks of SE, including denoising, dereverberation, and removal of other types of distortions, [83]. It is important to note that the enhanced model output should also improve the results of subsequent tasks performed on this signal, for example, ASR [99]. Similar limitations can be observed in Section VI-A. The majority of selected papers focus on denoising, suggesting that there is still room for improvement in the use of GANs for other tasks, as well as in their combinations.

B. DATASETS

Another direction for future work concerns datasets. The authors of [89] indicated the need to experiment with more challenging datasets. In relation to the generalization problem, the dataset used for research should reflect complex conditions and cover multiple tasks. As observed in Section VI-B, the most common dataset [111] covers only 4 noise types under 4 conditions. There have been custom datasets, as discussed in Subsection VI-B and listed in Table 6, or datasets from challenges, they are not used as widely, so the need for a more challenging benchmark dataset is clear. The same conclusion could be applied to languages, even though some of the used datasets have contained non-English speech, English remains the most common. The last issue to note is the limited number of paired datasets containing both clean and noisy recordings. Although creating more datasets is always a possibility, some research has been conducted with unpaired data; for example, [20], [76], [79], and [93] employed cycleGAN, which was originally developed for image data translation when the datasets are unpaired. An unsupervised version of MetricGAN was proposed in [87]. Recently, another metric-based model, MOS-GAN [82], was introduced.

C. MODEL SIZE AND COMPUTATIONAL COST

GANs not only require large datasets but also typically have large numbers of parameters and complex architectures. These limitations include computational cost and model size, which may prevent their use in real-time. To address these limitations, some papers have proposed using dual-path Transformers [41], [94] and state-space models [106]. Nevertheless, with the increasing application of speech within technology, such as hearing aids and mobile communications, researchers continue to identify computational efficiency as an important direction for future research [52], [83], [95].

D. STATE-SPACE MODELS

The results of the comparative analysis indicate that state-space sequence models outperform Transformer- and Conformer-based models on *Noisy speech database for training speech enhancement algorithms and TTS models* [111]. These models offer nearly linear scalability with

TABLE 5. Table with reported results using dataset [111]. Notes: this table contains one model per paper, 2 papers were excluded, all metrics have been rounded to 2 decimals and STOI has been standardized into [1, 0] interval, "-" means that the metric was not reported in the paper. Best PESQ, CSIG, CBAK, COVL per year starting with 2020 are in bold.

Model	Year	PESQ	STOI	CSIG	CBAK	COVL
initial	-	1.97	0.92	3.35	2.44	2.63
SEGAN [3]	2017	2.16	-	3.48	2.94	2.80
CNN-GAN [35]	2018	2.34	0.93	3.55	2.95	2.92
MMSE-GAN [34]	2018	2.53	0.93	3.80	3.12	-
GSEGAN [56]	2019	2.28	-	-	-	-
MetricGAN [25]	2019	2.86	-	3.99	3.18	3.42
SERGAN [55]	2019	2.62	0.94	-	-	-
CP-GAN [29]	2020	2.64	0.94	3.93	**3.29**	3.28
DSEGAN [62]	2020	2.35	0.93	3.55	3.10	2.93
HiFi-GAN [48]	2020	**2.94**	-	4.07	3.07	3.49
HLGAN [61]	2020	2.48	-	3.65	3.19	3.05
M-CRGAN-MSE [37]	2020	2.92	0.94	4.16	3.24	**3.54**
Multi-Resolution GAN*[77]	2020	2.72	0.94	**4.20**	3.21	3.47
AIA-CycleGAN [79]	2021	2.74	0.94	3.96	3.25	3.29
CycleGAN-DCD [92]	2021	2.90	0.94	4.24	3.57	3.49
HiFi-GAN-2 [42]	2021	3.11	-	**4.37**	3.54	**3.74**
Lightweight GAN [64]	2021	2.50	0.94	3.55	3.64	3.13
Metricgan+ [85]	2021	**3.15**	-	4.14	3.16	3.64
Progressive G multi-scale D GAN*[63]	2021	2.71	0.94	3.97	3.26	3.33
SASEGAN [30]	2021	2.36	0.93	3.54	3.08	2.93
Sinc-SEGAN [65]	2021	2.86	0.95	3.87	**3.66**	3.15
μ-law SGAN [86]	2021	2.86	-	4.16	3.42	-
CinCGAN [93]	2022	2.86	0.94	4.18	3.38	3.42
CMGAN [40]	2022	**3.41**	0.96	**4.63**	**3.94**	**4.12**
DCCRGAN [50]	2022	2.83	0.95	3.94	3.50	3.35
DPTGAN [90]	2022	2.86	0.94	3.98	3.49	3.42
MetricGAN-U [87]	2022	2.45	-	3.47	2.63	2.91
PGGAN [66]	2022	2.81	0.94	3.99	3.59	3.36
PSMGAN [38]	2022	2.92	-	3.88	3.45	3.52
CGA-MGAN [94]	2023	3.47	0.96	4.56	3.86	4.06
DSEGAN-Self-Attention [68]	2023	2.71	0.93	3.58	3.15	3.11
GAN-in-GAN [45]	2023	3.49	0.96	**4.77**	3.92	4.19
HiFi++ [21]	2023	2.90	0.95	-	-	-
MAMGAN [31]	2023	3.30	0.95	4.53	3.64	3.95
Metricgan-OKD [88]	2023	3.24	-	4.23	3.07	3.73
PAMGAN+/- [96]	2023	3.04	0.93	4.16	2.93	3.61
SCP-CMGAN [44]	2023	**3.52**	0.96	4.75	**3.97**	**4.25**
TPTGAN [41]	2023	3.35	-	4.59	3.83	4.02
Convolutional multi-timescale GAN*[70]	2024	2.67	-	3.83	3.36	3.28
MambaGAN+PCS [106]	2024	**3.72**	0.96	**4.82**	3.65	**4.40**
M-DGAN [46]	2024	3.52	0.96	4.68	**4.05**	4.21
MSCTGAN [101]	2024	3.42	0.96	4.66	3.95	4.12
PHASEGAN [107]	2024	3.19	0.95	4.62	3.83	3.97
Revised CMGAN*[102]	2024	3.21	0.95	-	-	-
SASEGAN-TCN [71]	2024	2.16	0.93	3.41	2.83	2.76
TSMGAN-II [100]	2024	3.40	-	4.54	3.88	4.03
UPB-CMGAN [98]	2024	3.55	0.96	4.78	-	4.28
CorrGAN [105]	2025	3.51	0.96	4.75	**4.00**	4.27
CRG-MGAN [103]	2025	3.48	0.96	4.66	3.95	4.15
DisCoGAN [43]	2025	2.86	-	-	-	-
MetricGAN+KAN [89]	2025	3.30	-	4.02	3.04	3.63
MOS-GAN [82]	2025	2.40	-	-	-	-
PASEGAN [108]	2025	2.76	0.93	3.79	3.32	3.54
Phase-aware MetricGAN*[109]	2025	3.58	0.95	4.77	3.66	4.29
TFDense-GAN [83]	2025	**3.62**	0.97	**4.80**	3.86	**4.33**

