A Survey of Advancing Audio Super-Resolution and Bandwidth Extension from Discriminative to Generative Models

Anonymous authors
Paper under double-blind review

Abstract

Audio super-resolution (SR), also referred to as bandwidth extension (BWE), aims to reconstruct high-fidelity signals from low-resolution (LR) or band-limited (BL) observations, an inherently ill-posed task due to the ambiguity of missing high-frequency (HF) content. This survey provides a comprehensive overview of the field, with a particular focus on the paradigm shift from discriminative mapping to modern generative modeling. We first review early discriminative deep neural network (DNN) models, which formulate BWE/SR as a deterministic mapping problem and are prone to regression-to-the-mean effects and spectral over-smoothing. We then systematically review generative approaches, including autoregressive (AR) models, variational autoencoders (VAEs), generative adversarial networks (GANs), diffusion and score-based models, flow-based methods, and Schrödinger bridges. Across these approaches, we examine key design aspects, including representation domain, architecture, conditioning mechanisms, and trade-offs among reconstruction fidelity, perceptual quality, robustness, and computational efficiency. Furthermore, we discuss emerging directions involving large language models (LLMs) and multimodal foundation models, and highlight open challenges in perceptual evaluation, phase modeling, and real-world generalization. By providing a structured taxonomy and unified perspective, this survey establishes a comprehensive foundation and offers a practical roadmap for advancing BWE/SR from deterministic point estimation toward distribution-aware generative modeling.

1 Introduction

Audio super-resolution (SR) (Kuleshov et al., 2017) and bandwidth extension (BWE) (Abel & Fingscheidt, 2017) are essential tasks in artificial intelligence and audio research communities. They are widely used in applications such as telecommunications (Li & Lee, 2015), hearing aids (Van Eeckhoutte et al., 2020), and legacy recording restoration (Moliner et al., 2024), and can benefit downstream tasks including speech and speaker recognition (Li et al., 2019; Yamamoto et al., 2019). Although these terms are defined from different perspectives and arise from different historical contexts, they share the common objective of restoring high-fidelity audio from bandwidth-constrained observations. From a learning perspective, BWE/SR can be formulated as ill-posed reconstruction problems in which missing high-frequency (HF) content must be inferred from band-limited (BL) or low-resolution (LR) observations. Because the missing HF content is not uniquely determined by the input, a single degraded waveform may correspond to multiple perceptually plausible broadband (BB) or high-resolution (HR) reconstructions. This inherent one-to-many ambiguity makes BWE/SR a useful testbed for conditional distribution modeling beyond deterministic point estimation.

Traditional techniques, including source-filter models (Makhoul & Berouti, 1979), codebook mapping (Unno & McCree, 2005), Gaussian mixture models (GMMs) (Ohtani et al., 2014), and hidden Markov models (HMMs) (Chen & Parsa, 2004), relied on structured priors and simplified acoustic assumptions, often producing over-smoothed spectra and audible artifacts due to limited modeling capacity. The advent of deep learning (DL) shifted BWE/SR toward data-driven modeling of complex spectral relationships. Early deep neural network (DNN) architectures, such as multilayer perceptrons (MLPs) (Li et al., 2015; Liu et al., 2015; Wang et al., 2015), recurrent neural networks (RNNs) (Schmidt & Edler, 2018; Birnbaum et al., 2019; Hou et al., 2020; Edraki et al., 2024), and convolutional neural networks (CNNs) (Kuleshov et al., 2017; Lim et al.,

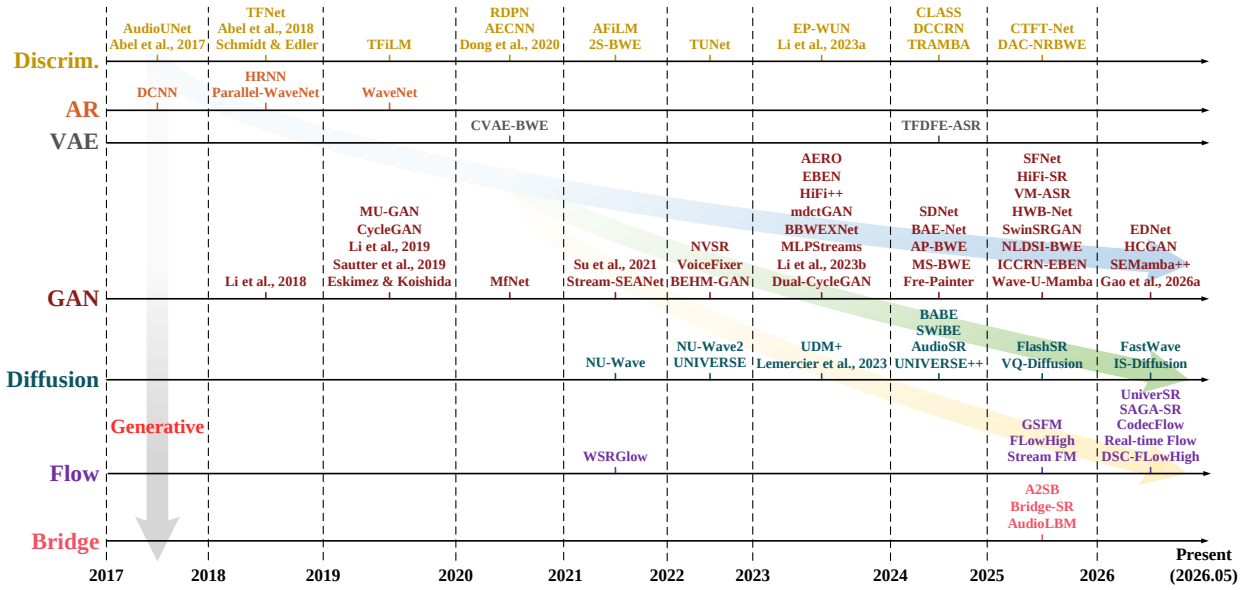


Figure 1: **Timeline of methodological evolution in BWE and SR (2017–present).** The trajectory highlights a clear recent generative tendency: after the early dominance of deterministic models, modern likelihood-based or score-based generative approaches are increasingly shaping the state-of-the-art paradigm, reflecting a generative shift from point estimation to conditional distribution matching for perceptually plausible HF reconstruction.

2018; Wang & Wang, 2020; Lin et al., 2021; Tamiti et al., 2025; Li et al., 2026), demonstrated superior performance over statistical models by directly learning non-linear mappings from degraded observations to clean targets. Subsequent research introduced more sophisticated structures, including U-Net variants (Kuleshov et al., 2017), Transformers (Vaswani et al., 2017), and Mambas (Shams et al., 2024), to better capture fine-grained spectral details and model long-range temporal dependencies. However, despite their success, these discriminative models frequently suffer from regression-to-the-mean behavior when trained with standard distance-based losses, resulting in spectral over-smoothing and a lack of HF richness. These limitations are especially pronounced in BWE/SR, where the missing HF content is inherently ambiguous and may admit multiple plausible reconstructions.

Recent advances in generative modeling have provided new paths to address this ambiguity by modeling the conditional distribution of plausible HF content rather than predicting a single deterministic target. Earlier generative frameworks—including autoregressive (AR) models (Gupta et al., 2019), variational autoencoders (VAEs) (Bachhav et al., 2020), and generative adversarial networks (GANs) (Hauret et al., 2023)—introduced non-deterministic reconstruction. Yet these methods face distinct trade-offs: AR models provide fine-grained temporal modeling at the cost of high inference latency, while VAEs offer structured latent spaces but often lack HF fidelity. Although GANs sharpen spectral details via adversarial training, they are frequently hindered by training instability and mode collapse. Conversely, more recent research leverages likelihood- and score-based paradigms, such as diffusion probabilistic models (Ho et al., 2020; Nichol & Dhariwal, 2021), flow-based methods (Lipman et al., 2022), and Schrödinger bridges (De Bortoli et al., 2021), provide alternative ways to model conditional audio distributions through iterative denoising, continuous-time dynamics, or stochastic bridges between degraded and clean signals. These methods have broadened the design space for BWE/SR, introducing new trade-offs among sample quality, diversity, controllability, computational cost, and inference latency. To contextualize this modeling shift from discriminative formulations to generative designs, Fig. 1 summarizes representative BWE/SR studies over recent years.

Despite the rapid advancements in BWE/SR, the field currently lacks a systematic work that articulates the significant paradigm shift from discriminative mapping to modern generative modeling frameworks. Existing surveys (Prasad & Kumar, 2016) remain fragmented and increasingly outdated: they primarily emphasize

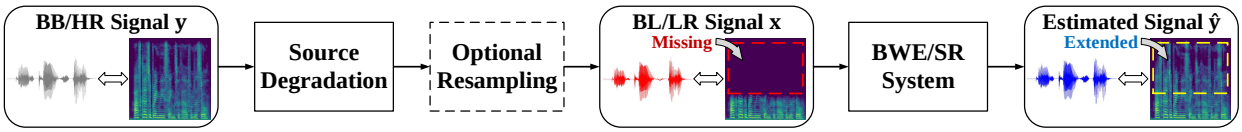


Figure 2: **Signal flow diagram of BWE/SR.** The degradation process removes HF spectral bandwidth from a reference signal $\mathbf{y}$, followed by an optional resampling stage, to produce the observation $\mathbf{x}$. In practical applications, this reference signal is not available. The BWE/SR system then estimates the reconstruction $\hat{\mathbf{y}}$ from the observation $\mathbf{x}$. Waveforms and spectrograms at each stage visualize the transition from the reference signal to the degraded input and finally to the recovered signal.

classical signal-processing approaches and provide limited coverage of early DNN-based models. Moreover, the definitional relationship between SR and BWE remains underexplored, leaving their distinctions and common objectives ambiguous across communities. To the best of our knowledge, this survey addresses this critical lack by providing the ***first comprehensive overview*** that tracks the evolution of both BWE and SR through this lens. We systematically organize the major classes of generative models, highlighting the shift from simple point estimation to modeling complex, high-dimensional audio distributions. By distilling the architectural principles, conditioning mechanisms, and inherent trade-offs of these varied approaches, this work provides researchers with practical guidance for designing high-fidelity audio systems. Furthermore, we identify emerging trends—such as large language models (LLMs) and multimodal foundation models—and highlight open challenges in perceptual evaluation and computational efficiency to inform future research directions in the generative era. Our contributions are summarized below:

- We formulate BWE and SR under a unified framework of HF reconstruction and clarify the relationship and distinctions between the two terms.
- We organize the literature by modeling paradigms, highlighting the progression from discriminative neural networks to modern generative models.
- We compare the major generative families and emphasize their conditioning mechanisms, training objectives, sampling procedures, and practical trade-offs.
- We discuss emerging trends and highlight open challenges to guide future research.

The remainder of this paper is organized as follows. Section 2 introduces the formal problem formulation and survey scope. Section 3 establishes the learning framework for BWE/SR, including input representations, training targets & objective functions, and spectral mapping paradigms. Section 4 reviews benchmark datasets and evaluation metrics. Section 5 surveys discriminative modeling approaches, while Section 6 provides an in-depth analysis of generative methods. Section 7 presents a comparative analysis of these paradigms, highlighting their fundamental trade-offs, as well as remaining challenges and future research directions. Finally, Section 8 concludes the paper.

2 Background

2.1 Problem Formulation

The tasks of BWE, SR, and speech super-resolution (SSR) can each be formulated as learning a function

$$f: \mathbf{x} \mapsto \mathbf{y}, \tag{1}$$

where $\mathbf{x} = [x[1], \ldots, x[N_1]]^\top$ and $\mathbf{y} = [y[1], \ldots, y[N_2]]^\top$ are one-dimensional (1-D) discrete audio sequences with N_1 and N_2 samples, respectively, and $x[n]$ and $y[n]$ denote the corresponding scalar-valued samples of $\mathbf{x}$ and $\mathbf{y}$. For BWE, the core objective is to restore missing HF spectral components, where the input $\mathbf{x}$ and target $\mathbf{y}$ correspond to BL and BB signals, respectively. In simulation, $\mathbf{x}$ is typically obtained by applying a low-pass filter to $\mathbf{y}$ to remove spectral content above a specific cutoff frequency. Some studies (Wang et al.,

Table 1: **Comparison of BWE, SR, and SSR** in terms of their underlying degradation processes, input-target relationships, sampling rate changes, primary reconstruction objectives, and application domains.

Task	Source Degradation	Input → Target	Rate	Primary Objective	Signal Scope
BWE	Bandwidth Reduction	BL → BB Signal	↑ / =	Restoring Bandwidth	Primarily Speech
SR	Temporal Downsampling	LR → HR Signal	↑	Restoring Time Resolution	General Audio
SSR	Temporal Downsampling	LR → HR Signal	↑	Restoring Time Resolution	Speech Only

2015; Liu et al., 2016; Abel et al., 2018; Li et al., 2021; Hauret et al., 2023; Li et al., 2023a) perform BWE by maintaining identical sampling rates for inputs and targets, resulting in equal sequence lengths ($N_1 = N_2$). In this configuration, the BL signal is represented on the same temporal grid as the target BB signal. Conversely, others (Liu et al., 2015; Li & Lee, 2015; Gu et al., 2016; Dong et al., 2020; Bachhav et al., 2020; Su et al., 2021) use a lower sampling rate for the input, resulting in $N_1 < N_2$. This requires the model to resolve temporal resolution mismatches in addition to spectral recovery, thereby aligning the formulation closely with SR. Conventional BWE assumes a known, fixed degradation process (e.g., a predefined band-limiting filter), whereas blind BWE (Moliner et al., 2024) relaxes this assumption and must infer unknown degradations from the observed signal. Although BWE has historically focused on speech signals to improve intelligibility in telecommunications, recent advancements have extended these techniques to musical applications (Moliner & Välimäki, 2022; Moliner et al., 2024; Hernandez-Olivan et al., 2026).

Conversely, the term SR emerged more recently as a direct conceptual parallel to image super-resolution (Dong et al., 2015; Yue et al., 2016; Yang et al., 2019b) in computer vision. In the SR setting, the input $\mathbf{x}$ is a LR signal sampled at rate R, and the target $\mathbf{y}$ is the corresponding HR signal sampled at a higher rate $R' > R$, with the upsampling ratio defined as

$$r = \frac{R'}{R}. \tag{2}$$

Consequently, SR is explicitly framed as a temporal interpolation problem where the model must increase the sampling rate, mapping the input sequence of length N_1 to an output of length $N_2 = rN_1$. Under the assumption that $\mathbf{x}$ and $\mathbf{y}$ are uniformly sampled from finite-bandwidth continuous-time signals, the Nyquist-Shannon sampling theorem (Jerri, 1977) guarantees that the maximum representable bandwidth is $R/2$ and $R'/2$, respectively. Therefore, SR inherently requires the restoration of HF content, in contrast to a standard upsampling operation that only increases the sampling rate. Without additional HF recovery, simply performing BL interpolation would provide no new information above the original Nyquist limit $R/2$. Additionally, SSR is defined similarly to SR but is restricted to speech signals, whereas SR targets general audio, including speech, music, and sound effects.

Fig. 2 illustrates the signal flow for the degradation and reconstruction processes of BWE and SR, where both waveforms and spectrograms are shown at each stage to visualize the signal transformations. Table 1 summarizes the key differences and commonalities between BWE, SR, and SSR. While BWE is defined by the restoration of spectral bandwidth, SR and SSR are characterized by temporal interpolation. Despite these differing definitions, all three tasks can be unified as ill-posed inverse problems—specifically spectral or temporal inpainting. In this framework, missing HF information is statistically reconstructed through data-driven modeling to approximate the natural audio characteristics learned from training data. Throughout this paper, we use BWE, SR, and SSR to refer to specific works, while the notation BWE/SR collectively denotes the broader research field.

2.2 Survey Scope

To maintain a focused survey of BWE and SR, we establish specific scope constraints regarding the input signal and restoration task. When the input $\mathbf{x}$ is both BL and corrupted by additive noise, the task is defined as noise-robust BWE (Lin et al., 2023; Yang et al., 2024; Lin et al., 2024; Liu et al., 2025a). Furthermore, universal speech enhancement (SE), also referred to as speech restoration, has recently emerged

as a prominent research direction, as promoted by the URGENT challenge (Zhang et al., 2024). Such approaches (Serrà et al., 2022; Liu et al., 2022b; Andreev et al., 2023; Lemercier et al., 2023; Scheibler et al., 2024; Ku et al., 2025; Welker et al., 2025; Kwak et al., 2026; Gao et al., 2026a; Lee & Choi, 2026; Hsieh & Braun, 2026) are designed to handle diverse distortions, including additive noise, reverberation, clipping, and bandwidth limitation. We include these works as they remain within the single-observation inverse problem framework. Conversely, we exclude methods that leverage multiple input modalities, such as the fusion of bone-conducted and air-conducted speech (Yu et al., 2020; Wang et al., 2022). We categorize these approaches as multi-modal SE rather than pure BWE/SR, as their performance relies on sensor fusion rather than the data-driven spectral modeling considered in this work.

3 Learning Framework

3.1 Input Representations

Early DNN-based methods have primarily relied on spectral-domain features that explicitly characterize the short-time spectral shape and excitation information of speech signals. Typical examples include the log-power spectrum (LPS) (Li & Lee, 2015; Li et al., 2015; Liu et al., 2015; Gu et al., 2016; Liu et al., 2016; Abel et al., 2018; Schmidt & Edler, 2018; Eskimez & Koishida, 2019; Bachhav et al., 2020), line spectral frequency (LSF) (Liu et al., 2015; 2016; Li et al., 2018), mel-frequency cepstral coefficient (MFCC) (Liu et al., 2015; 2016), and band-pass voicing coefficient (BPVC) (Liu et al., 2015; 2016). Beyond these frequency-domain descriptors, other works (Abel et al., 2016; Abel & Fingscheidt, 2017) explore time-domain and modulation-domain features such as the autocorrelation function, zero-crossing rate, gradient index, and normalized relative frame energy. In addition to these handcrafted features, some approaches instead learn compact spectral-envelope representations directly from data, using methods such as the Gaussian-Bernoulli restricted Boltzmann machine (Wang et al., 2015), STRAIGHT vocoder (Gu & Ling, 2015), and a specially structured DNN (Gu et al., 2016).