* The acronym was not provided in original paper.

respect to sequence length. However, Mamba-based designs have only been adopted in one study for GAN-based SE [106]. This suggests that state-space models are a promising yet under-explored direction for future research.

E. METRIC-DRIVEN TRAINING

A quantitative analysis of recent literature shows that metric-driven training strategies have become increasingly popular. A growing number of state-of-the-art models now rely on directly optimizing metric scores [88], [89], [94]. Examples include the Perceptual Evaluation of Speech Quality (PESQ) and the Non-Intrusive Perceptual Objective Speech Quality Metric (DNSMOS). While using perceptual losses improves speech quality, these methods either rely on paired clean speech or fail to leverage the full effectiveness of non-intrusive perceptual objective speech quality metrics [82]. The primary research directions include developing models that work with only noisy recordings and design

TABLE 6. Table with reported results using custom datasets for denoising task. Notes: this table contains one model per paper, some papers were excluded due to non-averaged reporting, all metrics have been rounded to 2 decimal places, and STOI has been standardized to the [1, 0] interval. "-" means that the metric was not reported in the paper. Composite metrics have not been included due to lack of papers that used them.

Model	Year	Dataset		PESQ	STOI
Denoising					
NG-Pix2Pix†[4]	2017	RSR2015 + Librispeech + white Gaussian noise + real-life records + OCTAVE project	initial	2.00	0.72
			enhanced	2.24	0.72
WCGAN-GP [28]	2018	TIMIT + NOISEX-92	initial	-	-
			enhanced	1.70	-
Knowledge Distillation GAN†[60]	2019	CHiME4	initial	2.15	-
			enhanced	2.34	-
MetricGAN [25]	2019	TIMIT + 100 nonspeech environmental sounds	initial	1.63	0.70
			enhanced	2.13	0.76
SEGAN+ [59]	2019	VCTK Corpus + DEMAND	initial	1.92	0.74
			enhanced	2.37	0.75
UNetGAN [57]	2019	TIMIT + NOISEX-92	initial	1.32	0.58
			enhanced	2.14	0.80
WGAN [54]	2019	VoiceBank + BUTReverbDB	initial	2.65	0.97
			enhanced	2.81	0.99
Forked GAN with mask-learning*†[47]	2020	TIMIT + 100 nonspeech environmental sounds	initial	1.86	0.79
			enhanced	2.90	0.94
HLGAN†[61]	2020	VoiceBank + 100 nonspeech environmental sounds	initial	1.07	-
			enhanced	1.23	-
MOCG†[76]	2020	TIMIT + NOISEX-92 + 100 nonspeech environmental sounds	initial	2.23	0.72
			enhanced	2.62	0.78
PAGAN†[84]	2020	TIMIT + 100 nonspeech environmental sounds + real-life records	initial	2.23	0.83
			enhanced	3.24	0.92
Transformer MetricGAN*†[78]	2020	DNS	initial	2.45	0.92
			enhanced	3.10	0.95
AeGAN†[32]	2021	TIMIT + QUT-NOISE-TIMIT	initial	1.72	0.76
			enhanced	2.64	0.88
CycleGAN-DCD†[92]	2021	WSJ0-SI84 + DNS + NOISEX-92	initial	1.57	0.73
			enhanced	2.42	0.91
PSMGAN [38]	2022	TIMIT + NOISEX-92	initial	1.73	0.68
			enhanced	2.57	0.88
SEGWGAN-HP†[67]	2022	TIMIT + 100 nonspeech environmental sounds + NOISEX-92	initial	2.26	0.76
			enhanced	2.75	0.84
SEGWGAN-HP†[67]	2022	Tibetan Corpus + 100 nonspeech environmental sounds + NOISEX-92	initial	2.13	0.75
			enhanced	2.49	0.83
VSEGAN†[80]	2022	GRID + real-life records	initial	1.14	0.57
			enhanced	2.99	0.88
HA-MGAN [95]	2023	DNS + NHANES	initial	-	-
			enhanced	2.53	0.94
MAMGAN†[31]	2023	TIMIT + DNS + NOISEX-92	initial	2.19	0.83
			enhanced	3.15	0.93
SEFGAN [69]	2023	WSJ0-CHiME3	initial	-	-
			enhanced	2.94	-
N2N2N†[91]	2024	Librispeech + DEMAND + URBANSOUND 8k + AudioSET	initial	-	-
			enhanced	2.00	0.90
N2N2N†[91]	2024	TIMIT + DEMAND + URBANSOUND 8k + AudioSET	initial	-	-
			enhanced	1.98	0.90
N2N2N†[91]	2024	SUBAK.KO + DEMAND + URBANSOUND 8k + AudioSET	initial	-	-
			enhanced	1.98	0.90
SASEGAN-TCN [71]	2024	THCHS30 + NOISEX-92	initial	1.40	0.80
			enhanced	1.81	0.84
TSMGAN-II [100]	2024	TIMIT + NOISEX-92	initial	1.48	-
			enhanced	2.31	-
TSMGAN-II [100]	2024	VoiceBank + DEMAND	initial	1.63	-
			enhanced	3.01	-
TSMGAN-II [100]	2024	VoiceBank + NOISEX-92	initial	1.38	-
			enhanced	2.57	-
CA-Res-SEGAN [72]	2025	TIMIT + NOISEX-92	initial	1.97	0.91
			enhanced	2.68	0.95
MOS-GAN [82]	2025	CHiME4	initial	1.27	-
			enhanced	1.52	-
MRGAN [104]	2025	TIMIT + NOISEX-92	initial	-	-
			enhanced	3.39	0.95
TFDense-GAN [83]	2025	DNS	initial	1.97	0.92
			enhanced	3.66	0.98

* The acronym was not provided in original paper.
† Approximate results averaged from provided metrics.

TABLE 7. Table with reported results using custom datasets for non-denoising SE tasks. Notes: this table contains one model per paper, some papers were excluded due to non-averaged reporting, all metrics have been rounded to 2 decimal places, and STOI has been standardized to the [1, 0] interval. "-" means that the metric was not reported in the paper. Composite metrics have not been included due to lack of papers that used them.

Model	Year	Dataset		PESQ	STOI
Denoising + dereverberation					
HiFi-GAN [48]	2020	Sythetic dataset	initial	1.92	0.91
			enhanced	2.78	0.94
HiFi-GAN [48]	2020	DAPS clean set + MIT Impulse Response Survey + REVERB Challenge + ACE Challenge	initial	1.41	0.87
			enhanced	2.00	0.89
CDGAN†[39]	2021	TIMIT + NOISEX-92	initial	1.72	0.68
			enhanced	2.62	0.81
HiFi-GAN-2 [42]	2021	DAPS clean set + MIT Impulse Response Survey + REVERB Challenge + ACE Challenge	initial	1.41	0.87
			enhanced	2.23	0.92
Deformable convolution GAN*[52]	2025	REVERB	initial	1.50	-
			enhanced	3.02	-
Dereverberation					
MetricGAN-U [87]	2022	VoiceBank + OpenSLR	initial	1.98	-
			enhanced	2.07	-
SkipConvGAN†[51]	2022	REVERB	initial	1.50	-
			enhanced	2.91	-
Near-end SE					
Multi-Metric GAN*[36]	2021	Hurricane Challenge + MS-SNSD	initial	-	-
			enhanced	3.46	-
Multi-Metric GAN*†[36]	2021	Harvard sentences + MS-SNSD	initial	-	-
			enhanced	3.55	-
Bone-conducted SE					
CycleGAN-DAL [81]	2022	AEUCHSAC&BC-2017	initial	-	0.62
			enhanced	-	0.80
CycleGAN-DAL [81]	2022	TMHINT	initial	-	0.69
			enhanced	-	0.84