The advent of DNN architectures has motivated a shift toward learning directly from time-domain and frequency-domain audio representations. In the literature, numerous works operate directly on the raw waveform (Kuleshov et al., 2017; Gu & Ling, 2017; Lim et al., 2018; Ling et al., 2018; Birnbaum et al., 2019; Dong et al., 2020; Hou et al., 2020; Wang & Wang, 2020; Wang & Wang, 2021; Rakotonirina, 2021; Li et al., 2021; Nguyen et al., 2022; Kim et al., 2024a), leveraging convolutional encoders to learn rich temporal-spectral representations while implicitly preserving phase information and avoiding the constraints and preprocessing artifacts associated with fixed time-frequency transforms. A special case is EBEN (Hauret et al., 2023), which operates on subband waveforms derived from Pseudo Quadrature Mirror Filter (PQMF) decomposition. In addition to taking waveforms as the input, (Lim et al., 2018) and (Dong et al., 2020) further incorporate the short-time Fourier transform (STFT) magnitude spectrogram and enable dual-branch modeling in the time and frequency domains. Additionally, log-mel spectrograms and mel-spectrograms (Gupta et al., 2019; Liu et al., 2022a;b; Andreev et al., 2023; Kim et al., 2024b; Liu et al., 2024a; Im & Nam, 2025; Zhao et al., 2025; Yun et al., 2025) have been widely used due to their strong alignment with human auditory perception. They are computed by mapping STFT magnitudes through perceptually spaced mel filterbanks, with or without an additional logarithmic transformation, respectively. However, both STFT magnitude spectrograms and mel-spectrograms lack the phase information and therefore require additional algorithms for phase recovery.

To address this limitation, several studies instead use STFT complex spectrograms (Mandel et al., 2023; Soltanmohammadi et al., 2023; Edraki et al., 2024; Lu et al., 2024a;b; Yu et al., 2024; Tamiti et al., 2025; Ku et al., 2025; Zhang & Ruan, 2025), capturing both real and imaginary components—or equivalently, log-magnitude and phase spectra—to enable more accurate waveform reconstruction and phase-aware modeling. Moreover, recent studies (Liu et al., 2021; Shuai et al., 2023) employ the modified discrete cosine transform (MDCT) and leverage the real-valued spectrogram representation. This technique supports phase-aware reconstruction by encoding phase information into the sign of the MDCT coefficients. MDCT removes inherent redundancy in the complex spectrogram and operates with fewer coefficients for the same temporal resolution, providing a compact and computationally efficient solution. Finally, recent trends (Fang et al., 2025; Liu et al., 2025a; Zhang et al., 2026) employ discrete latent tokens extracted from a neural audio codec (NAC) as features, producing a compact and semantically rich representation that enables high-fidelity

and perceptually consistent audio reconstruction. Overall, the evolution of input representations reflects a trajectory from handcrafted spectral features toward richer, phase-aware, and semantically grounded representations, driven by both architectural advances and increasing demands of high-fidelity reconstruction.

3.2 Training Targets and Objective Functions

Early BWE methods have relied on spectral-domain regression, which predicts the HF or fullband LPS (Li & Lee, 2015; Li et al., 2015; Liu et al., 2015; Gu et al., 2016; Liu et al., 2016; Eskimez & Koishida, 2019), HF cepstral vector (Abel & Fingscheidt, 2017; Abel et al., 2018), HF posterior probabilities modeled by HMM (Abel et al., 2016), or other spectral-envelope representations (Wang et al., 2015; Gu & Ling, 2015; Schmidt & Edler, 2018; Li et al., 2018; Bachhav et al., 2020). These approaches are typically trained using the minimum mean-square error (MMSE) criterion. The missing phase information is reconstructed by mirroring the low-frequency (LF) phase, followed by inverse STFT (iSTFT) and overlap-add synthesis. Such spectral-domain training targets provide a smooth and meaningful representation that is easier to model and leads to stable training. However, they discard phase information and cannot capture the stochastic structure of HF components, often resulting in over-smoothed spectra and limited perceptual realism.

Given these limitations, BWE/SR research has shifted toward direct waveform prediction (Kuleshov et al., 2017; Gu & Ling, 2017; Wang et al., 2018; Lim et al., 2018; Ling et al., 2018; Birnbaum et al., 2019; Gupta et al., 2019; Li et al., 2019; Wang & Wang, 2020; Hou et al., 2020; Dong et al., 2020; Hao et al., 2020; Wang & Wang, 2021; Li et al., 2021; Su et al., 2021; Rakotonirina, 2021; Nguyen et al., 2022; Li et al., 2023a; Andreev et al., 2023; Kim et al., 2024a; Zhao et al., 2025; Lee & Kim, 2025), where the model implicitly learns both amplitude and phase and can capture fine-grained HF details without relying on explicit spectral synthesis. Typical time-domain loss functions include the mean squared error (MSE) loss (Kuleshov et al., 2017; Lim et al., 2018; Birnbaum et al., 2019; Hao et al., 2020; Rakotonirina, 2021; Kim et al., 2024a) and mean absolute error (MAE) loss (Li et al., 2019; Wang & Wang, 2020; Dong et al., 2020; Wang & Wang, 2021; Su et al., 2021; Li et al., 2023a) applied directly to the waveform, scale-invariant signal-to-distortion ratio (SI-SDR) loss (Hou et al., 2020), and cross-entropy (CE) loss (Gu & Ling, 2017; Wang et al., 2018; Ling et al., 2018; Gupta et al., 2019). In addition, several studies incorporate frequency-domain losses for waveform targets, including MSE loss on MFCC (Li et al., 2019), MAE loss on STFT spectrograms (Wang & Wang, 2020; Wang & Wang, 2021; Su et al., 2021), MAE loss on mel or log-mel spectrograms (Hao et al., 2020; Su et al., 2021; Zhao et al., 2025; Lee & Kim, 2025), multi-scale spectral energy loss (Li et al., 2023a), and multi-scale STFT loss (Nguyen et al., 2022; Kim et al., 2024a; Lee & Kim, 2025). Later research has shown that combining time-domain and frequency-domain losses provides complementary constraints that enforce both sample-level waveform accuracy and perceptually important spectral structure.

Parallel to waveform prediction, another major class of approaches uses the STFT complex spectrograms (Mandel et al., 2023; Soltanmohammadi et al., 2023; Yang et al., 2024; Edraki et al., 2024; Lu et al., 2024a;b; Yu et al., 2024; Tamiti et al., 2025; Zhang & Ruan, 2025) or MDCT spectrograms (Shuai et al., 2023; Yuan et al., 2025) as training targets, allowing models to explicitly or implicitly account for phase information in the spectral domain. Similar loss functions are used, including the MAE loss (Soltanmohammadi et al., 2023; Edraki et al., 2024; Yu et al., 2024) and SI-SDR loss (Tamiti et al., 2025) applied to the waveform, while the most widely adopted frequency-domain objective is the multi-scale STFT loss (Mandel et al., 2023; Soltanmohammadi et al., 2023; Yang et al., 2024; Yu et al., 2024; Tamiti et al., 2025; Zhang & Ruan, 2025). Moreover, AP-BWE (Lu et al., 2024a) and its subsequent extension (Lu et al., 2024b) introduce a combination of log-amplitude MSE loss, phase anti-wrapping loss, and short-time complex spectral MSE loss to jointly model amplitude and phase. Additionally, several approaches (Liu et al., 2022a;b; 2024a; Kim et al., 2024b; Im & Nam, 2025; Yun et al., 2025) adopt a two-stage pipeline where the model first predicts a mel-spectrogram as an intermediate representation and then applies a pretrained vocoder to synthesize the waveform. In the first stage, the MAE loss is typically used to achieve accurate mel-spectrogram prediction.

Regarding diffusion and score-based models (Lee & Han, 2021; Han & Lee, 2022; Yu et al., 2023; Lin et al., 2024; Liu et al., 2024a; Im & Nam, 2025; Fang et al., 2025), the training targets are typically the additive noise or an equivalent denoising objective at each diffusion step. In this case, the model predicts either the noise or velocity vector in the waveform domain, spectrogram domain, or a learned latent space. These models are optimized using MAE or MSE losses applied in the corresponding domain. For GAN-based approaches,

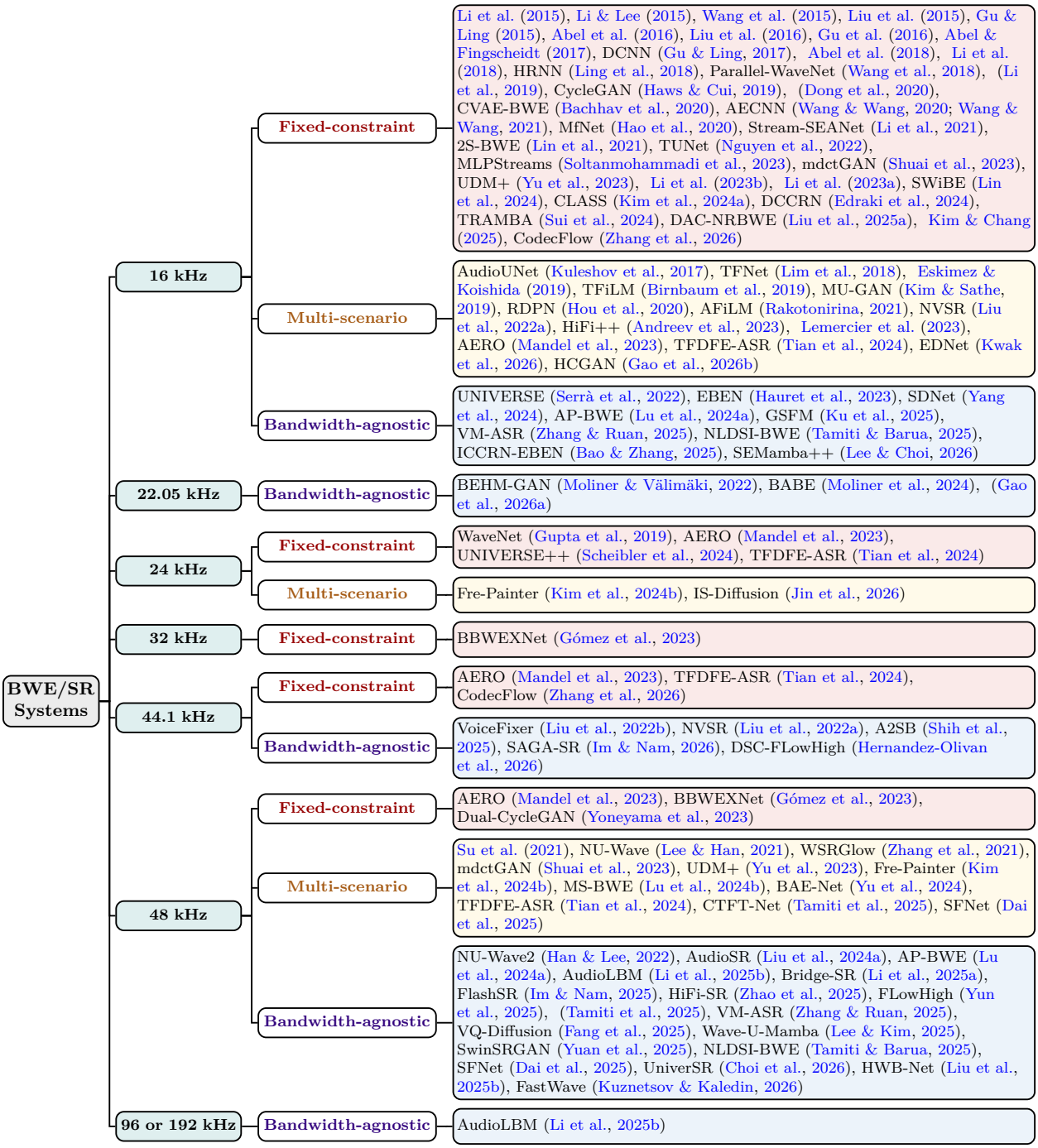


Figure 3: **Taxonomy of BWE/SR Literature.** Existing methods are organized by target sampling rates $\{16, 22.05, 24, 44.1, 48, 96, 192\}$ kHz and further categorized according to their spectral mapping paradigm {fixed-constraint, multi-scenario, bandwidth-agnostic} in training settings.

the training target can be either waveforms or spectrograms, and a combination of adversarial loss and reconstruction loss is employed to enhance perceptual quality. In flow-based approaches (Yun et al., 2025; Ku et al., 2025), training is performed by regressing an analytically defined velocity field that transports a simple base distribution to the target BB/HR signal distribution using the MSE loss. In bridge models (Li et al., 2025a;b), training aims to estimate a drift field that bridges a source and a target distribution by minimizing

the Kullback-Leibler (KL) divergence between the induced path measure and a reference stochastic process under boundary constraints, either directly in signal space or within a learned latent representation space. Finally, other methods use alternative training targets such as PQMF-decomposed subband signals (Hauret et al., 2023) or latent token sequences (Fang et al., 2025; Liu et al., 2025a; Zhang et al., 2026). Taken together, the evolution of training targets mirrors the shift in model paradigms: early spectral-regression objectives gave way to direct waveform prediction, which in turn has been complemented by phase-aware spectral targets and the objective functions of modern generative frameworks.

3.3 Spectral Mapping Paradigms

Existing BWE/SR frameworks can be categorized by their spectral mapping paradigm, which describes how the input cutoff frequency is configured during training—whether fixed at a single value, varied across a discrete set, or treated as a continuous free parameter. Three paradigms can be distinguished: *fixed-constraint*, *multi-scenario*, and *bandwidth-agnostic*. The fixed-constraint paradigm restricts the experimental protocol to a unique spectral mapping, in which a single cutoff frequency is used for both training and evaluation. This setting is typical of specialized systems such as legacy telephony restoration, where the speech bandwidth is extended from narrowband (NB) (e.g., 300 Hz–3.4 kHz or 0–4 kHz) to wideband (WB) (e.g., 0–8 kHz). The multi-scenario paradigm trains and evaluates separate models for multiple discrete input cutoff frequencies, enabling improved robustness across heterogeneous bandwidth conditions compared to a single fixed-cutoff setting. For instance, AudioUNet (Kuleshov et al., 2017) employs distinct models for upscaling ratios of $2\times, 4\times$, and $6\times$ at a target sampling rate of 16 kHz, each optimized specifically for its corresponding input bandwidth. The bandwidth-agnostic paradigm employs a single universal model trained to handle a wide range of input cutoff frequencies, enabling many-to-one spectral bandwidth mapping within a unified framework. In this regime, input cutoff frequencies are either sampled from a continuous stochastic range or selected from a discrete set, as in NVSR (Liu et al., 2022a), which samples cutoff frequencies over 1–16 kHz, and AP-BWE (Lu et al., 2024a), which uses a discrete set $\{4, 6, 8, 12\}$ kHz. Recent works have favored this paradigm for its architectural robustness: the model implicitly infers the input spectral support and reconstructs the missing HF components across arbitrary frequency gaps without prior knowledge of the degradation parameters. Fig. 3 illustrates a hierarchical taxonomy of BWE/SR methods based on target sampling rates and spectral mapping paradigms.

4 Evaluations

4.1 Audio Datasets

Most existing BWE/SR approaches are primarily designed for speech signals and are commonly evaluated using standard speech corpora. In particular, the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1988) and the CSTR Voice Cloning Toolkit (VCTK) (Yamagishi, 2012) are the most frequently adopted benchmarks due to their high-quality recordings and phonetic diversity. The TIMIT database contains approximately 5.4 hours of English speech from 630 speakers recorded at a sampling rate of 16 kHz with detailed phonetic and word-level annotations. The VCTK corpus consists of roughly 44 hours of English speech from 109 speakers with diverse accents recorded at a sampling rate of 48 kHz. Other studies employ a wider range of speech datasets, including LibriSpeech (Panayotov et al., 2015), Wall Street Journal (WSJ0) corpus (Corpus, 1992), LibriTTS (Zen et al., 2019), HiFi-TTS (Bakhturina et al., 2021), VoiceBank-DEMAND (Botinhao et al., 2016), and LJSpeech (Ito & Johnson, 2017). The BL/LR signals are typically synthesized from these ground-truth BB/HR sources by applying low-pass filtering or downsampling to simulate bandwidth-limited acquisition.

Among SSR-focused works, VCTK at 48 kHz has emerged as the de facto benchmark, used both for training and evaluation by systems such as NU-Wave (Lee & Han, 2021), AudioSR (Liu et al., 2024a), and FlashSR (Im & Nam, 2025), making it the most consistently adopted corpus for SSR despite its original design as a multi-speaker text-to-speech resource. Conversely, several bone-conducted BWE approaches incorporate real-world sensor data. Edraki et al. (2024) recorded a proprietary corpus of own-voice signals using paired air and in-

ear microphones, while Li et al. (2023a;b) use the Elevoc Simultaneously-recorded Microphone/Bone-sensor (ESMB) speech corpus[1].

Beyond speech, BWE/SR has expanded to music and general sounds, requiring datasets that capture greater spectral complexity. Several studies (Kuleshov et al., 2017; Lim et al., 2018; Rakotonirina, 2021; Mandel et al., 2023; Liu et al., 2024a; Moliner et al., 2024; Im & Nam, 2025; Li et al., 2025b) incorporate diverse music datasets to evaluate performance on signals with richer spectral density and polyphonic structure than speech. Representative datasets include MUSDB18-HQ (Rafii et al., 2019), MoisesDB (Pereira et al., 2023), MedleyDB (Bittner et al., 2014), SDS (Manco et al., 2023), MAESTRO (Hawthorne et al., 2018), COCOChorales (Wu et al., 2022), and FMA-small (Defferrard et al., 2016). These datasets span a wide range of musical content, from solo recordings to multi-instrument mixtures, enabling rigorous evaluation of HF reconstruction and polyphonic detail. A subset of studies (Tian et al., 2024; Liu et al., 2024a; Im & Nam, 2025; Li et al., 2025b) further considers datasets of general sounds, such as ESC-50 (Piczak, 2015), FreeSound (Mei et al., 2024), and ShipsEar (Santos-Domínguez et al., 2016). These datasets cover diverse non-stationary and stochastic acoustic conditions—from urban noise to underwater signals—often lacking the harmonic regularity found in speech or music.