* The acronym was not provided in original paper.
† Approximate results averaged from provided metrics.

robust perceptually guided training strategies that provide stable optimization. Further experimentation with focus on metrics is also possible. Some authors, for example, those of paper [49], have identified the gap between the most common objective metrics such as PESQ and human subjective evaluation. If listening tests and DNSMOS are present in the paper, they are usually included in the evaluation of performance, thus, additional exploration and advancement is needed in order to include perceptual evaluation within the training step.

F. PHASE INFORMATION

Although waveform-based GANs have always modeled phase, recent SE methods have increasingly focused on modeling phase using the time-frequency domain. This approach goes beyond magnitude-only learning by incorporating noisy phase reuse. Phase modeling is primarily achieved through complex-domain learning strategies [40], [41], [105], or through combined T-domain and TF-domain prediction [43], [46]. Only a few recent methods perform phase prediction [107], [108], [109]. These studies have shown that the full potential of phase information has yet to be realized and have identified phase-related information as a subject for further investigation.

G. DISCRIMINATOR ARCHITECTURES

The majority of recent architectural innovations focus on increasing the capacity of generator-side modeling, while discriminator architectures often remain relatively shallow. Recent models introduce discriminator-side architectural refinements such as deformable convolution [52], multi-scale inputs [43], and auxiliary attention modules [62], [71]. Model comparative results presented in this survey indicate that these models achieve competitive performance, suggesting that refinement of the discriminator architecture is a promising research direction.

VIII. CONCLUSION

This paper overviews GANs and how they have been applied for SE task. Survey includes 87 papers from the very first attempts to use GANs for SE to the late 2025. The collected publications show that GANs retain their popularity over the years. The main goal of this review is to introduce GANs for SE and provide sufficient information to assist researchers with interested in employing GANs for their SE tasks.

Key factors are summarized and trends are presented. Firstly, the taxonomy of signal representations and training targets is created. Overall classification as well as the trends are provided, they reveal that the researchers have moved on from conventional representations, such as waveform or magnitude, to more advanced ones that jointly model magnitude and phase. Secondly, adversarial losses are investigated. The analysis shows that the MetricGAN loss, which was intended for speech domain, and its variations are still widely used. Thirdly, taxonomy of generator architectures is provided. The annual 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] 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.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[3] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Aug. 2017, pp. 3642–3646.

[4] 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.

[5] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. J. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends Hearing*, vol. 27, Jan. 2023, Art. no. 23312165231209913.

[6] S. T. Yousif and B. M. Mahmmod, "Speech enhancement algorithms: A systematic literature review," *Algorithms*, vol. 18, no. 5, p. 272, May 2025. [Online]. Available: https://www.mdpi.com/1999-4893/18/5/272

[7] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis," *Artif. Intell. Rev.*, vol. 56, no. S3, pp. 3651–3703, Dec. 2023.

[8] A. R. Yuliani, M. F. Amri, E. Suryawati, A. Ramdan, and H. F. Pardede, "Speech enhancement using deep learning methods: A review," *Jurnal Elektronika dan Telekomunikasi*, vol. 21, no. 1, pp. 19–26, Aug. 2021.

[9] C. Jannu and S. D. Vanambathina, "An overview of speech enhancement based on deep learning techniques," *Int. J. Image Graph.*, vol. 25, no. 1, Jan. 2025, Art. no. 2550001.

[10] D. O'Shaughnessy, "Speech enhancement—A review of modern methods," *IEEE Trans. Hum.-Mach. Syst.*, vol. 54, no. 1, pp. 110–120, 2024.

[11] A. Azarang and N. Kehtarnavaz, "A review of multi-objective deep learning speech denoising methods," *Speech Commun.*, vol. 122, pp. 1–10, Sep. 2020.

[12] J. Agrawal, M. Gupta, and H. Garg, "A review on speech separation in cocktail party environment: Challenges and approaches," *Multimedia Tools Appl.*, vol. 82, no. 20, pp. 31035–31067, Aug. 2023.

[13] A. dos Santos, P. de Oliveira, and B. Masiero, "A retrospective on multichannel speech and audio enhancement using machine and deep learning techniques," in *Proc. 24th Int. Congr. Acoust.*, 2022, pp. 173–184.

[14] S. Drgas, "A survey on low-latency DNN-based speech enhancement," *Sensors*, vol. 23, no. 3, p. 1380, Jan. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1380

[15] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, M. Usama, and J. Qadir, "Transformers in speech processing: A survey," 2023, *arXiv:2303.11607*.