4.2 Evaluation Methods

4.2.1 Objective Metrics

Signal-to-Noise Ratio (SNR): Given N samples of a reference signal $\mathbf{s} = [s[1], \ldots, s[N]]^\top$ and its estimate $\hat{\mathbf{s}} = [\hat{s}[1], \ldots, \hat{s}[N]]^\top$, SNR (Box, 1988) is defined as

$$\mathrm{SNR}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left(\frac{\|\mathbf{s}\|^2}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2} \right). \tag{3}$$

SNR is a standard metric in the signal processing literature, commonly used to evaluate waveform distortion in the time domain. To better capture frame-level variations, the segmental signal-to-noise ratio (SegSNR) (Paliwal & Basu, 1987) computes the average SNR over short-time segments of an audio signal. Let us denote the l-th frame of $\mathbf{s}$ as $\mathbf{s}_l$ and the l-th frame of $\hat{\mathbf{s}}$ as $\hat{\mathbf{s}}_l$, $l = 1, \ldots, L$, where L represents the total number of frames. SegSNR is defined as

$$\mathrm{SegSNR}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{L} \sum_{l=1}^{L} 10 \cdot \log_{10} \left(\frac{\|\mathbf{s}_l\|^2}{\|\mathbf{s}_l - \hat{\mathbf{s}}_l\|^2} \right). \tag{4}$$

Both metrics are widely used but have limited sensitivity to perceptual quality in BWE/SR: because the generated HF components are stochastic and may not align sample-by-sample with the reference even when perceptually convincing, high SNR does not guarantee good HF reconstruction quality.

Scale-invariant Signal-to-Distortion Ratio (SI-SDR): SI-SDR (Kolbæk et al., 2020) evaluates reconstruction fidelity while explicitly removing sensitivity to global gain mismatch by optimally projecting the estimated signal onto the subspace spanned by the reference signal, which is defined as

$$\text{SI-SDR}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \cdot \log_{10} \left(\frac{\|\alpha \mathbf{s}\|^2}{\|\hat{\mathbf{s}} - \alpha \mathbf{s}\|^2} \right), \tag{5}$$

where

$$\alpha = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle}{\|\mathbf{s}\|^2}. \tag{6}$$

This metric evaluates reconstruction quality by measuring the energy ratio between the optimally scaled target signal and the residual distortion. Owing to its scale-invariant formulation and robustness to gain mismatch, SI-SDR has become a standard evaluation criterion for speech separation, SE, and BWE/SR in contemporary audio research. However, it shares the same fundamental limitation as SNR for BWE/SR:

[1] Available at: https://github.com/elevoctech/ESMB-corpus.

sample-level alignment between generated HF content and the reference is not required for perceptual fidelity, so SI-SDR can underrate perceptually high-quality reconstructions.

Log-Spectral Distance (LSD): LSD (Gray & Markel, 2003) measures reconstruction quality by computing the average discrepancy between the logarithmic magnitude spectra of the reference and reconstructed signals across frequencies. Let $\mathbf{S} \in \mathbb{C}^{T\times F}$ and $\hat{\mathbf{S}} \in \mathbb{C}^{T\times F}$ denote the STFT of $\mathbf{s}$ and $\hat{\mathbf{s}}$, respectively, where T and F denote the numbers of time frames and frequency bins in the spectrogram. Their magnitude spectra are defined as $\mathbf{X} = |\mathbf{S}| \in \mathbb{R}^{T\times F}$ and $\hat{\mathbf{X}} = |\hat{\mathbf{S}}| \in \mathbb{R}^{T\times F}$. LSD is then computed as

$$\mathrm{LSD}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{F}\sum_{f=1}^{F} D_{t,f}^2}, \tag{7}$$

with the per-frame, per-frequency log-spectral difference

$$D_{t,f} = \log_{10}\left(\frac{\hat{\mathbf{X}}_{t,f}}{\mathbf{X}_{t,f}}\right). \tag{8}$$

To evaluate different spectral regions separately, the LF LSD (LSD-LF) is used to assess reconstruction quality of the preserved LF region, while the HF LSD (LSD-HF) measures the quality of the extended HF region. They are defined as

$$\text{LSD-HF}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{F_{\mathrm{HF}}}\sum_{f\in\mathcal{F}_{\mathrm{HF}}} D_{t,f}^2}, \tag{9}$$

$$\text{LSD-LF}(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{F_{\mathrm{LF}}}\sum_{f\in\mathcal{F}_{\mathrm{LF}}} D_{t,f}^2}, \tag{10}$$

where $\mathcal{F}_{\mathrm{HF}}$ and $\mathcal{F}_{\mathrm{LF}}$ denote the sets of HF and LF bins, with F_{HF} and F_{LF} representing the numbers of bins in HF and LF regions, respectively. Lower scores in LSD, LSD-HF, and LSD-LF imply a more similar spectrum to the target. While LSD correlates better with perceptual quality than time-domain measures like SNR, it often fails to account for the quality of reconstructed signals that are perceptually authentic but not mathematically identical to the reference spectra.

Virtual Speech Quality Objective Listener (ViSQOL): ViSQOL (Hines et al., 2015) is a signal-based, full-reference metric that estimates perceptual quality using a spectro-temporal measure of similarity between reference and degraded signals. It outputs a Mean Opinion Score-Listening Quality Objective (MOS-LQO), where higher values indicate better quality. ViSQOL supports both speech and general audio: in speech mode (16 kHz), MOS-LQO ranges from 1 to 5, while in audio mode (48 kHz), it ranges from 1 to 4.75. Unlike SNR-based metrics, ViSQOL is better suited for BWE/SR evaluation as it reflects perceptual similarity rather than sample-level fidelity.

Perceptual Evaluation of Speech Quality (PESQ): PESQ (Beerends et al., 2002) is an ITU-T standardized metric that estimates perceptual speech quality by comparing reference and degraded signals. The original PESQ targets NB speech (300 Hz–3.4 kHz) with scores ranging from -0.5 to 4.5, while its WB extension (WB-PESQ)[2] supports 16 kHz speech with scores between 1.04 and 4.64. In both cases, higher values indicate better quality. However, PESQ is designed for distortions typical of telecommunications channels (e.g., coding artifacts, packet loss) rather than HF reconstruction, and can therefore be misleading for evaluating BWE/SR systems.

Short-Time Objective Intelligibility (STOI): STOI (Taal et al., 2010) is an objective metric that estimates speech intelligibility by correlating short-time temporal envelopes of reference and degraded signals, producing scores between 0 and 1, where higher values indicate better intelligibility. Its extension, ESTOI (Jensen & Taal, 2016), incorporates longer temporal dependencies to better match human perception under noisy and reverberant conditions. STOI/ESTOI are most relevant for NB speech BWE, where intelligibility is the primary goal, but are less informative for music SR or high-fidelity speech reconstruction, where perceptual quality is more critical.

[2] https://www.itu.int/rec/T-REC-P.862.2.

4.2.2 Subjective Evaluations

Mean Opinion Score (MOS): MOS (Wester et al., 2015) is a subjective evaluation method standardized by the ITU-T for assessing perceived speech and audio quality. It is computed by averaging listener ratings, typically collected on a five-point Absolute Category Rating scale, where 1 denotes bad quality and 5 denotes excellent quality. As a direct measure of human perception, MOS serves as a reliable benchmark and is widely used to complement objective metrics in speech and audio research. In BWE/SR, MOS evaluations are increasingly conducted via crowdsourcing platforms such as Amazon Mechanical Turk, while automatic MOS predictors, such as MOSNet (Lo et al., 2019), DNSMOS (Reddy et al., 2021), NISQA (Mittag et al., 2021), and WV-MOS (Andreev et al., 2022), have emerged as scalable alternatives to full listening tests.

Preference Tests: Preference-based listening tests are subjective evaluations in which listeners compare audio samples produced by different systems and indicate their preference. In A/B tests, two audio samples are directly compared in terms of perceived quality or intelligibility, while ABX tests require listeners to identify which of two candidates is closer to a reference signal. These methods enable direct pairwise comparisons and are often used alongside MOS evaluations, providing higher sensitivity to subtle differences between high-performing BWE/SR systems.

Multiple Stimuli with Hidden Reference and Anchor (MUSHRA): The MUSHRA test[3] is a standardized subjective evaluation methodology designed to assess intermediate-quality audio systems. In this test, listeners are presented with multiple stimuli, including a hidden reference and one or more anchor signals, and are asked to rate the perceived audio quality on a continuous scale from 0 to 100. Compared with MOS-based evaluations, MUSHRA provides finer resolution and improved sensitivity for distinguishing perceptual differences among systems. In BWE/SR studies, the BL/LR input signal is typically used as the low-quality anchor, grounding the rating scale in the degradation being addressed.

In practice, LSD and MOS are the most consistently co-reported metrics across the BWE/SR literature—LSD for its direct spectral relevance to the HF reconstruction task and MOS for its alignment with human perceptual judgment—while MUSHRA and ViSQOL are increasingly adopted in higher-fidelity music and general audio settings.

5 Discriminative Models

Discriminative models for BWE and SR learn direct mappings from BL/LR representations to their BB/HR counterparts. This section categorizes existing approaches into MLP, RNN, CNN, Transformer, and Mamba-based frameworks, reflecting the progression from foundational to more advanced architectures. Table 2 summarizes representative methods across these paradigms.

5.1 Multilayer Perceptron (MLP)

Early studies (Li et al., 2015; Liu et al., 2015; Wang et al., 2015; Gu & Ling, 2015; Li & Lee, 2015) formulate BWE as a high-dimensional nonlinear mapping problem and employ MLPs to model the complex relationship between lower-band (LB) spectral envelopes and upper-band (UB) components. A common strategy is to first perform unsupervised pre-training using restricted Boltzmann machines (RBMs) (Ackley et al., 1985) to initialize the DNN weights and then discriminative fine-tuning with the MMSE criterion to estimate UB spectral parameters. The UB phase is typically recovered via inverse mirroring of the phase from the BL signal. Abel *et al.* (Abel et al., 2016; Abel & Fingscheidt, 2017; Abel et al., 2018) compare HMM-GMM, DNN-HMM hybrid, and MLP-only approaches for BWE, finding that direct MLP-based cepstral prediction with LB phase-copying achieves more accurate UB energy estimation and improved perceptual quality at low computational cost. The MLP-based approaches have outperformed conventional GMM-based methods across both objective and subjective evaluations.

[3] https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/en.

Table 2: **Taxonomy of discriminative models for BWE/SR** according to their architectural design. Some methods span multiple categories as they combine architectures (e.g., hybrid CNN-RNN designs).

Architecture	Literature
MLP	Li et al. (2015), Li & Lee (2015), Wang et al. (2015), Gu & Ling (2015), Liu et al. (2015), Abel et al. (2016), Abel & Fingscheidt (2017), Abel et al. (2018),
RNN	Liu et al. (2016), Gu et al. (2016), Schmidt & Edler (2018), TFiLM (Birnbaum et al., 2019), RDPN (Hou et al., 2020), 2S-BWE (Lin et al., 2021), DCCRN (Edraki et al., 2024)
CNN	AudioUNet (Kuleshov et al., 2017), TFNet (Lim et al., 2018), Schmidt & Edler (2018), TFiLM (Birnbaum et al., 2019), RDPN (Hou et al., 2020), Dong et al. (2020), AECNN (Wang & Wang, 2020; Wang & Wang, 2021), AFiLM (Rakotonirina, 2021), 2S-BWE (Lin et al., 2021), TUNet (Nguyen et al., 2022), EP-WUN (Lin et al., 2023), Li et al. (2023a), DCCRN (Edraki et al., 2024), CLASS (Kim et al., 2024a), CTFT-Net (Tamiti et al., 2025),
Transformer	AFiLM (Rakotonirina, 2021), TUNet (Nguyen et al., 2022), TRAMBA (Sui et al., 2024), CLASS (Kim et al., 2024a), DAC-NRBWE (Liu et al., 2025a), CTFT-Net (Tamiti et al., 2025)
Mamba	TRAMBA (Sui et al., 2024)

5.2 Recurrent Neural Network (RNN)

RNNs constitute an important model family for audio signal processing because they are designed to capture both short-range and long-range temporal dependencies. Gu et al. (2016) enhance BWE by incorporating linguistic information through bottleneck features extracted from an MLP-based HMM state classifier, providing a compact representation of both acoustic and phonetic cues. To improve the reduction of discontinuities from frame-independent mapping, this method employs a deep RNN with stacked long short-term memory (LSTM) layers, outperforming both GMM-based and MLP-based approaches. Liu et al. (2016) propose a two-stage BWE approach that uses a bidirectional LSTM (BLSTM) to estimate the UB LPS from LB acoustic features, followed by exemplar-based sparse refinement to mitigate over-smoothing before combining with LB components.

Subsequent works (Schmidt & Edler, 2018; Birnbaum et al., 2019; Hou et al., 2020; Lin et al., 2021; Edraki et al., 2024), further explore hybrid CNN-RNN architectures that combine convolutional feature extraction with recurrent temporal modeling to improve BWE/SR performance. Schmidt & Edler (2018) present a BWE system that estimates UB spectral envelope from LB power spectra using a regressive CNN-LSTM architecture, which improves speech quality without additional algorithmic delay and can be integrated into modern speech and audio codecs. Birnbaum et al. (2019) introduce Temporal Feature-Wise Linear Modulation (TFiLM), a mechanism that captures long-range dependencies by modulating convolutional activations with contextual information extracted by an RNN. This hybrid design effectively expands the receptive field of convolutional sequence models with minimal computational overhead, achieving superior SR performance over the MLP-based model. Hou et al. (2020) combine acoustic features with a low-dimensional feature vector that summarizes speaker-specific characteristics and phonetic posteriorgrams (PPGs) and propose a residual dual-path network (RDPN) for BWE. The RDPN segments the input sequence into overlapping chunks and applies BLSTM layers along intra-chunk and inter-chunk dimensions to capture utterance-level temporal dependencies while mitigating gradient-vanishing issues. Lin et al. (2021) introduce a two-stage BWE (2S-BWE) framework that leverages the complementary strengths of frequency-domain and time-domain modeling to reconstruct high-fidelity speech. The approach employs a spectrogram-based network—utilizing either a convolutional recurrent network (CRN) (Tan & Wang, 2018) or a temporal convolutional network (TCN) (Bai et al., 2018)—to estimate the HF spectral envelope, followed by a Wave-U-Net-based (Stoller et al., 2018) refinement stage trained with a multi-resolution STFT loss (Defossez et al., 2020) to improve phase consistency and enhance temporal details. Edraki et al. (2024) investigate BWE for bone-conducted speech using a deep complex convolutional-recurrent network (DCCRN) (Hu et al., 2020), which comprises a pair of complex two-dimensional (2-D) convolutional encoder-decoder and stacked

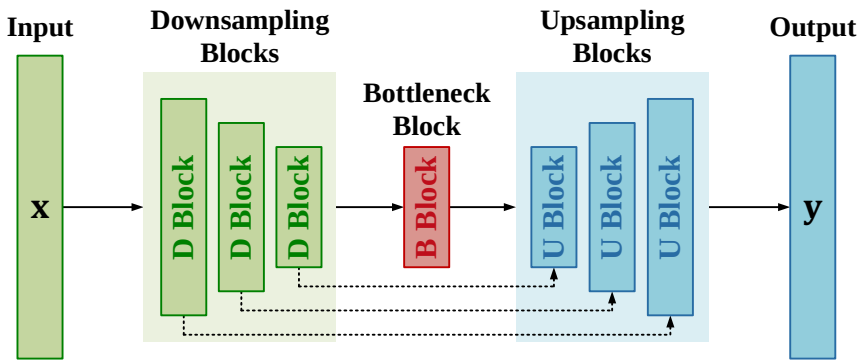


Figure 4: **The U-Net architecture.** It employs a symmetric encoder-decoder structure with multi-scale skip connections that align the corresponding stages, while the bottleneck block forms the most compact latent representation.

LSTM layers. To improve generalization in this highly speaker-dependent setting, the authors introduce a bottleneck module that disentangles speaker-specific characteristics from speech content, therefore enabling the architecture to adapt more effectively to unseen speakers.

5.3 Convolutional Neural Network (CNN)

CNN-based BWE/SR models can be broadly grouped into three categories: plain feed-forward CNNs (Schmidt & Edler, 2018), autoencoder-based CNNs (Dong et al., 2020; Wang & Wang, 2020; Hou et al., 2020; Wang & Wang, 2021; Edraki et al., 2024), and U-Net-based CNNs (Kuleshov et al., 2017; Lim et al., 2018; Birnbaum et al., 2019; Lin et al., 2021; Rakotonirina, 2021; Nguyen et al., 2022; Lin et al., 2023; Li et al., 2023a; Kim et al., 2024a; Tamiti et al., 2025). Earlier research primarily adopted plain feed-forward CNNs that process features at a single scale via sequential convolutions or autoencoder-based CNNs that employ an encoder-decoder structure to compress features into a bottleneck representation before reconstruction. Dong et al. (2020) propose a BWE method based on a time-frequency network whose dual branches are constructed from an autoencoder architecture enhanced with channel-attention and non-local modules to achieve richer channel-wise and spatial feature modeling. By jointly predicting waveform-level phase in the time branch and spectral-magnitude information in the frequency branch, the model enables concurrent phase-magnitude modeling and avoids invalid STFT reconstruction. Wang & Wang (2020); Wang & Wang (2021) employ an autoencoder architecture called AECNN (Pandey & Wang, 2019) and introduce cross-domain training objectives for SSR, combining time-domain waveform losses with frequency-domain STFT losses to achieve balanced reconstruction of temporal and spectral structures. In addition, the transposed-convolution upsampling layers in AECNN are replaced with subpixel layers to mitigate artifacts and improve computational efficiency. As discussed in the previous section, Hou et al. (2020); Edraki et al. (2024) integrate LSTM modules into an autoencoder-based BWE framework. Although the above methods are effective, plain feed-forward CNNs are limited by single-scale receptive fields, while autoencoder-based models may suffer from information loss due to heavy bottleneck compression.

To address these limitations, later designs have evolved toward U-Net architectures that employ multi-scale encoder-decoder pathways with aligned skip connections, as illustrated in Fig. 4. Through progressive downsampling and upsampling across multiple resolutions, the multi-scale encoder-decoder pathways capture local details while integrating long-range contextual information, whereas the aligned skip connections facilitate direct cross-resolution feature propagation, mitigating information loss caused by compression and facilitating more accurate reconstruction. AudioUNet (Kuleshov et al., 2017), the first U-Net architecture applied to SR, operates directly in the raw waveform and is trained end-to-end with a time-domain MSE loss. In addition, the model incorporates 1-D subpixel-shuffling layers, enabling high-fidelity reconstruction and reducing artifacts relative to conventional upsampling across multiple upscaling ratios. Lim et al. (2018) introduce an end-to-end Time-Frequency Network (TFNet), in which both time-domain and frequency-domain

branches follow the architecture design patterns from AudioUNet (Kuleshov et al., 2017). By jointly optimizing the time-domain waveform and frequency-domain magnitude representations through a spectral fusion layer, the model implicitly models phase in the time domain and the magnitude in the frequency domain. This cross-domain formulation yields significantly higher-quality reconstructions than prior single-domain SR approaches. Li et al. (2023a) propose a U-Net-based bone-conducted speech BWE framework in which a shifted window-based attention mechanism (Liu et al., 2021) is integrated into each residual 1-D convolutional encoder-decoder block, enabling effective modeling of long-range temporal dependencies. In addition, the method replaces the conventional time-domain MSE loss with a multi-scale mel-spectral energy distance objective, resulting in improved HF restoration performance. EP-WUN (Lin et al., 2023) integrates a Wave-U-Net backbone (Stoller et al., 2018) with a triplet loss-based embedding polishing stage to simultaneously perform noise-robust BWE and latent-space signal purification. More recent works (Rakotonirina, 2021; Nguyen et al., 2022; Kim et al., 2024a; Sui et al., 2024; Tamiti et al., 2025; Liu et al., 2025a) integrate Transformers (Vaswani et al., 2017) into U-Net architectures, which are analyzed in the following section.

5.4 Transformer

Transformers have increasingly been adopted for BWE/SR due to their ability to capture long-range temporal and spectral dependencies through a global self-attention mechanism. Unlike convolutional models, which are constrained by local receptive fields, Transformers can model interactions between distant time-frequency regions, which are crucial for restoring HF harmonics and capturing phoneme-level context. Transformer-based discriminative models (Rakotonirina, 2021; Nguyen et al., 2022; Kim et al., 2024a; Tamiti et al., 2025; Liu et al., 2025a) are commonly embedded within U-Net structures, where self-attention modules replace or augment convolutional blocks at the bottleneck or intermediate feature levels, enabling global context modeling while retaining the multi-scale benefits of U-Net feature fusion. AFiLM (Rakotonirina, 2021) is among the first to introduce Transformers into SR, replacing RNN-based conditioning with a self-attention mechanism through an Attention-based Feature-Wise Linear Modulation layer. AFiLM enables the model to capture long-range temporal dependencies beyond the limits of local convolutions, which leads to faster training and improved reconstruction accuracy. Subsequently, Nguyen et al. (2022) propose Transformer-aided U-Net (TUNet), which employs a lightweight 1-D U-Net backbone with a Transformer encoder at the bottleneck, enabling global-dependency modeling while maintaining computational efficiency through aggressive time-downsampling. While the Transformer refines the compressed latent representation, the U-Net handles hierarchical feature extraction and reconstruction, resulting in a compact yet highly effective BWE architecture. Kim et al. (2024a) introduce a continual learning approach for SSR (CLASS), which extends TUNet (Nguyen et al., 2022) by integrating self-supervised pretraining and continual learning to prevent catastrophic forgetting when transitioning from masked speech modeling to SSR. This strategy preserves prior representational knowledge while improving reconstruction performance and cross-condition generalization. Tamiti et al. (2025) propose a complex-domain time-frequency network (CTFT-Net) that incorporates a conformer bottleneck into the U-Net to jointly recover the magnitude and phase from complex spectrograms. The complex global attention block (CGAB) effectively captures long-range time-frequency dependencies, resulting in superior HF reconstruction for SSR, particularly under extreme upsampling conditions. Liu et al. (2025a) propose a neural codec-based framework for noise-robust BWE in which a conformer-based neural network learns a discriminative mapping from noisy codec embeddings to clean codebook indices. The predicted discrete tokens are then decoded by a pre-trained neural audio codec to reconstruct the enhanced BB speech, achieving state-of-the-art performance among noise-robust BWE methods.

5.5 Mamba

Mamba (Gu & Dao, 2024) is a sequence modeling architecture based on selective state-space models (SSMs) that enables efficient modeling of long-range dependencies with linear computational complexity in sequence length. Compared with CNNs, which rely on finite receptive fields or deep stacking to capture long-term context, Mamba naturally models global temporal dependencies through its state-space formulation. In contrast to Transformer-based models, whose self-attention incurs quadratic complexity with respect to sequence length, Mamba achieves scalable and memory-efficient sequence modeling while maintaining strong expressive power. TRAMBA (Sui et al., 2024) presents a hybrid Transformer-Mamba SSR framework specifically for

deployment on mobile and wearable platforms. By pretraining on large-scale speech datasets and fine-tuning with only a small amount of user-specific bone-conduction data, TRAMBA effectively addresses data scarcity and reduces the burden of modality-specific data collection. Extensive evaluations show that TRAMBA achieves substantial improvements in speech quality and intelligibility while maintaining a compact model size and enabling highly efficient inference. Furthermore, real-system integration and user studies demonstrate that TRAMBA delivers robust SSR performance in noisy environments and significantly improves wearable battery life by reducing sensing and data transmission requirements. Beyond purely discriminative settings, Mamba-based designs have also been adopted within generative frameworks: Wave-U-Mamba (Lee & Kim, 2025), VM-ASR (Zhang & Ruan, 2025), EDNet (Kwak et al., 2026), HCGAN (Gao et al., 2026b), and SEMamba++ (Lee & Choi, 2026) integrate Mamba blocks into GAN generators (discussed in Section 6), and VQ-Diffusion (Fang et al., 2025) employs a Mamba-2 backbone within a discrete diffusion model for token-level BWE. This cross-paradigm adoption underscores the versatility of state-space sequence modeling as an efficient alternative to self-attention across the full spectrum of BWE/SR architectures.

Across the discriminative paradigm, the progression from MLPs to RNNs, CNNs, Transformers, and Mambas reflects a consistent drive toward richer temporal context, greater spectral detail, and more efficient computation. While these models achieve strong objective performance, they share a fundamental limitation: trained with distance-based losses, they tend to produce over-smoothed reconstructions that lack the fine stochastic HF structure of natural audio. This motivates the shift to generative modeling frameworks, surveyed in the following section.

6 Generative Models

Generative models for BWE/SR model conditional distributions to synthesize diverse and natural HF content, providing a probabilistic alternative to deterministic mappings. This section categorizes the literature by modeling paradigm, including AR, VAE, GAN, diffusion, flow, and bridge models, with representative approaches summarized in Table 3.

6.1 Autoregressive (AR) Model

WaveNet (Van Den Oord et al., 2016) is a fully probabilistic AR model originally developed for audio generation tasks such as multi-speaker speech synthesis, text-to-speech (TTS), and musical audio modeling, where each sample is predicted from all previous samples. This framework is then extended to BWE/SR (Wang et al., 2018; Gupta et al., 2019) by adopting a conditional WaveNet to model the conditional distribution of a BB/HR waveform $\mathbf{y} = [y[1], \ldots, y[T]]^\top$ given a BL/LR waveform $\mathbf{x}$. The formulation is given as

$$p(\mathbf{y} \mid h(\mathbf{x})) = \prod_{t=1}^{T} p\left(y[t] \mid y[1], \ldots, y[t-1], h(\mathbf{x})\right), \tag{11}$$

where the conditioning term $h(\mathbf{x})$ can be either a raw waveform $h(\mathbf{x}) = \mathbf{x}$ or a log-mel spectrogram

$$h(\mathbf{x}) = \log\left(\mathrm{MelFilter}\left(|\mathrm{STFT}(\mathbf{x})|\right)\right). \tag{12}$$

The log-mel spectrogram conditioning has been shown to outperform raw waveform conditioning (Wang et al., 2018; Gupta et al., 2019) and is widely recognized as an effective intermediate representation for high-fidelity audio generation (Shen et al., 2018).

WaveNet is built on causal convolutions that enforce temporal causality by conditioning each sample $y[t]$ only on past samples, while enabling parallel computation during training and efficient modeling of long sequences. To expand temporal context without excessively deep stacks or large kernels, WaveNet employs dilated causal convolutions, where the dilation factor increases exponentially across layers, resulting in exponential receptive-field expansion with minimal parameter growth. Fig. 5 depicts dilated causal convolutions for dilation = 1, 2, 4, and 8. Each layer uses gated activation units (Van den Oord et al., 2016), which outperform standard ReLU activations in audio modeling. The output distribution is modeled via a categorical softmax over μ-law-companded samples with 256-level quantization (Recommendation, 1988), and it is trained using the CE loss. Residual and skip connections further stabilize optimization in deep architectures.

Table 3: **Taxonomy of generative approaches for BWE/SR** according to their modeling paradigms.

Method	Literature
AR	DCNN (Gu & Ling, 2017), HRNN (Ling et al., 2018), Parallel-WaveNet (Wang et al., 2018), WaveNet (Gupta et al., 2019)
VAE	CVAE-BWE (Bachhav et al., 2020), FDFE-ASR (Tian et al., 2024)
GAN	Li et al. (2018), Eskimez & Koishida (2019), Li et al. (2019), CycleGAN (Haws & Cui, 2019), Sautter et al. (2019), MU-GAN (Kim & Sathe, 2019), MfNet (Hao et al., 2020), Su et al. (2021), Stream-SEANet (Li et al., 2021), BEHM-GAN (Moliner & Välimäki, 2022), NVSR (Liu et al., 2022a), VoiceFixer (Liu et al., 2022b), EBEN (Hauret et al., 2023), Dual-CycleGAN (Yoneyama et al., 2023), HiFi++ (Andreev et al., 2023), BBWEXNet (Gómez et al., 2023), mdctGAN (Shuai et al., 2023), MLPStreams (Soltanmohammadi et al., 2023), AERO (Mandel et al., 2023), Li et al. (2023b), MS-BWE (Lu et al., 2024b), Fre-Painter (Kim et al., 2024b), AP-BWE (Lu et al., 2024a), SDNet (Yang et al., 2024), BAE-Net (Yu et al., 2024), ICCRN-EBEN (Bao & Zhang, 2025) SwinSRGAN (Yuan et al., 2025), NLDSI-BWE (Tamiti & Barua, 2025), SFNet (Dai et al., 2025), Wave-U-Mamba (Lee & Kim, 2025), VM-ASR (Zhang & Ruan, 2025), HiFi-SR (Zhao et al., 2025), HWB-Net (Liu et al., 2025b), HCGAN (Gao et al., 2026b), SEMamba++ (Lee & Choi, 2026), EDNet (Kwak et al., 2026), Gao et al. (2026a),
Diffusion	NU-Wave (Lee & Han, 2021), UNIVERSE (Serrà et al., 2022), NU-Wave2 (Han & Lee, 2022), Lemercier et al. (2023), UDM+ (Yu et al., 2023), UNIVERSE++ (Scheibler et al., 2024), SWiBE (Lin et al., 2024), AudioSR (Liu et al., 2024a), BABE (Moliner et al., 2024), FlashSR (Im & Nam, 2025), VQ-Diffusion (Fang et al., 2025), IS-Diffusion (Jin et al., 2026), FastWave (Kuznetsov & Kaledin, 2026)
Flow	WSRGlow (Zhang et al., 2021), GSFM (Ku et al., 2025), Stream FM (Welker et al., 2025), FLowHigh (Yun et al., 2025), UniverSR (Choi et al., 2026), SAGA-SR (Im & Nam, 2026), CodecFlow (Zhang et al., 2026), Real-time Flow (Hsieh & Braun, 2026), DSC-FLowHigh (Hernandez-Olivan et al., 2026)
Bridge	A2SB (Shih et al., 2025), Bridge-SR (Li et al., 2025a), AudioLBM (Li et al., 2025b)

In (Gupta et al., 2019), the conditioning log-mel spectrogram is processed to match the temporal resolution of the waveform and injected into WaveNet layers to guide generation. While WaveNet produces high-quality audio, its sample-by-sample generation leads to slow inference. To address this limitation, Wang et al. (2018) propose a parallel WaveNet trained via teacher-student distillation, where a non-AR student model approximates the AR teacher. By leveraging inverse AR flows, the student enables parallel generation and significantly improves inference speed. The training objective combines a KL divergence term with a spectral-domain power loss to preserve perceptual quality.

Following the WaveNet framework, Gu & Ling (2017) investigate both causal and non-causal dilated convolutional architectures for waveform-level BWE. The causal design supports real-time processing, while the non-causal variant, inspired by ByteNet (Kalchbrenner et al., 2016), exploits future context for offline scenarios. Unlike the original WaveNet, which performs time-shifted next-sample prediction, both architectures directly generate the BB waveform. For the causal configuration, the conditional distribution is modeled as

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} p\left(y[t] \mid x[t-N+1], \ldots, x[t]\right), \tag{13}$$

where $\mathbf{x} = [x[1], x[2], \ldots, x[T]]^\top$ is the input BL sequence, $\mathbf{y} = [y[1], y[2], \ldots, y[T]]^\top$ is the output BB sequence, and N is the length of the receptive field. For the non-causal version, the distribution becomes

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} p\left(y[t] \mid x[t-N/2], \ldots, x[t+N/2]\right), \tag{14}$$

where $N+1$ is the length of the corresponding receptive field. As in WaveNet, both waveforms are μ-law quantized and the model is trained using the CE loss.

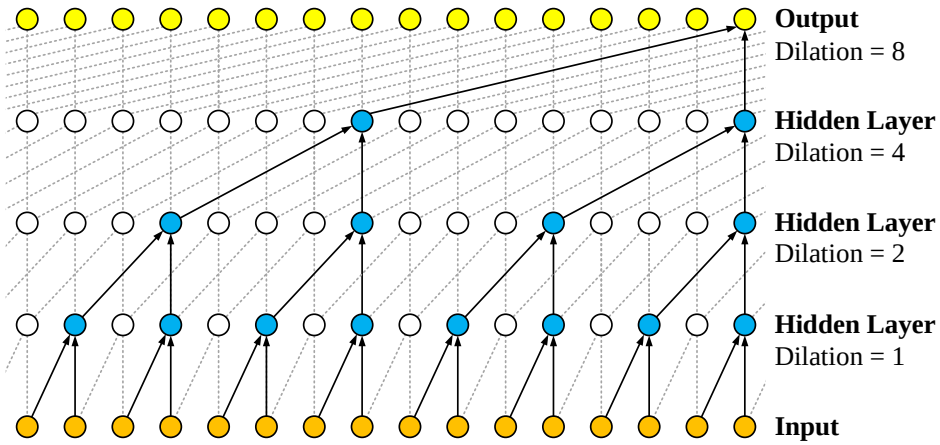


Figure 5: **Visualization of a stack of dilated causal convolutional layers.** Dilated causal convolutions with dilation factors of 1, 2, 4, and 8 are shown, where dilation specifies the spacing between consecutive filter taps, allowing the temporal receptive field to grow exponentially while preserving causality.

Following (Gu & Ling, 2017), Ling et al. (2018) propose a BWE model based on a hierarchical RNN (HRNN) architecture. Instead of stacked dilated convolutions, HRNNs employ multiple recurrent layers operating at different temporal resolutions, enabling efficient modeling of long-range dependencies. The model directly predicts BB waveforms conditioned on BL input, and can incorporate additional conditioning, such as bottleneck features from a DNN-based state classifier, to further improve speech quality. Results show that HRNN-based modeling achieves better perceptual quality and faster generation than dilated CNNs.

6.2 Variational Autoencoder (VAE)

An unconditional variational autoencoder (VAE) (Doersch, 2016) models the generation of an input signal $\mathbf{x} \in \mathbb{R}^T$ through a latent variable $\mathbf{z} \in \mathbb{R}^D$. The generative process is formulated as

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \mid \mathbf{z})\, p_\theta(\mathbf{z})\, d\mathbf{z}, \tag{15}$$

where $p_\theta(\mathbf{x})$ denotes the marginal likelihood of the input signal, $p_\theta(\mathbf{x} \mid \mathbf{z})$ represents the conditional likelihood that maps the latent representation to the input, and $p_\theta(\mathbf{z})$ is the prior distribution over the latent variable, which is typically chosen as a standard Gaussian distribution to regularize the latent space. During training, an encoder network $q_\phi(\mathbf{z} \mid \mathbf{x})$ approximates the intractable posterior over latent variables, while a decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ reconstructs the input from the latent representation. The model parameters θ and ϕ are learned by maximizing the evidence lower bound (ELBO) function $\mathcal{L}(\theta, \phi; \mathbf{x})$:

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta(\mathbf{z}))\,, \tag{16}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the KL divergence. The first term of $\mathcal{L}(\theta, \phi; \mathbf{x})$ encourages accurate reconstruction of the input signal, while the second term regularizes the latent distribution to match the prior. Sampling of the latent variable is enabled by the reparameterization trick

$$\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}, \tag{17}$$

where $\boldsymbol{\mu}_\phi \in \mathbb{R}^D$ and $\boldsymbol{\sigma}_\phi \in \mathbb{R}^D$ denote the mean and standard deviation predicted by the encoder network, respectively, $\boldsymbol{\epsilon} = [\epsilon^{(1)}, \ldots, \epsilon^{(D)}]^\top \in \mathbb{R}^D$ is an auxiliary noise vector with independent standard normal entries

$$\epsilon^{(d)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad d = 1, \ldots, D, \tag{18}$$

and $\odot$ denotes element-wise multiplication. Unlike AR models which generate samples sequentially, VAEs reconstruct the entire signal $\mathbf{x}$ in a single forward pass, enabling efficient and fully parallel inference. Fig. 6 depicts the architecture of an unconditional VAE.

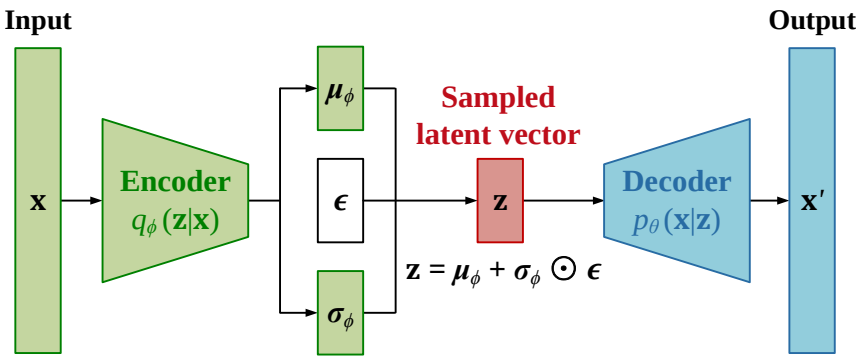


Figure 6: **Architecture of an unconditional VAE,** where an encoder infers the latent distribution $\mathbf{z}$ parameterized by mean $\boldsymbol{\mu}_\phi$ and variance $\boldsymbol{\sigma}_\phi$, and a decoder reconstructs the input signal $\mathbf{x}$ via latent sampling using the reparameterization trick.