[16] S. Gul and M. S. Khan, "A survey of audio enhancement algorithms for music, speech, bioacoustics, biomedical, industrial, and environmental sounds by image U-net," *IEEE Access*, vol. 11, pp. 144456–144483, 2023.

[17] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M. B. Ali, M. Adan, and M. Mujtaba, "Generative adversarial networks for speech processing: A review," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101308.

[18] D. Skariah and J. Thomas, "Review of speech enhancement methods using generative adversarial networks," in *Proc. Int. Conf. Control, Commun. Comput. (ICCC)*, 2023, pp. 1–4.

[19] N. Elgiriyewithana and N. D. Kodikara, "A comprehensive review on generative models for speech enhancement," in *Proc. 4th Int. Conf. Robot., Autom. Artif. Intell. (RAAI)*, Dec. 2024, pp. 236–252.

[20] Z. Meng, J. Li, Y. Gong, and B.-H.-F. Juang, "Cycle-consistent speech enhancement," in *Proc. Interspeech*, Sep. 2018, pp. 1165–1169.

[21] 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.

[22] 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.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Aug. 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5769–5779.

[25] 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

[26] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," 2018, *arXiv:1807.00734*.

[27] J. H. Lim and J. C. Ye, "Geometric GAN," 2017, *arXiv:1705.02894*.

[28] S. Qin and T. Jiang, "Improved Wasserstein conditional generative adversarial network speech enhancement," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 181, Dec. 2018.

[29] G. Liu, K. Gong, X. Liang, and Z. Chen, "CP-GAN: Context pyramid generative adversarial network for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6624–6628.

[30] H. Phan, H. L. Nguyen, O. Y. Chén, P. Koch, N. Q. K. Duong, I. McLoughlin, and A. Mertins, "Self-attention generative adversarial network for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7103–7107.

[31] H. Guo, H. Jian, Y. Wang, H. Wang, X. Zhao, W. Zhu, and Q. Cheng, "MAMGAN: Multiscale attention metric GAN for monaural speech enhancement in the time domain," *Appl. Acoust.*, vol. 209, Jun. 2023, Art. no. 109385.

[32] S. Abdulatif, K. Armanious, K. Guirguis, J. T. Sajeev, and B. Yang, "AeGAN: Time-frequency speech denoising via generative adversarial networks," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 451–455.

[33] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5024–5028.

[34] 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.

[35] 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.

[36] H. Li and J. Yamagishi, "Multi-metric optimization using generative adversarial networks for near-end speech intelligibility enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3000–3011, 2021.

[37] Z. Zhang, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, H. Song, and X. Li, "On loss functions and recurrency training for GAN-based speech enhancement systems," in *Proc. Interspeech*, Oct. 2020, pp. 3266–3270.

[38] S. Routray and Q. Mao, "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network," *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101270.

[39] Y. Li, W.-T. Zhang, and S.-T. Lou, "Generative adversarial networks for single channel separation of convolutive mixed speech signals," *Neurocomputing*, vol. 438, pp. 63–71, May 2021.

[40] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *Proc. Interspeech*, Sep. 2022, pp. 936–940.

[41] Z. Liu, Z. Jiang, W. Luo, Z. Fan, H. Di, Y. Long, and H. Wang, "TPTGAN: Two-path transformer-based generative adversarial network using joint magnitude masking and complex spectral mapping for speech enhancement," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2023, pp. 48–61.

[42] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2021, pp. 166–170.

[43] S. S. Shetu, E. A. P. Habets, and A. Brendel, "GAN-based speech enhancement for low SNR using latent feature conditioning," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[44] V. Zadorozhnyy, Q. Ye, and K. Koishida, "SCP-GAN: Self-correcting discriminator optimization for training consistency preserving metric GAN on speech enhancement tasks," in *Proc. INTERSPEECH*, Aug. 2023, pp. 2463–2467.

[45] Y. Duan, J. Ren, H. Yu, and X. Jiang, "GAN-in-GAN for monaural speech enhancement," *IEEE Signal Process. Lett.*, vol. 30, pp. 853–857, 2023.

[46] X. Lin, Y. Zhang, and S. Wang, "Mixed T-domain and TF-domain magnitude and phase representations for GAN-based speech enhancement," *Sci. Rep.*, vol. 14, no. 1, p. 17698, Jul. 2024.

[47] J. Lin, S. Niu, A. J. V. Wijngaarden, J. L. McClendon, M. C. Smith, and K.-C. Wang, "Improved speech enhancement using a time-domain GAN with mask learning," in *Proc. Interspeech*, Oct. 2020, pp. 3286–3290.

[48] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. Interspeech*, Oct. 2020, pp. 4506–4510.

[49] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "Multi-CMGAN+/+: Leveraging multi-objective speech quality metric prediction for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 351–355.