Tian et al. (2024) propose a time-frequency fusion-based SR framework that jointly leverages waveform and spectrogram representations. Their approach combines a VAE-based time-domain module, a U-Net-based spectrogram module, and an attention-based fusion mechanism to integrate complementary information from both domains. This design preserves both temporal and spectral characteristics and demonstrates strong robustness across acoustic domains, including underwater scenarios. Bachhav et al. (2020) propose a conditional VAE (CVAE)-based BWE method that models the conditional distribution of BB speech given the BL input features. By incorporating adversarial learning into the CVAE framework, the approach improves both perceptual speech quality and downstream speech recognition performance.

6.3 Generative Adversarial Network (GAN)

GANs constitute one of the most prominent families of generative models, demonstrating substantial success across image domains, including text-to-image synthesis (Reed et al., 2016; Zhu et al., 2019; Kang et al., 2023), image SR (Ledig et al., 2017; Bulat et al., 2018; Mahapatra et al., 2019), and image style transfer (Azadi et al., 2018; Chen et al., 2018; Yang et al., 2019a). Motivated by these advances, GANs have been increasingly adopted in audio processing applications, including SE (Donahue et al., 2018a; Soni et al., 2018; Cao et al., 2022), audio synthesis (Donahue et al., 2018b; Engel et al., 2019; Morrison et al., 2021), and BWE/SR (Kim & Sathe, 2019; Su et al., 2021; Liu et al., 2022a; Hauret et al., 2023; Lu et al., 2024a; Lee & Kim, 2025). Within the BWE/SR paradigm, a generator network G reconstructs the BB/HR feature representation or waveform from the corresponding BL/LR input $\mathbf{x}$, while a discriminator network D is trained to distinguish between the generated output $\hat{\mathbf{y}}$ and BB/HR target $\mathbf{y}$, thereby establishing a min-max adversarial optimization framework. This training mechanism encourages the generator to produce BB/HR signals that preserve natural harmonic structures and exhibit spectral characteristics consistent with authentic audio.

GAN-based BWE/SR approaches can be systematically categorized along three principal design dimensions: (i) generator specification, (ii) discriminator configuration, and (iii) training objective formulation. The generator specification addresses whether the generator operates as a standalone model or as a submodule within a composite framework, as well as the processing domain (e.g., waveform-based or spectrogram-based) and the neural architecture (e.g., MLP, RNN, CNN, Transformer, or Mamba). Table 4 provides a taxonomy of GAN-based BWE/SR frameworks according to the generator's processing domain and architectures. The discriminator configuration specifies whether a single discriminator or multiple discriminators are employed; the former evaluates audio realism independently, whereas the latter operate in parallel to jointly assess perceptual fidelity across multiple representations. Finally, the training objective formulation specifies the loss function design. Collectively, these design factors govern the perceptual quality and spectral fidelity of the reconstructed signals, as well as the system robustness.

Table 4: **Taxonomy of GAN-based BWE/SR approaches** according to their generators' architectural design and processing domain.

Generator	Waveform-based Network	Spectrogram-based Network
MLP	-	Li et al. (2018), Sautter et al. (2019), MLPStreams (Soltanmohammadi et al., 2023)
RNN	HWB-Net (Liu et al., 2025b)	CycleGAN (Haws & Cui, 2019), BAE-Net (Yu et al., 2024), SFNet (Dai et al., 2025), ICCRN-EBEN (Bao & Zhang, 2025)
CNN	Li et al. (2019), MU-GAN (Kim & Sathe, 2019), MfNet (Hao et al., 2020), Su et al. (2021), Stream-SEANet (Li et al., 2021), EBEN (Hauret et al., 2023), BBWEXNet (Gómez et al., 2023), Gao et al. (2026a)	Eskimez & Koishida (2019), CycleGAN (Haws & Cui, 2019), BEHM-GAN (Moliner & Välimäki, 2022), NVSR (Liu et al., 2022a), VoiceFixer (Liu et al., 2022b), Li et al. (2023b), Dual-CycleGAN (Yoneyama et al., 2023), HiFi++ (Andreev et al., 2023), AP-BWE (Lu et al., 2024a), Fre-Painter (Kim et al., 2024b), MS-BWE (Lu et al., 2024b), SDNet (Yang et al., 2024), NLDSI-BWE (Tamiti & Barua, 2025),
Transformer	-	AERO (Mandel et al., 2023), mdctGAN (Shuai et al., 2023), SwinSRGAN (Yuan et al., 2025), HiFi-SR (Zhao et al., 2025)
Mamba	Wave-U-Mamba (Lee & Kim, 2025)	VM-ASR (Zhang & Ruan, 2025), EDNet (Kwak et al., 2026), SEMamba++ (Lee & Choi, 2026), HCGAN (Gao et al., 2026b)

6.3.1 Generator

Based on the signal representation, GAN-based approaches can be broadly categorized into waveform-based networks which operate directly on raw waveforms and spectrogram-based networks which operate on time-frequency representations. A special instance within waveform-based methods is EBEN (Hauret et al., 2023), which operates on PQMF-decomposed subband signals: the lowest subband is used as input to a U-Net generator to predict all subbands, followed by PQMF synthesis to reconstruct the BB waveform. For spectrogram-based approaches, a variety of representations have been explored, including LSF (Li et al., 2018), power or STFT magnitude spectrograms (Li et al., 2018; Eskimez & Koishida, 2019), STFT complex spectrograms (Moliner & Välimäki, 2022; Mandel et al., 2023; Soltanmohammadi et al., 2023; Yang et al., 2024; Lu et al., 2024a;b; Yu et al., 2024; Zhang & Ruan, 2025; Dai et al., 2025; Tamiti & Barua, 2025; Gao et al., 2026b), mel-spectrograms (Haws & Cui, 2019; Liu et al., 2022a;b; Li et al., 2023b; Yoneyama et al., 2023; Andreev et al., 2023; Kim et al., 2024b; Zhao et al., 2025; Lee & Choi, 2026; Kwak et al., 2026), MDCT spectrograms (Shuai et al., 2023; Yuan et al., 2025), and cepstral-domain features (Bao & Zhang, 2025).

Architecturally, most GAN frameworks employ a generator as a standalone model for signal reconstruction. In contrast, several works (Liu et al., 2022a;b; Li et al., 2023b; Andreev et al., 2023; Kim et al., 2024b; Zhao et al., 2025) incorporate a pretrained generator—such as TFGAN (Tian et al., 2020) or HiFi-GAN (Kong et al., 2020)—as a submodule within a larger architecture. In the first category, an earlier study (Li et al., 2018) employs a simple four-layer MLP generator, while subsequent approaches evolve toward recurrent architectures and convolutional designs, including autoencoder and U-Net architectures. More recent approaches adopt other high-capacity convolutional backbones, such as ConvNeXt (Lu et al., 2024a;b; Tamiti & Barua, 2025) and Deep Complex Networks (Moliner & Välimäki, 2022). In addition, Transformer-based U-Net generators leverage self-attention to capture global temporal-spectral dependencies in the spectrogram domain. Specifically, mdctGAN (Shuai et al., 2023) introduces a Transformer bottleneck stack that performs global attention on frequency-consistent features, AERO (Mandel et al., 2023) incorporates a Frequency Transformer Block before each encoder stage, and SwinSRGAN (Yuan et al., 2025) adopts multiple Residual Swin Transformer Blocks in the bottleneck. More recently, Mamba-based generators propose state-space sequence modeling for efficient long-context reconstruction. Wave-U-Mamba (Lee & Kim, 2025) inserts Mamba blocks throughout a waveform U-Net to efficiently model long-range temporal structure, while VM-ASR (Zhang & Ruan, 2025) applies VMamba-style Visual State Space blocks (Liu et al., 2024b)

within a dual-stream complex-spectrogram U-Net to better capture global-local spectro-temporal features and improve harmonic reconstruction. EDNet (Kwak et al., 2026) adopts a dual-stream architecture where Gating Mamba modules integrate magnitude masking and mapping to guide phase reconstruction through long-range temporal–spectral modeling. HCGAN (Gao et al., 2026b) employs a dual-branch generator consisting of a U-Net-based spectral subgenerator with a Mamba bottleneck and a dedicated harmonic branch that utilizes temporal-harmonic modules to estimate HF structures for final spectral-harmonic fusion. SE-Mamba++ (Lee & Choi, 2026) adopts an encoder-decoder architecture, where the bottleneck employs multi-resolution time-frequency dual-processing (TFDP) blocks with Time Mamba layers for temporal modeling and a Fourier-based frequency module to capture global, local, and periodic spectral patterns.

In the second category, a complementary design incorporates a pretrained vocoder generator as a submodule within the overall architecture. NVSR (Liu et al., 2022a) and VoiceFixer (Liu et al., 2022b) adopt a two-stage SSR pipeline, in which a ResUNet-based analysis network (Kong et al., 2021a) first predicts the HR mel-spectrogram from its LR counterpart, followed by a pretrained TFGAN vocoder (Tian et al., 2020) that synthesizes the HR waveform. As errors may propagate across stages, NVSR (Liu et al., 2022a) introduces a LF replacement step to preserve LB fidelity. Li et al. (2023b) propose another two-stage framework in which the first stage employs a U-Net-based generator with adversarial training to perform BWE on bone-conducted speech in the log-mel spectrogram domain. In the second stage, a DiffWave vocoder (Kong et al., 2021b) is trained separately to reconstruct high-fidelity waveforms from the extended spectral features. HiFi++ (Andreev et al., 2023) further employs a staged generator architecture: SpectralUNet enhances the mel-spectrogram, a HiFi-GAN backbone (Kong et al., 2020) generates the waveform, and WaveUNet and SpectralMaskNet subsequently refine it in the time and frequency domains, respectively. Fre-Painter (Kim et al., 2024b) integrates a pretrained masked-autoencoder encoder (Huang et al., 2022) with a HiFi-GAN vocoder (Kong et al., 2020), leveraging masking-based training to enhance robustness across varying sampling rates. HiFi-SR (Zhao et al., 2025) employs a Transformer-convolutional generator that encodes LR mel-spectrograms via a hybrid MossFormer-recurrent network (Zhao et al., 2024) to capture long-range dependencies, followed by HiFi-GAN-based waveform synthesis with joint optimization.

6.3.2 Discriminator

Early GANs (Li et al., 2018; Eskimez & Koishida, 2019; Li et al., 2019; Sautter et al., 2019; Hao et al., 2020) typically employ a single discriminator operating on either spectral features or raw waveforms to evaluate signal realism and perceptual fidelity. While MLP-based discriminators (Li et al., 2018) operate on spectral parameters such as LSF and energy, CNN-based architectures leverage a variety of input representations, including full-band LPS (Eskimez & Koishida, 2019), MFCC (Li et al., 2019), mel-spectrograms (Haws & Cui, 2019), and raw waveforms (Kim & Sathe, 2019; Hao et al., 2020). Subsequent works adopt multi-discriminator architectures to provide complementary adversarial guidance from multiple perceptual perspectives, enabling joint modeling of global structure and fine-grained temporal details. The majority of works (Li et al., 2021; Su et al., 2021; Moliner & Välimäki, 2022; Andreev et al., 2023; Gómez et al., 2023; Hauret et al., 2023; Mandel et al., 2023; Shuai et al., 2023; Soltanmohammadi et al., 2023; Yang et al., 2024; Yu et al., 2024; Kim et al., 2024b; Lu et al., 2024a;b; Lee & Kim, 2025; Yuan et al., 2025; Zhao et al., 2025; Zhang & Ruan, 2025; Gao et al., 2026a; Lee & Choi, 2026) employ either the multi-scale discriminator (MSD) from MelGAN (Kumar et al., 2019a) or the multi-period discriminator (MPD) from HiFi-GAN (Kong et al., 2020). MSD evaluates waveforms at multiple temporal resolutions (full, $2\times$, and $4\times$ downsampled), allowing lower-resolution branches to capture LF structures such as coarse spectral envelopes and prosodic patterns, while the full-resolution branch preserves HF details and transient characteristics. In contrast, MPD analyzes periodically sampled waveform segments via 2-D convolutions, explicitly modeling pitch-dependent harmonic regularity and periodic structure that conventional time-domain discriminators often fail to capture.

Alternative designs (Su et al., 2021; Yoneyama et al., 2023; Hauret et al., 2023; Lu et al., 2024a;b; Zhao et al., 2025; Yuan et al., 2025; Tamiti & Barua, 2025; Lee & Choi, 2026) further diversify discriminator architectures. Su et al. (2021) introduce a spectral-domain discriminator operating on log-mel spectrograms to enhance HF reconstruction. Yoneyama et al. (2023) combine waveform-level (Yamamoto et al., 2020) and multi-band spectral discriminators (Kumar et al., 2020) for independent frequency-band analysis. EBEN (Hauret et al., 2023) employs a multi-scale ensemble discriminator on PQMF-decomposed subbands

with a full-scale MSD sub-discriminator to enforce cross-band coherence. Lu et al. (2024a;b) introduce a multi-resolution amplitude discriminator (MRAD) and a multi-resolution phase discriminator (MRPD), inspired by the multi-resolution discriminator (MRD) (Jang et al., 2021), operating across multiple STFT resolutions to alleviate spectral over-smoothing and phase distortion artifacts. HiFi-SR (Zhao et al., 2025) addresses the limited phase sensitivity of MRD via a multi-band, multi-scale time-frequency discriminator (MBD) that jointly analyzes complex STFT components across time scales and frequency subbands. Yuan et al. (2025) design a high-band multi-band discriminator (HB-MBD) that focuses on reconstructing HF content in the MDCT domain. Tamiti & Barua (2025) introduce a multi-scale recurrence discriminator (MSRD) and a multi-resolution Lyapunov discriminator (MRLD), which impose nonlinear dynamical-system constraints to enable lightweight, depthwise-separable discriminator designs. In addition, SEMamba++ (Lee & Choi, 2026) employs a multi-scale sub-band constant-Q transform (MS-SB-CQT) discriminator (Gu et al., 2024) to better capture periodic and harmonic structures.

6.3.3 Training Objective

GAN-based BWE/SR methods typically optimize a generator G that predicts a BB/HR estimate $\hat{\mathbf{y}} = G(\mathbf{x})$ from a BL/LR observation $\mathbf{x}$, together with discriminator(s) D that distinguish $\hat{\mathbf{y}}$ from the ground truth $\mathbf{y}$. In practice, the generator objective is commonly formulated as a weighted combination of (i) an adversarial term that promotes perceptual realism, (ii) a reconstruction term that enforces signal fidelity, and (iii) an feature matching (FM) term that stabilizes adversarial optimization. The standard adversarial loss is defined as the min-max optimization by

$$\min_{G} \max_{D} \quad \mathbb{E}_{\mathbf{y}}[\log D(\mathbf{y})] + \mathbb{E}_{\mathbf{x}}[\log(1 - D(G(\mathbf{x})))] \,, \tag{19}$$

where $D(\mathbf{y})$ estimates the likelihood of the ground-truth BB/HR signal, and $D(G(\mathbf{x}))$ measures the perceptual plausibility of the generated output. Through this dual-objective optimization, GAN-based approaches recover sharper spectral details and more realistic HF contents compared to deterministic regression models.

The adversarial term drives distribution matching and perceptual realism, using binary CE loss (Li et al., 2018; Eskimez & Koishida, 2019; Li et al., 2019; Kim & Sathe, 2019; Sautter et al., 2019; Haws & Cui, 2019; Hao et al., 2020; Shuai et al., 2023; Yoneyama et al., 2023; Gao et al., 2026b), hinge loss (Li et al., 2021; Hauret et al., 2023; Mandel et al., 2023; Soltanmohammadi et al., 2023; Gómez et al., 2023; Yang et al., 2024; Lu et al., 2024a;b; Bao & Zhang, 2025), or least-squares GAN (LSGAN) loss (Andreev et al., 2023; Kim et al., 2024b; Yu et al., 2024; Zhao et al., 2025; Yuan et al., 2025; Zhang & Ruan, 2025; Dai et al., 2025; Liu et al., 2025b; Gao et al., 2026a; Lee & Choi, 2026).

The binary CE loss is expressed as

$$\mathcal{L}_G^{\text{CE}} = -\mathbb{E}_{\hat{\mathbf{y}}}[\log D(\hat{\mathbf{y}})], \tag{20}$$

$$\mathcal{L}_D^{\text{CE}} = -\mathbb{E}_{\mathbf{y}}[\log D(\mathbf{y})] - \mathbb{E}_{\hat{\mathbf{y}}}[\log(1 - D(\hat{\mathbf{y}}))]. \tag{21}$$

The hinge loss is formulated as

$$\mathcal{L}_G^{\text{hinge}} = -\mathbb{E}_{\hat{\mathbf{y}}}[D(\hat{\mathbf{y}})], \tag{22}$$

$$\mathcal{L}_D^{\text{hinge}} = \mathbb{E}_{\hat{\mathbf{y}}}[\max(0, 1 + D(\hat{\mathbf{y}}))] + \mathbb{E}_{\mathbf{y}}[\max(0, 1 - D(\mathbf{y}))]. \tag{23}$$

The LSGAN loss is defined as

$$\mathcal{L}_G^{\text{LS}} = \mathbb{E}_{\hat{\mathbf{y}}}[(D(\hat{\mathbf{y}}) - 1)^2], \tag{24}$$

$$\mathcal{L}_D^{\text{LS}} = \mathbb{E}_{\mathbf{y}}[(D(\mathbf{y}) - 1)^2] + \mathbb{E}_{\hat{\mathbf{y}}}[D(\hat{\mathbf{y}})^2]. \tag{25}$$

Early GAN-based designs often adopt the CE loss formulation, whereas later works favor hinge loss or LSGAN objectives for improved gradient behavior and training stability.

To enforce fidelity to the target signal and preserve spectral structure, GAN-based BWE/SR methods augment adversarial training with time-domain and/or frequency-domain reconstruction losses. A widely adopted choice is the multi-resolution STFT (MR-STFT) loss (Yamamoto et al., 2020), which combines

spectral convergence and log-magnitude terms across multiple STFT configurations (e.g., varying FFT sizes, window lengths, and hop sizes):

$$\mathcal{L}_G^{\text{STFT}} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\mathbf{y},\hat{\mathbf{y}}} \left[\mathcal{L}_{\text{sc}}^m(\mathbf{y}, \hat{\mathbf{y}}) + \mathcal{L}_{\text{mag}}^m(\mathbf{y}, \hat{\mathbf{y}}) \right], \tag{26}$$

where M is the number of STFT configurations, G denotes the generator. The spectral convergence loss $\mathcal{L}_{\text{sc}}$ and log-magnitude loss $\mathcal{L}_{\text{mag}}$ are defined as

$$\mathcal{L}_{\text{sc}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\| \, |\text{STFT}(\mathbf{y})| - |\text{STFT}(\hat{\mathbf{y}})| \, \|_F}{\| \, |\text{STFT}(\mathbf{y})| \, \|_F}, \tag{27}$$

$$\mathcal{L}_{\text{mag}}(\mathbf{y}, \hat{\mathbf{y}}) = \| \log |\text{STFT}(\mathbf{y})| - \log |\text{STFT}(\hat{\mathbf{y}})| \|_1, \tag{28}$$

where $|\text{STFT}(\cdot)|$ denotes the magnitude of STFT, and $\| \cdot \|_F$ and $\| \cdot \|_1$ denote the Frobenius norm and l_1 norm, respectively.

Another widely adopted objective is the mel-spectrogram MAE loss, which evaluates reconstruction accuracy in a perceptually weighted mel-frequency domain. Its formulation is given by

$$\mathcal{L}_G^{\text{mel}} = \| \text{Mel}(\mathbf{y}) - \text{Mel}(\hat{\mathbf{y}}) \|_1, \tag{29}$$

where $\text{Mel}(\cdot)$ denotes the mel-spectrogram transformation. Both objectives act as perceptually motivated reconstruction losses, reflecting the nonlinear frequency sensitivity of human hearing and encouraging perceptually plausible spectral detail in the reconstructed signals.

To stabilize adversarial optimization and suppress unnatural artifacts, the FM loss aligns intermediate discriminator features between real and generated samples:

$$\mathcal{L}_G^{\text{FM}} = \mathbb{E}_{\mathbf{y},\hat{\mathbf{y}}} \left[\sum_{i=1}^{K} \frac{1}{N_i} \| D_i(\mathbf{y}) - D_i(\hat{\mathbf{y}}) \|_1 \right], \tag{30}$$

where K denotes the number of discriminator layers, $D_i(\cdot)$ represents the intermediate feature map at the i-th layer, and N_i denotes the number of elements in that feature map. These components are typically combined into a weighted generator objective:

$$\mathcal{L}_G = \lambda_{\text{adv}} \mathcal{L}_G^{\text{adv}} + \lambda_{\text{rec}} \mathcal{L}_G^{\text{rec}} + \lambda_{\text{FM}} \mathcal{L}_G^{\text{FM}}, \tag{31}$$

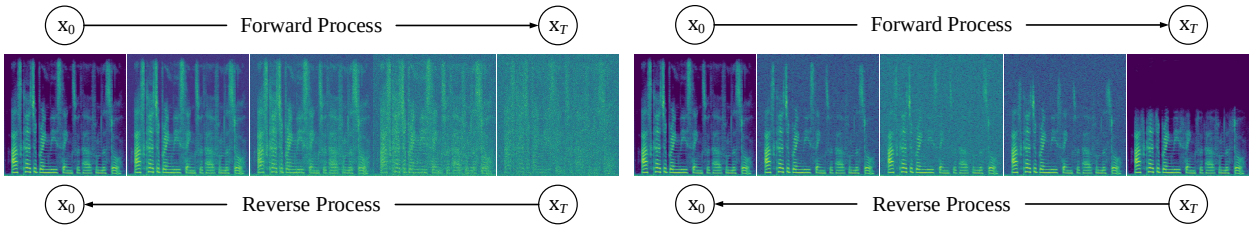
where $\mathcal{L}_G^{\text{adv}}$, $\mathcal{L}_G^{\text{rec}}$, and $\mathcal{L}_G^{\text{FM}}$ denote the adversarial, reconstruction, and FM losses, respectively, and $\lambda_{\text{adv}}, \lambda_{\text{rec}}, \lambda_{\text{FM}}$ are weighting coefficients that balance perceptual naturalness and spectral fidelity. Simultaneously, the discriminator is optimized to minimize $\mathcal{L}_D$. Moreover, some works incorporate task-aware perceptual constraints, such as matching embeddings from a pretrained automatic speech recognition network (Li et al., 2019), to enhance downstream performance.

6.4 Diffusion Model

Diffusion models (Ho et al., 2020; Rombach et al., 2022; Ho & Salimans, 2022; Gao et al., 2023; Li et al., 2025c; Lu et al., 2025), also known as diffusion-based generative models or score-based generative models (SGMs), constitute a class of generative models that synthesize data by learning to invert a progressively applied Gaussian noising process. A diffusion model comprises two coupled stochastic processes: a forward diffusion process and a reverse sampling process. In the forward process, progressively increasing Gaussian noise is added to a clean signal $\mathbf{x}_0$ according to

$$\mathbf{x}_t = \sqrt{\alpha_t} \, \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \, \boldsymbol{\epsilon}, \tag{32}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad t = 1, \ldots, T, \tag{33}$$



(a) Diffusion (b) Bridge

Figure 7: **Illustration of diffusion and bridge processes on audio spectrograms. (a)** Diffusion: the forward process progressively corrupts the clean audio spectrogram $\mathbf{x}_0$ by injecting noise, whereas the learned reverse process iteratively denoises a heavily perturbed sample $\mathbf{x}_T$ to recover $\mathbf{x}_0$. **(b)** Bridge: the forward bridge process degrades the HR spectrogram $\mathbf{x}_0$ into a LR spectrogram $\mathbf{x}_T$ through progressive bandwidth reduction with stochastic perturbations, while the reverse bridge process reconstructs $\mathbf{x}_0$ from $\mathbf{x}_T$.

where $\mathbf{x}_t$ denotes the intermediate signal at the diffusion step t, $\alpha_t = 1-\beta_t$, and $\{\beta_t\}_{t=1}^{T}$ denotes a predefined noise variance schedule. This forward process progressively transforms the data distribution into an isotropic Gaussian distribution. The reverse process is parameterized by a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ that predicts the injected noise, or equivalently the denoising direction, at each diffusion step t, and generates samples via iterative denoising according to

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}, \tag{34}$$

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad t = T, \ldots, 1, \tag{35}$$

with $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t}$. By learning to estimate the denoising direction at each step accurately, the reverse process enables the generation of samples following the underlying training data distribution. The forward and reverse diffusion processes are illustrated in Fig. 7 (a). Due to their training stability, strong mode coverage, and high perceptual fidelity, diffusion models have emerged as state-of-the-art approaches across a wide range of audio generation and restoration tasks.

NU-Wave (Lee & Han, 2021) introduces waveform-to-waveform conditional diffusion for SSR, adopting a vocoder-inspired residual bidirectional dilated convolution (Bi-DilConv) backbone from DiffWave (Kong et al., 2021b) and injecting the LR waveform via a redesigned conditioner with an enlarged receptive field to enable accurate upsampling from 16 kHz or 24 kHz to 48 kHz. NU-Wave2 (Han & Lee, 2022) further improves harmonic reconstruction by introducing short-time Fourier convolution (STFC) to provide stronger global spectral context and bandwidth spectral feature transform (BSFT) to explicitly condition on the retained bandwidth, enabling a single model to handle diverse input sampling rates with fewer parameters than NU-Wave (Lee & Han, 2021). FastWave (Kuznetsov & Kaledin, 2026) further lightens NU-Wave2 (Han & Lee, 2022) with architectural modifications to reduce the computational complexity for any-to-48 kHz SR.

To mitigate computational costs of waveform-level diffusion, AudioSR (Liu et al., 2024a) performs SR in a compact latent space obtained via a VAE encoder. Inspired by AudioLDM (Liu et al., 2023), the latent diffusion model (LDM) predicts HR mel-spectrograms from their LR counterparts, followed by waveform synthesis using a HiFi-GAN vocoder (Kong et al., 2020). Building on this framework, FlashSR (Im & Nam, 2025) distills the multi-step LDM into a single-step model via Flash Diffusion (Chadebec et al., 2025) and distribution-matching objectives, enabling efficient sampling while preserving perceptual quality. It further integrates a BigVGAN-based vocoder (Lee et al., 2022), jointly conditioned on the predicted mel-spectrogram and input LR waveform, eliminating explicit LF refinement and enhancing HF reconstruction. Jin et al. (2026) propose an inference-time scaling diffusion (IS-Diffusion) SR framework built on AudioSR (Liu et al., 2024a), which generates multiple HR candidates and leverages task-specific verifiers with zero-order search to select optimal samples, thereby reducing stochastic variance and improving spectral fidelity.

UDM+ (Yu et al., 2023) reframes SSR as a frequency-domain inpainting problem and proposes a reverse-sampling strategy that propagates LR information through the sampling trajectory, rather than relying

solely on auxiliary conditioning of the denoiser. Combined with an unconditional diffusion model (UDM), this formulation improves robustness across varying upsampling ratios and downsampling filters. SWiBE (Lin et al., 2024) targets noise-robust BWE by parameterizing a stochastic diffusion trajectory that progressively expands bandwidth in the complex spectrogram domain, using a U-Net-based SGM (NCSN++ (Song et al., 2020)) to estimate score functions along the trajectory. Vector Quantized Diffusion (VQ-Diffusion) (Fang et al., 2025) performs BWE directly in the discrete token space derived from a NAC, mitigating HF loss associated with mel-spectrogram front-ends. It combines Descript Audio Codec (DAC)-based tokenization, a discrete denoising diffusion model (D3PM) for token reconstruction, and a Mamba-2 backbone to capture long-range temporal dependencies for high-fidelity BWE.

Lemercier et al. (2023) extend diffusion-based generative modeling to speech restoration in the complex time-frequency domain, comparing it with a discriminative counterpart using the same DNN architecture. This method employs iterative denoising to capture complex degradations and shows consistent improvements over discriminative models, particularly for BWE and dereverberation. UNIVERSE (Serrà et al., 2022) proposes a unified generative framework for universal SE by decoupling restoration into a multi-resolution conditioning network and a score-based diffusion generator, where mixture density networks (MDNs) provide robust spectral guidance. Building on this, UNIVERSE++ (Scheibler et al., 2024) enhances stability and fidelity via anti-aliasing, HiFi-GAN-based adversarial loss, and low-rank adaptation (LoRA) (Hu et al., 2022) with phoneme-level constraints to reduce hallucinations and improve linguistic consistency.

Beyond supervised conditional modeling, diffusion priors can be leveraged for inverse problems with uncertain or unknown degradations. BABE (Moliner et al., 2024) addresses zero-shot blind BWE of historical recordings by combining a pretrained unconditional diffusion prior with generalized diffusion posterior sampling, jointly estimating a parametric low-pass degradation operator during iterative HF content reconstruction.

6.5 Flow-based Model

Flow-based generative models (Ho et al., 2019; Kumar et al., 2019b; Lipman et al., 2022; Tong et al., 2023; Gat et al., 2024) learn transport mappings that transform a simple source distribution into the conditional distribution of BB/HR audio given a BL/LR observation. In the current BWE/SR literature, this family can be broadly divided into two branches: (i) Glow-style discrete normalizing flows, which optimize exact conditional likelihood via a finite sequence of invertible transformations, and (ii) continuous normalizing flows trained with flow matching, which learn a continuous transport governed by an ordinary differential equation (ODE) by regressing a target velocity field (Lipman et al., 2023; Pooladian et al., 2023). Although both branches are flow-based, they occupy distinct design regimes. Glow-style models emphasize exact likelihood computation and one-pass invertibility. In contrast, flow matching provides greater flexibility in transport path design, conditioning mechanisms, and representation learning, but requires solving a continuous-time ODE during sampling. Their relationship and trade-offs are summarized in Table 5.

A Glow-style conditional normalizing flow defines an invertible mapping

$$\mathbf{y} = f_\theta(\mathbf{z} \mid \mathbf{c}), \tag{36}$$

where $\mathbf{y}$ denotes the target BB/HR audio, $\mathbf{z}$ is a latent variable drawn from a simple base distribution $p(\mathbf{z})$, which is typically chosen as $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{c}$ represents the conditioning information extracted from the BL/LR observation. The corresponding exact conditional likelihood is obtained via the change-of-variables formula given as

$$\log p_\theta(\mathbf{y} \mid \mathbf{c}) = \log p(\mathbf{z}) + \log \left| \det \frac{\partial f_\theta^{-1}(\mathbf{y} \mid \mathbf{c})}{\partial \mathbf{y}} \right|, \tag{37}$$

where $\det(\cdot)$ denotes the determinant of the Jacobian matrix associated with the inverse mapping. WSR-Glow (Zhang et al., 2021) serves as a representative Glow-style model for SR, integrating invertible flow-based transformations with WaveNet-style affine coupling to maximize the exact conditional likelihood of the HR waveforms given the LR observations. Its dual-conditioning design combines a time-domain LR encoder with an STFT-based encoder, enabling the fusion of complementary temporal and spectral cues for enhanced HF reconstruction.

Table 5: **Key distinctions between Glow-style discrete normalizing flows and flow matching.** Glow-style models use constrained invertible architectures for exact likelihoods, while flow matching adopts simulation-free vector-field learning for flexible probability paths with iterative inference.

Modeling	Mapping	Training	Inference
Glow-style Flow	Invertible Transformation	Exact Likelihood Maximization	Single Forward Pass
Flow Matching	Continuous-time ODE	Velocity Field Regression	Multi-step ODE

More recent work has largely shifted toward flow matching, which trains a continuous normalizing flow by learning a conditional velocity field along a prescribed probability path (Lipman et al., 2023; Pooladian et al., 2023; Yun et al., 2025; Ku et al., 2025). The transport dynamics are controlled by

$$\frac{d}{dt}\mathbf{y}_t = v_\theta(\mathbf{y}_t, t \mid \mathbf{c}), \tag{38}$$

where $\mathbf{y}_t \sim p_t(\cdot \mid \mathbf{c})$ denotes the intermediate state at time t, v_θ is the learned conditional velocity field, and $\mathbf{c}$ represents the condition derived from the degraded input. The process evolves from a source distribution $p_0(\cdot \mid \mathbf{c})$ to the target conditional HR audio distribution $p_1(\cdot \mid \mathbf{c})$ at $t = 1$. In contrast to Glow-style flows, which optimize exact likelihood, recent BWE/SR systems adopt flow matching to learn this transport via conditional velocity regression.

Within this branch, FLowHigh (Yun et al., 2025) demonstrates that carefully designed conditional paths can significantly improve the efficiency of flow matching for SSR. Specifically, it introduces a data-dependent source prior and a post-processing step that restores the observed LF band, enabling high-quality single-step sampling. Hernandez-Olivan et al. (2026) extend this paradigm to controllable music BWE by incorporating Dynamic Spectral Contours and classifier-free guidance, enabling finer control over the reconstructed bandwidth content. A complementary direction scales flow matching through enhanced representations and training strategies. Ku et al. (2025) pretrain a generative speech foundation model (GSFM) directly on complex-valued STFT coefficients, eliminating the need for a separate vocoder and supporting multiple restoration tasks, including BWE. Along a related vocoder-free direction, UniverSR (Choi et al., 2026) models the conditional distribution of complex spectral coefficients via flowmatching and reconstructs audio using iSTFT, yielding a unified SR framework without a dedicated neural vocoder.

Recent work further expands conditioning mechanisms and improves computational efficiency. SAGA-SR (Im & Nam, 2026) augments flow matching with semantic guidance from text and acoustic cues derived from spectral roll-off, employing a diffusion-Transformer backbone to enable any-rate upsampling to 44.1 kHz. CodecFlow (Zhang et al., 2026), in contrast, performs BWE in a neural codec latent space, where a voicing-aware conditional flow converter and a structure-constrained residual vector quantization (RVQ) module mitigate latent mismatch and improve the compute-quality trade-off. To facilitate real-time deployment, Stream FM (Welker et al., 2025) and Real-time Flow (Hsieh & Braun, 2026) adopt few-step sampling strategies within their flow matching formulations, effectively bridging the gap between generative quality and low-latency requirements.

Overall, flow-based BWE/SR has evolved from Glow-style exact-likelihood models toward flow matching, which favors more flexible conditional transport and reduced sampling complexity. Consequently, flow matching has become increasingly attractive for modern high-fidelity systems, where computational efficiency and representational flexibility are as critical as likelihood tractability.

6.6 Bridge Model

Bridge models, often described as Schrödinger bridges in recent audio machine learning literature (Chen et al., 2023; Wang et al., 2024a;b; Li et al., 2025a;b; Shih et al., 2025), are closely related to diffusion and score-based models but instead learn a stochastic transport between endpoint marginals. Diffusion models start from an approximately Gaussian endpoint induced by progressive noising, whereas bridge models directly

connect prescribed source and target distributions, as illustrated in Fig. 7. However, unlike the classical Schrödinger bridge (Léonard, 2013; De Bortoli et al., 2024)—an unsupervised optimal transport formulation solved via iterative proportional fitting on unpaired distributions—most current BWE/SR approaches (Li et al., 2025a;b; Shih et al., 2025) adopt supervised training with paired data and single-pass regression, making them more closely related to *bridge matching* (Shi et al., 2023; Zhou et al., 2023; Peluchetti, 2023). This formulation is particularly appealing for BWE/SR because the observed BL/LR signal itself provides an informative starting point for reconstructing the BB/HR target.