[50] H. Huang, R. Wu, J. Huang, J. Lin, and J. Yin, "DCCRGAN: Deep complex convolution recurrent generator adversarial network for speech enhancement," in *Proc. Int. Symp. Electr., Electron. Inf. Eng. (ISEEIE)*, Feb. 2022, pp. 30–35.

[51] V. Kothapally and J. H. L. Hansen, "SkipConvGAN: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1600–1613, 2022.

[52] 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.

[53] S. Pascual, M. Park, J. Serra, A. Bonafonte, and K.-H. Ahn, "Language and noise transfer in speech enhancement generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5019–5023.

[54] N. Adiga, Y. Pantazis, V. Tsiaras, and Y. Stylianou, "Speech enhancement for noise-robust speech synthesis using Wasserstein GAN," in *Proc. Interspeech*, Sep. 2019, pp. 1821–1825.

[55] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 106–110.

[56] C. Fan, B. Liu, J. Tao, J. Yi, Z. Wen, and Y. Bai, "Noise prior knowledge learning for speech enhancement via gated convolutional generative adversarial network," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 662–666.

[57] X. Hao, X. Su, Z. Wang, H. Zhang, and Batushiren, "UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," in *Proc. Interspeech*, 2019, pp. 1786–1790.

[58] S. Pascual, J. Serra, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *Proc. Interspeech*, Sep. 2019, pp. 1791–1795.

[59] S. Pascual, J. Serra, and A. Bonafonte, "Time-domain speech enhancement using generative adversarial networks," *Speech Commun.*, vol. 114, pp. 10–21, Nov. 2019.

[60] J. Wu, Y. Hua, S. Yang, H. Qin, and H. Qin, "Speech enhancement using generative adversarial network by distilling knowledge from statistical method," *Appl. Sci.*, vol. 9, no. 16, p. 3396, Aug. 2019.

[61] F. Yang, Z. Wang, J. Li, R. Xia, and Y. Yan, "Improving generative adversarial networks for speech enhancement through regularization of latent representations," *Speech Commun.*, vol. 118, pp. 1–9, Apr. 2020.

[62] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1700–1704, 2020.

[63] H. Y. Kim, J. W. Yoon, S. J. Cheon, W. H. Kang, and N. S. Kim, "A multi-resolution approach to GAN-based speech enhancement," *Appl. Sci.*, vol. 11, no. 2, p. 721, Jan. 2021.

[64] L. Li, Z. Lu, T. Watzel, L. Kurzinger, and G. Rigoll, "Light-weight self-attention augmented generative adversarial networks for speech enhancement," *Electronics*, vol. 10, no. 13, p. 1586, Jun. 2021.

[65] L. Li, Wudamu, L. Kurzinger, T. Watzel, and G. Rigoll, "Lightweight end-to-end speech enhancement generative adversarial network using sinc convolutions," *Appl. Sci.*, vol. 11, no. 16, p. 7564, Aug. 2021.

[66] Y. Li, M. Sun, and X. Zhang, "Perception-guided generative adversarial network for end-to-end speech enhancement," *Appl. Soft Comput.*, vol. 128, Oct. 2022, Art. no. 109446.

[67] X. Zhu and H. Huang, "Using hybrid penalty and gated linear units to improve Wasserstein generative adversarial networks for single-channel speech enhancement," *Comput. Model. Eng. Sci.*, vol. 135, no. 3, pp. 2155–2172, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1526149222002387

[68] B. K. Asiedu Asante, C. Broni-Bediako, and H. Imamura, "Exploring multi-stage GAN with self-attention for speech enhancement," *Appl. Sci.*, vol. 13, no. 16, p. 9217, Aug. 2023.

[69] M. Strauss, N. Pia, N. K. S. Rao, and B. Edler, "SEFGAN: Harvesting the power of normalizing flows and GANs for efficient high-quality speech enhancement," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2023, pp. 1–5.

[70] C. Yang and Z. Wu, "Speech enhancement method based on generative adversarial network and convolutional block attention module," in *Proc. 6th Int. Conf. Commun., Inf. Syst. Comput. Eng. (CISCE)*, May 2024, pp. 682–686.

[71] R. Lv, N. Chen, S. Cheng, G. Fan, L. Rao, X. Song, W. Lv, and D. Yang, "SASEGAN-TCN: Speech enhancement algorithm based on self-attention generative adversarial network and temporal convolutional network," *Math. Biosciences Eng.*, vol. 21, no. 3, pp. 3860–3875, 2024.

[72] J. Huang, "Improved SEGAN speech enhancement by fusing residual and convolutional attention mechanisms," in *Proc. 5th Int. Conf. Neural Netw., Inf. Commun. Eng. (NNICE)*, Jan. 2025, pp. 543–546.

[73] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1581–1585.

[74] J. Lin, S. Niu, Z. Wei, X. Lan, A. J. V. Wijngaarden, M. C. Smith, and K.-C. Wang, "Speech enhancement using forked generative adversarial networks with spectral subtraction," in *Proc. Interspeech*, Sep. 2019, pp. 3163–3167.