Let p_{src} and p_{tgt} denote the source and target distributions, and let $\{\mathbf{z}_t\}_{t\in[0,T]}$ denote the corresponding transport trajectory. For BWE/SR, p_{src} is induced by degraded BL/LR observations, while p_{tgt} corresponds to the BB/HR data distribution. Starting from a reference diffusion process,

$$d\mathbf{z}_t = f(\mathbf{z}_t, t)\, dt + g(t)\, d\mathbf{w}_t, \qquad t \in [0, T], \tag{39}$$

the Schrödinger bridge problem seeks a path measure $p \in \mathcal{P}[0, T]$ that matches the endpoint marginals while remaining close to a reference measure p_{ref}:

$$\min_{p\in\mathcal{P}[0,T]} D_{\text{KL}}(p \,\|\, p_{\text{ref}}) \quad \text{s.t.} \quad p_0 = p_{\text{src}},\ p_T = p_{\text{tgt}}. \tag{40}$$

This problem admits an equivalent characterization via paired forward and backward stochastic differential equations (SDEs) with corrected drifts:

$$d\mathbf{z}_t = \left[f(\mathbf{z}_t, t) + g^2(t)\nabla_{\mathbf{z}} \log \Psi_t(\mathbf{z}_t)\right] dt + g(t)\, d\mathbf{w}_t, \tag{41}$$

$$d\mathbf{z}_t = \left[f(\mathbf{z}_t, t) - g^2(t)\nabla_{\mathbf{z}} \log \hat{\Psi}_t(\mathbf{z}_t)\right] dt + g(t)\, d\bar{\mathbf{w}}_t, \tag{42}$$

where Ψ_t and $\hat{\Psi}_t$ are time-dependent Schrödinger potentials. Compared with diffusion models that start from an uninformative noise prior, bridge models exploit the informative source marginal, resulting in a shorter and more efficient generation trajectory.

Bridge-SR (Li et al., 2025a) is the first bridge-based framework for SR. Operating directly in the waveform domain, it uses the observed LR waveform as an informative prior for the HR target and learns the corresponding score functions for efficient data-to-data generation. The study further shows compelling quality-efficiency trade-offs with a lightweight backbone and few-step synthesis. AudioLBM (Li et al., 2025b) extends this idea from waveform-space transport to continuous latent-space transport. By compressing audio into a continuous latent representation and performing latent-to-latent bridging, it better aligns the generative process with LR-to-HR upsampling. It further introduces frequency-aware conditioning on the source and target bands to enable any-to-any upsampling during training, while cascaded latent bridge models and prior augmentation support SR beyond 48 kHz, including 96 and 192 kHz, across speech, music, and general sound effects. A2SB (Shih et al., 2025) formulates BWE and audio inpainting as a unified spectrogram inpainting problem within the Schrödinger bridge framework, enabling end-to-end waveform generation without relying on a separate vocoder. The model further incorporates a factorized magnitude-phase representation and sliding-window inference to support HR audio restoration at 44.1 kHz over long music segments.

7 Challenges and Overlook

7.1 Discussion

Discriminative DNN-based models have driven major progress in BWE/SR, yet their core limitation is structural: they implement a *deterministic* mapping from a BL/LR observation to a single BB/HR estimate. Since HF recovery is an inherently one-to-many problem, point-estimation training—often dominated by sample-level or spectral-point losses—tends to collapse conditional uncertainty, leading to a regression-to-the-mean effect. Empirically, the reconstructed audio signals often exhibit spectral over-smoothing and reduced harmonic, transient, and aperiodic micro-structure. Consequently, such models may preserve LF fidelity while under-delivering perceptual brightness, detail, and naturalness.

In contrast, generative modeling—including AR models, VAEs, GANs, diffusion and score-based models, flow-based models, and bridge models—reframes BWE/SR as learning a conditional distribution rather than a deterministic mapping. This probabilistic perspective better reflects the inherent ambiguity of the task: the missing HF band is not uniquely determined by the observed LF content. A principled model should therefore represent conditional uncertainty and support multiple plausible reconstructions. Consequently, generative approaches can synthesize HF details that are consistent with semantic content, timbral structure, and event characteristics across speech, music, and general sounds, rather than regressing toward an average spectrum. This explains the shift toward conditional generative modeling, where conditioning enforces fidelity to the observed band while distribution modeling enhances perceptual realism.

Within the generative family, GANs were early attempts to improve perceptual sharpness through adversarial learning. However, they have become less dominant in recent BWE/SR research because adversarial objectives can be unstable and prone to generating HF details that sound sharp but are weakly constrained by the observed LF content. These trade-offs have motivated increasing interest in generative paradigms with more explicit sampling dynamics and better-controlled conditional generation.

Diffusion, flow, and bridge-based models address this limitation in different ways. Diffusion and score-based models capture complex conditional distributions through iterative denoising, often achieving high perceptual quality and strong mode coverage, but at the cost of multi-step sampling latency (Han & Lee, 2022; Liu et al., 2024a). Flow-based models instead learn continuous transformations between simple and data distributions, offering a more direct path toward efficient generation and few-step sampling for low-latency inference (Welker et al., 2025; Hsieh & Braun, 2026). Bridge models further generalize this view by learning stochastic paths between degraded and clean signals, which is conceptually well aligned with restoration tasks, but they often introduce additional modeling and sampling complexity (Li et al., 2025a;b). In this landscape, flow-based approaches are particularly favorable for BWE/SR as shown in Fig. 1, because they offer a practical balance between generative flexibility and inference efficiency: they retain the distributional modeling benefits needed for plausible HF synthesis while enabling faster sampling than standard diffusion models, making them attractive for real-time audio reconstruction.

7.2 Challenges and Open Problems

Despite the growing success of diffusion, bridge, and flow-based models for BWE/SR, several challenges remain open. First, high perceptual quality often comes with nontrivial inference cost, especially for iterative diffusion and bridge models, making low-latency or real-time deployment difficult; although flow-based and flow-matching methods improve efficiency, their quality-speed trade-offs remain task- and design-dependent. Second, evaluation remains a major bottleneck, as common objective metrics may not reliably capture perceptual naturalness, artifact suppression, or the realism of hallucinated HF details; this limitation is especially pronounced for music and general audio, where perceived quality depends on timbre, texture, and event realism rather than waveform fidelity alone. Third, robustness to real-world degradations remains limited, since many systems are trained on simplified low-pass or downsampling operators and may degrade under codec artifacts, noise, device coloration, unknown bandwidth constraints, and broader domain shifts. Finally, the design space of diffusion schedules, bridge endpoints, flow paths, conditioning mechanisms, and neural representations remains underexplored, leaving open questions about how to balance fidelity, diversity, controllability, and efficiency in next-generation generative BWE/SR systems.

7.3 Future Outlook

Recent significant advances in LLMs (Touvron et al., 2023; DeepSeek-AI, 2024; Nie et al., 2025; Li et al., 2025d; Wang et al., 2025b; Wu et al., 2025), large audio-language models (LALMs) (Chen et al., 2025; Chu et al., 2024; Ghosh et al., 2025; 2026; Wu et al., 2025; Rong et al., 2026; Gong et al., 2026), and multimodal pretraining (Xu et al., 2025; Wang et al., 2025a; Kong et al., 2025; Wang et al., 2025b) suggest several promising directions for BWE/SR. First, large Transformer-based foundation models trained on audio-text or broader multimodal corpora can provide semantically structured priors over speech, music, general sounds, and complex acoustic scenes. Conditioning BWE/SR models on such high-level representations may help constrain HF synthesis, reducing implausible hallucination while preserving perceptually meaningful detail.

Second, retrieval-augmented paradigms offer a possible route toward more controllable BWE/SR. Instead of relying solely on parametric model memory, future systems could retrieve high-quality exemplars matched by speaker identity, instrument family, acoustic scene, timbral profile, or caption-level semantics, and use these references to guide HF reconstruction conditioned on the degraded input. Such retrieval mechanisms may be especially useful when the missing bandwidth depends strongly on source-specific structure, but they also raise open questions about retrieval quality, reference mismatch, and computational overhead.

Additionally, self-supervised and multimodal representation learning may improve robustness and evaluation. Objectives such as contrastive learning, masked audio modeling, and audio-language alignment can produce representations that transfer across restoration tasks and degradation conditions, potentially improving generalization to noise, codec artifacts, device coloration, and unseen bandwidth constraints. In parallel, learned perceptual metrics derived from audio-language or audio-visual models may better capture human judgments of naturalness, timbre, texture, and event realism than conventional signal-based measures, making them useful for evaluation, perceptual fine-tuning, and guided generation.

Overall, future progress in BWE/SR will likely depend not only on stronger generative backbones, but also on improved conditioning, retrieval, perceptual evaluation, and deployment-aware optimization. Joint advances across these directions will be essential for developing robust, efficient, and perceptually faithful systems across speech, music, and general audio.

8 Conclusion

This survey presented a unified review of BWE and SR, treating both as ill-posed HF reconstruction problems while clarifying their relationship and distinctions. We traced the field's evolution from discriminative DNN-based models to modern generative paradigms, including AR models, VAEs, GANs, diffusion and score-based models, flow-based models, and Schrödinger bridges. From this perspective, we highlighted a central transition in BWE/SR: from point estimation, which often leads to regression-to-the-mean behavior and spectral over-smoothing, to distribution-aware generation, which better captures the one-to-many nature of missing HF content. We further summarized key trade-offs among existing methods in reconstruction fidelity, perceptual quality, robustness, controllability, and computational efficiency, and identified open challenges in scalable deployment, perceptual evaluation, and real-world generalization. We hope this survey provides a clear taxonomy and practical roadmap for advancing robust, efficient, and perceptually faithful BWE/SR systems across speech, music, and general audio.

References

Johannes Abel and Tim Fingscheidt. Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

Johannes Abel, Maximilian Strake, and Tim Fingscheidt. Artificial bandwidth extension using deep neural networks for spectral envelope estimation. In *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016.

Johannes Abel, Maximilian Strake, and Tim Fingscheidt. A simple cepstral domain dnn approach to artificial speech bandwidth extension. In *ICASSP*. IEEE, 2018.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 1985.

Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Dmitry Vetrov. Hifi++: a unified framework for bandwidth extension and speech enhancement. In *ICASSP*. IEEE, 2023.

Pavel Andreev et al. Wv-mos: Mos score prediction by fine-tuned wav2vec 2.0. *arXiv preprint arXiv:2203.13086*, 2022.

Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *CVPR*, 2018.

Pramod Bachhav, Massimiliano Todisco, and Nicholas Evans. Artificial bandwidth extension using conditional variational auto-encoders and adversarial learning. In *ICASSP*. IEEE, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*, 2021.

Hongtao Bao and Xueliang Zhang. Frequency-domain enhanced extreme bandwidth extension network with iccrn for superior speech quality. In *Proc. Interspeech 2025*, 2025.

John G Beerends, Andries P Hekstra, Antony W Rix, Michael P Hollier, et al. Perceptual evaluation of speech quality (pesq): The new itu standard for end-to-end speech quality assessment part ii-psychoacoustic model. *Journal of the Audio Engineering Society*, 50(10):765–778, 2002.

Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei W Koh, and Stefano Ermon. Temporal film: Capturing long-range sequence dependencies with feature-wise modulations. *NeurIPS*, 2019.

Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *Ismir*, volume 14, pp. 155–160, 2014.

Cassia Valentini Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA speech synthesis workshop*, pp. 159–165, 2016.

George Box. Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, 30(1):1–17, 1988.

Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, 2018.

Ruizhe Cao, Sherif Abdulatif, and Bin Yang. Cmgan: Conformer-based metric gan for speech enhancement. *arXiv preprint arXiv:2203.15149*, 2022.

Clement Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and EngSiong Chng. Audio large language models can be descriptive speech quality evaluators. *ICLR*, 2025.

Guo Chen and Vijay Parsa. Hmm-based frequency bandwidth extension for speech enhancement using line spectral frequencies. In *ICASSP*. IEEE, 2004.

Xinyuan Chen, Chang Xu, Xiaokang Yang, Li Song, and Dacheng Tao. Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Transactions on Image Processing*, 28(2):546–560, 2018.

Zehua Chen, Guande He, Kaiwen Zheng, Xu Tan, and Jun Zhu. Schrodinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023.

Woongjib Choi, Sangmin Lee, Hyungseob Lim, and Hong-Goo Kang. Universr: Unified and versatile audio super-resolution via vocoder-free flow matching. *ICASSP*, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

C Corpus. The design for the wall street journal-based. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, pp. 357. Morgan Kaufmann Publishers, 1992.

Lingling Dai, Yuxuan Ke, Andong Li, Xiaodong Li, and Chengshi Zheng. Sfnet: A two-stage source-filter-based neural network for real-time speech bandwidth extension. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *CVPR*, 34:17695–17709, 2021.

Valentin De Bortoli, Iryna Korshunova, Andriy Mnih, and Arnaud Doucet. Schrodinger bridge flow for unpaired data translation. *Advances in Neural Information Processing Systems*, 37:103384–103441, 2024.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*, 2020.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *ICASSP*. IEEE, 2018a.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018b.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

Yuanjie Dong, Yaxing Li, Xiaoqi Li, Shan Xu, Dan Wang, Zhihui Zhang, and Shengwu Xiong. A time-frequency network with channel attention and non-local modules for artificial bandwidth extension. In *ICASSP*. IEEE, 2020.

Amin Edraki, Wai-Yip Chan, Jesper Jensen, and Daniel Fogerty. Speaker adaptation for enhancement of bone-conducted speech. In *ICASSP*. IEEE, 2024.

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

Sefik Emre Eskimez and Kazuhito Koishida. Speech super resolution generative adversarial network. In *ICASSP*. IEEE, 2019.

Yuan Fang, Jinglin Bai, Jiajie Wang, and Xueliang Zhang. Vector quantized diffusion model based speech bandwidth extension. In *ICASSP*. IEEE, 2025.

Miaomiao Gao, Xiaojuan Zhang, and Xiaoxiao Xiang. Real-world speech recovery under multiple distortions: A two-stage framework with feature consistency and adversarial fine-tuning. *IEEE Signal Processing Letters*, 2026a.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. *ICCV*, 2023.

Yuan Gao, Xin Wang, Xiaotong Wang, Yibin Tang, Aimin Jiang, and Ying Chen. A harmonic-coupled generative adversarial network for speech super-resolution in low bandwidth scenarios. *IEEE Transactions on Audio, Speech and Language Processing*, 2026b.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database. *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, 107:16, 1988.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *NeurIPS*, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Esteban Gómez, Mohammad Hassan Vali, and Tom Bäckström. Low-complexity real-time neural network for blind bandwidth extension of wideband speech. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 31–35. IEEE, 2023.

Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, Ruixiao Li, Qingyuan Cheng, Shimin Li, and Xipeng Qiu. Moss-audio-tokenizer: Scaling audio tokenizers for future audio foundation models. *arXiv preprint arXiv:2602.10934*, 2026.

Augustine Gray and John Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2003.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *COLM*, 2024.

Yicheng Gu, Xueyao Zhang, Liumeng Xue, and Zhizheng Wu. Multi-scale sub-band constant-q transform discriminator for high-fidelity vocoder. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10616–10620. IEEE, 2024.

Yu Gu and Zhen-Hua Ling. Restoring high frequency spectral envelopes using neural networks for speech bandwidth extension. In *IJCNN*. IEEE, 2015.

Yu Gu and Zhen-Hua Ling. Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension. In *INTERSPEECH*, pp. 1123–1127, 2017.

Yu Gu, Zhen-Hua Ling, and Li-Rong Dai. Speech bandwidth extension using bottleneck features and deep recurrent neural networks. In *Interspeech*, 2016.

Archit Gupta, Brendan Shillingford, Yannis Assael, and Thomas C Walters. Speech bandwidth extension with wavenet. In *2019 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pp. 205–208. IEEE, 2019.

Seungu Han and Junhyeok Lee. Nu-wave 2: A general neural audio upsampling model for various sampling rates. *arXiv preprint arXiv:2206.08545*, 2022.

Xiang Hao, Chenglin Xu, Nana Hou, Lei Xie, Eng Siong Chng, and Haizhou Li. Time-domain neural network approach for speech bandwidth extension. In *ICASSP*. IEEE, 2020.

Julien Hauret, Thomas Joubaud, Véronique Zimpfer, and Éric Bavu. Eben: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones. In *ICASSP*. IEEE, 2023.

David Haws and Xiaodong Cui. Cyclegan bandwidth extension acoustic modeling for automatic speech recognition. In *ICASSP*. IEEE, 2019.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247*, 2018.

Carlos Hernandez-Olivan, Hendrik Vincent Koops, Hao Hao Tan, and Elio Quinton. Single-step controllable music bandwidth extension with flow matching. *ICASSP*, 2026.

Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International conference on machine learning*, pp. 2722–2730. PMLR, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Nana Hou, Chenglin Xu, Van Tung Pham, Joey Tianyi Zhou, Eng Siong Chng, and Haizhou Li. Speaker and phoneme-aware speech bandwidth extension with residual dual-path network. *Interspeech*, 2020.

Tsun-An Hsieh and Sebastian Braun. Towards real-time generative speech restoration with flow-matching. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15847–15851. IEEE, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *CVPR*, 2022.

Jaekwon Im and Juhan Nam. Flashsr: One-step versatile audio super-resolution via diffusion distillation. In *ICASSP*. IEEE, 2025.

Jaekwon Im and Juhan Nam. Saga-sr: Semantically and acoustically guided audio super-resolution. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1706–1710. IEEE, 2026.

Keith Ito and Linda Johnson. The LJ Speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*, 2021.

Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

Abdul J Jerri. The shannon sampling theorem—its various extensions and applications: A tutorial review. *Proceedings of the IEEE*, 65(11):1565–1596, 1977.

Yizhu Jin, Zhen Ye, Zeyue Tian, Haohe Liu, Qiuqiang Kong, Yike Guo, and Wei Xue. Inference-time scaling for diffusion-based audio super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14982–14990, 2026.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023.

Donghyun Kim and Joon-Hyuk Chang. Progressive subband modeling for artifacts-free speech super-resolution. In *ICASSP*. IEEE, 2025.

Donghyun Kim, Yungyeo Kim, and Joon-Hyuk Chang. Class: Continual learning approach for speech super-resolution. In *ICASSP*. IEEE, 2024a.

Seung-Bin Kim, Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee. Audio super-resolution with robust speech representation learning of masked autoencoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1012–1022, 2024b.

Sung Kim and Visvesh Sathe. Bandwidth extension on raw audio via generative adversarial networks. *arXiv preprint arXiv:1903.09027*, 2019.

Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen. On loss functions for supervised monaural time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:825–838, 2020.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *CVPR*, 2020.

Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. *arXiv preprint arXiv:2109.05418*, 2021a.

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, et al. Token reduction should go beyond efficiency in generative models – from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *ICLR*, 2021b.