[75] H. Li, S.-W. Fu, Y. Tsao, and J. Yamagishi, "IMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning," in *Proc. Interspeech*, Oct. 2020, pp. 1336–1340.

[76] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1826–1838, 2020.

[77] Y. Wang, G. Yu, J. Wang, H. Wang, and Q. Zhang, "Improved relativistic cycle-consistent GAN with dilated residual network and multi-attention for speech enhancement," *IEEE Access*, vol. 8, pp. 183272–183285, 2020.

[78] S.-W. Fu, C.-F. Liao, T.-A. Hsieh, K.-H. Hung, S.-S. Wang, C. Yu, H.-C. Kuo, R. E. Zezario, Y.-J. Li, S.-Y. Chuang, Y.-J. Lu, Y.-C. Lin, and Y. Tsao, "Boosting objective scores of a speech enhancement model by MetricGAN post-processing," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 455–459.

[79] 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.

[80] X. Xu, Y. Wang, D. Xu, Y. Peng, C. Zhang, J. Jia, and B. Chen, "VSEGAN: Visual speech enhancement generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7308–7311.

[81] Q. Pan, Y. Pan, J. Zhou, H. Wang, L. Tao, and H. K. Kwan, "CycleGAN with dual adversarial loss for bone-conducted speech enhancement," in *Proc. IEEE Region 10 Conf.*, Nov. 2022, pp. 1–4.

[82] W. Jiang, F. Wen, and K. Yu, "MOS-GAN: Mean opinion score GAN for unsupervised speech enhancement," *IEEE Signal Process. Lett.*, vol. 32, pp. 3465–3469, 2025.

[83] H. Chen, J. Zhang, Y. Fu, X. Zhou, R. Wang, Y. Xu, and D. Ke, "TFDense-GAN: A generative adversarial network for single-channel speech enhancement," *EURASIP J. Adv. Signal Process.*, vol. 2025, no. 1, p. 10, Mar. 2025.