Pin-Jui Ku, Alexander H Liu, Roman Korostik, Sung-Feng Huang, Szu-Wei Fu, and Ante Jukić. Generative speech foundation model pretraining for high-quality speech extraction and restoration. In *ICASSP*. IEEE, 2025.

Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. Audio super-resolution using neural nets. In *ICLR (Workshop Track)*, 2017.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *CVPR*, 2019a.

Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2 (5):3, 2019b.

Rithesh Kumar, Kundan Kumar, Vicki Anand, Yoshua Bengio, and Aaron Courville. Nu-gan: High resolution neural upsampling with gan. *arXiv preprint arXiv:2010.11362*, 2020.

Nikita Kuznetsov and Maksim Kaledin. Fastwave: Optimized diffusion model for audio super-resolution. *arXiv preprint arXiv:2603.04122*, 2026.

Doyeop Kwak, Youngjoon Jang, Seongyu Kim, and Joon Son Chung. Ednet: A versatile speech enhancement framework with gating mamba mechanism and phase shift-invariant training. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *Interspeech*, 2021.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

Yongjoon Lee and Jung-Woo Choi. Semamba++: A general speech restoration framework leveraging global, local, and periodic spectral patterns. *arXiv preprint arXiv:2603.11669*, 2026.

Yongjoon Lee and Chanwoo Kim. Wave-u-mamba: an end-to-end framework for high-quality and efficient speech super resolution. In *ICASSP*. IEEE, 2025.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Chang Li, Zehua Chen, Fan Bao, and Jun Zhu. Bridge-sr: Schrödinger bridge for efficient sr. In *ICASSP*. IEEE, 2025a.

Chang Li, Zehua Chen, Liyuan Wang, and Jun Zhu. Audio super-resolution with latent bridge models. *arXiv preprint arXiv:2509.17609*, 2025b.

Changtao Li, Feiran Yang, and Jun Yang. Restoration of bone-conducted speech with u-net-like model and energy distance loss. *IEEE Signal Processing Letters*, 31:166–170, 2023a.

Changtao Li, Feiran Yang, and Jun Yang. A two-stage approach to quality restoration of bone-conducted speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:818–829, 2023b.

Kehuang Li and Chin-Hui Lee. A deep neural network approach to speech bandwidth expansion. In *ICASSP*. IEEE, 2015.

Kehuang Li, Zhen Huang, Yong Xu, and Chin-Hui Lee. Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In *Interspeech*, 2015.

Sen Li, Stéphane Villette, Pravin Ramadas, and Daniel J Sinder. Speech bandwidth extension using generative adversarial networks. In *ICASSP*. IEEE, 2018.

Xinyu Li, Venkata Chebiyyam, Katrin Kirchhoff, and A Amazon. Speech audio super-resolution for speech recognition. In *INTERSPEECH*, pp. 3416–3420, 2019.

Yize Li, Yihua Zhang, Sijia Liu, and Xue Lin. Pruning then reweighting: Towards data-efficient training of diffusion models. *ICASSP*, 2025c.

Yunpeng Li, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek. Real-time speech frequency bandwidth extension. In *ICASSP*. IEEE, 2021.

Yuqi Li, Zijie Zhou, et al. A preference-driven methodology for efficient code generation. *IEEE Transactions on Artificial Intelligence*, 2025d.

Yuqi Li, Kai Li, Xin Yin, Zhifei Yang, Junhao Dong, Zeyu Dong, Chuanguang Yang, Yingli Tian, and Yao Lu. Sepprune: Structured pruning for efficient deep speech separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Teck Yian Lim, Raymond A Yeh, Yijia Xu, Minh N Do, and Mark Hasegawa-Johnson. Time-frequency networks for audio super-resolution. In *ICASSP*. IEEE, 2018.

Ju Lin, Yun Wang, Kaustubh Kalgaonkar, Gil Keren, Didi Zhang, and Christian Fuegen. A two-stage approach to speech bandwidth extension. In *Interspeech*, volume 2021, pp. 1689–1693, 2021.

Yin-Tse Lin, Bo-Hao Su, Chi-Han Lin, Shih-Chan Kuo, Jyh-Shing Roger Jang, and Chi-Chun Lee. Noise-robust bandwidth expansion for 8k speech recordings. In *Interspeech*, pp. 5107–5111, 2023.

Yin-Tse Lin, Shreya G Upadhyay, Bo-Hao Su, and Chi-Chun Lee. Swibe: A parameterized stochastic diffusion process for noise-robust bandwidth expansion. In *Proc. Interspeech 2024*, 2024.

Zhen-Hua Ling, Yang Ai, Yu Gu, and Li-Rong Dai. Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):883–894, 2018.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Bin Liu, Jianhua Tao, Zhengqi Wen, Ya Li, and Danish Bukhari. A novel method of artificial bandwidth extension using deep architecture. In *INTERSPEECH*, 2015.

Bin Liu, Jianhua Tao, et al. A novel research to artificial bandwidth extension based on deep blstm recurrent neural networks and exemplar-based sparse representation. In *Interspeech*, 2016.

Haohe Liu, Woosung Choi, Xubo Liu, Qiuqiang Kong, Qiao Tian, and DeLiang Wang. Neural vocoder is all you need for speech super-resolution. *arXiv preprint arXiv:2203.14941*, 2022a.

Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: A unified framework for high-fidelity speech restoration. *arXiv preprint arXiv:2204.05841*, 2022b.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley. Audiosr: Versatile audio super-resolution at scale. In *ICASSP*. IEEE, 2024a.

Xi Liu, Mu Yang, Szu-Jui Chen, and John HL Hansen. A neural codec approach for noise-robust bandwidth expansion. In *Proc. Interspeech 2025*, 2025a.

Xin Liu, Shulin He, and Xueliang Zhang. Hwb-net: A novel high-performance and efficient hybrid waveform bandwidth extension method. In *Proc. Interspeech 2025*, 2025b.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *CVPR*, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.

Lei Lu, Yize Li, Yanzhi Wang, Wei Wang, and Wei Jiang. Hdcompression: Hybrid-diffusion image compression for ultra-low bitrates. *PRICAI*, 2025.

Ye-Xin Lu, Yang Ai, Hui-Peng Du, and Zhen-Hua Ling. Towards high-quality and efficient speech bandwidth extension with parallel amplitude and phase prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024a.

Ye-Xin Lu, Yang Ai, Zheng-Yan Sheng, and Zhen-Hua Ling. Multi-stage speech bandwidth extension with flexible sampling rate control. *arXiv preprint arXiv:2406.02250*, 2024b.

Dwarikanath Mahapatra, Behzad Bozorgtabar, and Rahil Garnavi. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, 71:30–39, 2019.

John Makhoul and Michael Berouti. High-frequency regeneration in speech coding systems. In *ICASSP*. IEEE, 1979.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. The song describer dataset: a corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*, 2023.

Moshe Mandel, Or Tal, and Yossi Adi. Aero: Audio super resolution in the spectral domain. In *ICASSP*. IEEE, 2023.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 3339–3354, 2024.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*, 2021.

Eloi Moliner and Vesa Välimäki. Behm-gan: Bandwidth extension of historical music using generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:943–956, 2022.

Eloi Moliner, Filip Elvander, and Vesa Välimäki. Blind audio bandwidth extension: A diffusion-based zero-shot approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. Chunked autoregressive gan for conditional waveform synthesis. *arXiv preprint arXiv:2110.10139*, 2021.

Viet-Anh Nguyen, Anh HT Nguyen, and Andy WH Khong. Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining. In *ICASSP*. IEEE, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pp. 8162–8171. PMLR, 2021.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *NeurIPS*, 2025.

Yamato Ohtani, Masatsune Tamura, Masahiro Morita, and Masami Akamine. Gmm-based bandwidth extension using sub-band basis spectrum model. In *Interspeech*, pp. 2489–2493, 2014.

K Paliwal and Anjan Basu. A speech enhancement method based on kalman filtering. In *ICASSP*. IEEE, 1987.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*. IEEE, 2015.

Ashutosh Pandey and DeLiang Wang. A new framework for cnn-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

Stefano Peluchetti. Diffusion bridge mixture transports, schrödinger bridge problems and generative modeling. *Journal of Machine Learning Research*, 24(374):1–51, 2023.

Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl. Moisesdb: A dataset for source separation beyond 4-stems. *arXiv preprint arXiv:2307.15913*, 2023.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: straightening flows with minibatch couplings. *ICML*, 2023.

N Prasad and T Kishore Kumar. Bandwidth extension of speech signals: A comprehensive review. *International Journal of Intelligent Systems and Applications*, 8(2):45–52, 2016.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18-hq-an uncompressed version of musdb18. *(No Title)*, 2019.

Nathanaël Carraz Rakotonirina. Self-attention for audio super-resolution. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021.

CCITT Recommendation. Pulse code modulation (pcm) of voice frequencies. In *ITU*, 1988.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497. IEEE, 2021.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*. Pmlr, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Yan Rong, Chenxing Li, Dong Yu, and Li Liu. Audiogenie-reasoner: A training-free multi-agent framework for coarse-to-fine audio deep reasoning. *ICASSP*, 2026.

David Santos-Domínguez, Soledad Torres-Guijarro, Antonio Cardenal-López, and Antonio Pena-Gimenez. Shipsear: An underwater vessel noise database. *Applied Acoustics*, 113:64–69, 2016.

Jonas Sautter, Friedrich Faubel, Markus Buck, and Gerhard Schmidt. Artificial bandwidth extension using a conditional generative adversarial network with discriminative training. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7005–7009. IEEE, 2019.

Robin Scheibler, Yusuke Fujita, Yuma Shirahata, and Tatsuya Komatsu. Universal score-based speech enhancement with high content preservation. *arXiv preprint arXiv:2406.12194*, 2024.

Konstantin Schmidt and Bernd Edler. Blind bandwidth extension based on convolutional and recurrent deep neural networks. In *ICASSP*. IEEE, 2018.

Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.

Siavash Shams, Sukru Samet Dindar, Xilin Jiang, and Nima Mesgarani. Ssamba: Self-supervised audio representation learning with mamba state space model. *IEEE Spoken Language Technology (SLT) Workshop*, 2024.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*. IEEE, 2018.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

Kevin J Shih, Zhifeng Kong, Weili Nie, Arash Vahdat, Sang-gil Lee, Joao Felipe Santos, Ante Jukić, Rafael Valle, and Bryan Catanzaro. Audio-to-audio schrodinger bridges. *AI for Music Workshop at NeurIPS*, 2025.

Chenhao Shuai, Chaohua Shi, Lu Gan, and Hongqing Liu. mdctGAN: Taming transformer-based GAN for speech super-resolution with Modified DCT spectra. *arXiv preprint arXiv:2305.11104*, 2023.

Erfan Soltanmohammadi, Paris Smaragdis, and Michael M Goodwin. Low-complexity streaming speech super-resolution. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *ICASSP*. IEEE, 2018.

Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

Jiaqi Su, Yunyun Wang, Adam Finkelstein, and Zeyu Jin. Bandwidth extension is all you need. In *ICASSP*. IEEE, 2021.

Yueyuan Sui, Minghui Zhao, Junxi Xia, Xiaofan Jiang, and Stephen Xia. Tramba: A hybrid transformer and mamba architecture for practical audio and bone conduction speech super resolution and enhancement on mobile and wearable platforms. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29, 2024.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *ICASSP*. IEEE, 2010.

Tarikul Islam Tamiti and Anomadarshi Barua. Nldsi-bwe: Non linear dynamical systems-inspired multi resolution discriminators for speech bandwidth extension. *arXiv preprint arXiv:2510.01109*, 2025.

Tarikul Islam Tamiti, Biraj Joshi, Rida Hasan, Rashedul Hasan, Taieba Athay, Nursad Mamun, and Anomadarshi Barua. A high-fidelity speech super resolution network using a complex global attention module with spectro-temporal loss. *arXiv preprint arXiv:2507.00229*, 2025.

Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, volume 2018, pp. 3229–3233, 2018.

Qiao Tian, Yi Chen, Zewang Zhang, Heng Lu, Linghui Chen, Lei Xie, and Shan Liu. Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis. *arXiv preprint arXiv:2011.12206*, 2020.

Ye Tian, Zhe Wang, Jianguo Sun, and Liguo Zhang. Time-frequency domain fusion enhancement for audio super-resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2879–2887, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Takahiro Unno and Alan McCree. A robust narrowband to wideband extension system featuring enhanced codebook mapping. In *ICASSP*. IEEE, 2005.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12:1, 2016.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *CVPR*, 29, 2016.

Maaike Van Eeckhoutte, Paula Folkeard, Danielle Glista, and Susan Scollie. Speech recognition, loudness, and preference with extended bandwidth hearing aids for adult hearing aid users. *International Journal of Audiology*, 59(10):780–791, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

Heming Wang and Deliang Wang. Towards robust speech super-resolution. *IEEE/ACM transactions on audio, speech, and language processing*, 2021.

Heming Wang and Deliang Wang. Time-frequency loss for cnn based speech super-resolution. In *ICASSP*. IEEE, 2020.

Heming Wang, Xueliang Zhang, and DeLiang Wang. Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement. *IEEE/ACM transactions on audio, speech, and language processing*, 30:3134–3143, 2022.

Mu Wang, Zhiyong Wu, Shiyin Kang, Xixin Wu, Jia Jia, Dan Su, Dong Yu, and Helen Meng. Speech super-resolution using parallel wavenet. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 260–264. IEEE, 2018.

Siyi Wang, Siyi Liu, Andrew Harper, Paul Kendrick, Mathieu Salzmann, and Milos Cernak. Diffusion-based speech enhancement with schrödinger odinger bridge and symmetric noise schedule. *arXiv preprint arXiv:2409.05116*, 2024a.

Yingxue Wang, Shenghui Zhao, Wenbo Liu, Ming Li, and Jingming Kuang. Speech bandwidth expansion based on deep neural networks. In *INTERSPEECH*, 2015.

Yuji Wang, Zehua Chen, Xiaoyu Chen, Yixiang Wei, Jun Zhu, and Jianfei Chen. Framebridge: Improving image-to-video generation with bridge models. *arXiv preprint arXiv:2410.15371*, 2024b.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. Filter-and-refine: A mllm based cascade system for industrial-scale video content moderation. *ACL*, 2025a.

Zixuan Wang, Yu Sun, Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Luna Dong, Zhuolin Hao, Hongyu Xiong, and Yang Song. Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content governance. *EMNLP*, 2025b.

Simon Welker, Bunlong Lay, Maris Hillemann, Tal Peer, and Timo Gerkmann. Real-time streamable generative speech restoration with flow matching. *arXiv preprint arXiv:2512.19442*, 2025.

Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. Are we using enough listeners? no! an empirically-supported critique of interspeech 2014 tts evaluations. In *Interspeech 2015*. International Speech Communication Association, 2015.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

Yusong Wu, Josh Gardner, Ethan Manilow, Ian Simon, Curtis Hawthorne, and Jesse Engel. The chamber ensemble generator: Limitless high-quality mir data via generative modeling. *arXiv preprint arXiv:2209.14458*, 2022.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Junichi Yamagishi. English multi-speaker corpus for cstr voice cloning toolkit. *URL http://homepages. inf. ed. ac. uk/jyamagis/page3/page58/page58. html*, 2012.

Hitoshi Yamamoto, Kong Aik Lee, Koji Okabe, and Takafumi Koshinaka. Speaker augmentation and bandwidth extension for deep speaker embedding. In *Interspeech*, pp. 406–410, 2019.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP*. IEEE, 2020.

Junkang Yang, Hongqing Liu, Lu Gan, Yi Zhou, Xing Li, Jie Jia, and Jinzhuo Yao. Sdnet: Noise-robust bandwidth extension under flexible sampling rates. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–6. IEEE, 2024.

Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. Controllable artistic text style transfer via shape-matching gan. In *ICCV*, 2019a.

Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019b.

Reo Yoneyama, Ryuichi Yamamoto, and Kentaro Tachibana. Nonparallel high-quality audio super resolution with domain adaptation and resampling cyclegans. In *ICASSP*. IEEE, 2023.

Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung. Time-domain multi-modal bone/air conducted speech enhancement. *IEEE Signal Processing Letters*, 27:1035–1039, 2020.

Chin-Yun Yu, Sung-Lin Yeh, György Fazekas, and Hao Tang. Conditioning and sampling in variational diffusion models for speech super-resolution. In *ICASSP*. IEEE, 2023.

Guochen Yu, Xiguang Zheng, Nan Li, Runqiang Han, Chengshi Zheng, Chen Zhang, Chao Zhou, Qi Huang, and Bing Yu. Bae-net: A low complexity and high fidelity bandwidth-adaptive neural network for speech super-resolution. In *ICASSP*. IEEE, 2024.

Jiajun Yuan, Xiaochen Wang, Yuhang Xiao, et al. Swinsrgan: Swin transformer-based generative adversarial network for high-fidelity speech super-resolution. *arXiv preprint arXiv:2509.03913*, 2025.

Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal processing*, 128:389–408, 2016.

Jun-Hak Yun, Seung-Bin Kim, and Seong-Whan Lee. Flowhigh: Towards efficient and high-quality audio super-resolution with single-step flow matching. In *ICASSP*. IEEE, 2025.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Bowen Zhang, Junchuan Zhao, Ian McLoughlin, Ye Wang, and A S Madhukumar. Codecflow: Efficient bandwidth extension via conditional flow matching in neural codec latent space. *arXiv preprint arXiv:2603.02022*, 2026.

Kexun Zhang, Yi Ren, Changliang Xu, and Zhou Zhao. Wsrglow: A glow-based waveform generative model for audio super-resolution. *arXiv preprint arXiv:2106.08507*, 2021.

Ting-Wei Zhang and Shanq-Jang Ruan. Vm-asr: A lightweight dual-stream u-net model for efficient audio super-resolution. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, Zhaoheng Ni, Anurag Kumar, Jan Pirklbauer, Marvin Sach, Shinji Watanabe, et al. Urgent challenge: Universality, robustness, and generalizability for speech enhancement. *arXiv preprint arXiv:2406.04660*, 2024.

S Zhao, Y Ma, C Ni, C Zhang, H Wang, TH Nguyen, K Zhou, J Yip, D Ng, and B Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. *ICASSP*, 2024.

Shengkui Zhao, Kun Zhou, Zexu Pan, Yukun Ma, Chong Zhang, and Bin Ma. Hifi-sr: A unified generative transformer-convolutional adversarial network for high-fidelity speech super-resolution. In *ICASSP*. IEEE, 2025.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*, 2023.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019.