[84] P. Li, Z. Jiang, S. Yin, D. Song, P. Ouyang, L. Liu, and S. Wei, "PAGAN: A phase-adapted generative adversarial networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6234–6238.
[85] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, Aug. 2021, pp. 201–205.
[86] H. Li, Y. Xu, D. Ke, and K. Su, "μ-law SGAN for generating spectra with more details in speech enhancement," *Neural Netw.*, vol. 136, pp. 17–27, Apr. 2021.
[87] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7412–7416.
[88] W. Shin, B. H. Lee, J. S. Kim, H. J. Park, and S. W. Han, "MetricGAN-OKD: Multi-metric optimization of MetricGAN via online knowledge distillation for speech enhancement," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, Jul. 2023, pp. 31521–31538. [Online]. Available: https://proceedings.mlr.press/v202/shin23b.html
[89] Y. Mai and S. Goetze, "MetricGAN+KAN: Kolmogorov–Arnold networks in metric-driven speech enhancement systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
[90] D. Zhang, A. Dong, J. Yu, Y. Cao, C. Zhang, and Y. Zhou, "Speech enhancement generative adversarial network architecture with gated linear units and dual-path transformers," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2022, pp. 2563–2568.
[91] A. Deb, Asaduzzaman, R. Roy, A. Islam, C. Shahnaz, and M. Saquib, "N2N2N: A clean data independent speech enhancement approach with modified cGAN," in *Proc. IEEE Region 10 Conf. (TENCON)*, Dec. 2024, pp. 1474–1477.
[92] G. Yu, Y. Wang, H. Wang, Q. Zhang, and C. Zheng, "A two-stage complex network using cycle-consistent generative adversarial networks for speech enhancement," *Speech Commun.*, vol. 134, pp. 42–54, Nov. 2021.
[93] G. Yu, A. Li, Y. Wang, Y. Guo, H. Wang, and C. Zheng, "Joint magnitude estimation and phase recovery using cycle-in-cycle GAN for non-parallel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6967–6971.
[94] H. Chen and X. Zhang, "CGA-MGAN: Metric GAN based on convolution-augmented gated attention for speech enhancement," *Entropy*, vol. 25, no. 4, p. 628, Apr. 2023.
[95] J. Cheng, R. Liang, L. Zhao, C. Huang, and B. W. Schuller, "Speech denoising and compensation for hearing aids using an FTCRN-based metric GAN," *IEEE Signal Process. Lett.*, vol. 30, pp. 374–378, 2023.
[96] G. Close, T. Hain, and S. Goetze, "PAMGAN+/-: Improving phase-aware speech enhancement performance via expanded discriminator training," in *Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, 2023.
[97] Z. Hou, Q. Hu, T. Sun, Y. Hu, C. Zhu, and K. Chen, "Convolutional recurrent MetriCGAN with spectral dimension compression for full-band speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–2.
[98] S. Zhang, Z. Qiu, D. Takeuchi, N. Harada, and S. Makino, "Unrestricted global phase bias-aware single-channel speech enhancement with conformer-based metric GAN," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 1026–1030.
[99] S. Abdulatif, R. Cao, and B. Yang, "CMGAN: Conformer-based metric-GAN for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2477–2493, 2024.
[100] L. Lin, Y. Li, and H. Wang, "TSMGAN-II: Generative adversarial network based on two-stage mask transformer and information interaction for speech enhancement," in *Proc. Int. Conf. Intell. Comput.*, 2024, pp. 174–185.
[101] Z. Wu and N. Zheng, "MSCTGAN: Conformer improvements for speech enhancement," in *Proc. IEEE 17th Int. Conf. Signal Process. (ICSP)*, Oct. 2024, pp. 598–602.
[102] Y.-Z. Li, C.-W. Chang, A. N. Aung, and J.-W. Hung, "Integrating non-matching audio distance loss to the CMGAN speech enhancement," in *Proc. 10th Int. Conf. Appl. Syst. Innov. (ICASI)*, Apr. 2024, pp. 392–393.
[103] L. Yu, W. Zhang, F. Niu, and X. Li, "CRG-MGAN: A speech enhancement algorithm based on GAN," *AIP Adv.*, vol. 15, no. 9, pp. 1–18, Sep. 2025.
[104] C. Meng, G. Wei, Y. Long, C. Kong, and P. Ma, "MRGAN: LightWeight monaural speech enhancement using GAN network," in *Proc. Pattern Recognit. Comput. Vis.* Singapore: Springer, Jun. 2024, pp. 370–378.
[105] V. Zadorozhnyy, S. Amizadeh, Q. Ye, and K. Koishida, "CorrGAN: Simultaneous learning of speech enhancement and perceptual quality loss functions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
[106] T. Luo, F. Zhou, and Z. Bai, "MambaGAN: Mamba based metric GAN for monaural speech enhancement," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Aug. 2024, pp. 411–416.
[107] Y. Cheng, L. Zhou, Y. Cao, C. Zhuang, and Q. Wang, "A speech enhancement method based on dual-path phase-aware GAN networks," in *Proc. 13th Int. Conf. Commun., Circuits Syst. (ICCCAS)*, May 2024, pp. 315–320.
[108] Y. He, Q. Du, and L. Duo, "Phase-aware speech enhancement using multi-head attention and complex convolutions in SEGAN," in *Proc. IEEE 5th Int. Conf. Electron. Technol., Commun. Inf. (ICETCI)*, May 2025, pp. 1099–1103.
[109] H. Wang and X. Jiao, "Phase-aware speech enhancement with dual-stream architecture and MetricGAN," in *Proc. Int. Conf. Electr. Autom. Artif. Intell. (ICEAAI)*, Jan. 2025, pp. 1178–1184.
[110] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[111] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," School Inform., Centre Speech Technol. Res. (CSTR), Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., 2017, doi: 10.7488/ds/2117 .
[112] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," School Inform., Centre Speech Technol. Res. (CSTR), Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., 2019.
[113] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013, pp. 1–14.
[114] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1–1.1,"Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NIST IR 4930, 1993, doi: 10.6028/NIST.IR.4930.
[115] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[116] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[117] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, Oct. 2020, pp. 2492–2496.
[118] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2013, pp. 1–4.
[119] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.
[120] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[121] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

JUSTINA RAMONAITĖ (Graduate Student Member, IEEE) received the M.S. degree in mathematical sciences from the Faculty of Mathematics and Informatics, Vilnius University, Lithuania, in 2023. She is currently pursuing the Ph.D. degree in informatics with the Institute of Data Science and Digital Technologies, Vilnius University.

Her main research interest includes speech signal processing and enhancement.

GRAŽINA KORVEL (Member, IEEE) received the Ph.D. degree from the Institute of Data Science and Digital Technologies, Vilnius University, in 2013. Since 2022, she has been a member of the Young Academy of the Lithuanian Academy of Sciences. Currently, she is a Research Fellow and a Professor with the Institute of Data Science and Digital Technologies, Vilnius University. Her research interests include speech and music signal processing, natural language processing, the development of mathematical models, and the applications of computational intelligence. She received acknowledgment from the Prime Minister of Lithuania for her obtained scientific results, in 2013 and 2019, and was awarded the Rector's Science Award at Vilnius University for her outstanding scientific achievements, in 2025.

GINTAUTAS TAMULEVIČIUS (Senior Member, IEEE) received the Ph.D. degree in computer science, in 2003. He is currently a Senior Researcher with Vilnius University, Lithuania. His main activities are related to research, teaching, and research and development projects. His main responsibilities are administration, the management of research project teams, and the study process. His research interests include speech signal processing, digital signal processing, and neural network-based signal processing.