Computer Speech and Language 94 (2025) 101808

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Combined generative and predictive modeling for speech super-resolution

Heming Wang [a,*], Eric W. Healy [a,b], DeLiang Wang [a,b]

[a] The Ohio State University, 281 W Lane Ave, Columbus, 43210 OH, United States
[b] Center for Cognitive and Brain Science, 1835 Neil Ave, Columbus, 43210 OH, United States

ARTICLE INFO

Keywords:
Speech super-resolution
Multi-stage learning
Diffusion model
Bandwidth extension

ABSTRACT

Speech super-resolution (SR) is the task that restores high-resolution speech from low-resolution input. Existing models employ simulated data and constrained experimental settings, which limit generalization to real-world SR. Predictive models are known to perform well in fixed experimental settings, but can introduce artifacts in adverse conditions. On the other hand, generative models learn the distribution of target data and have a better capacity to perform well on unseen conditions. In this study, we propose a novel two-stage approach that combines the strengths of predictive and generative models. Specifically, we employ a diffusion-based model that is conditioned on the output of a predictive model. Our experiments demonstrate that the model significantly outperforms single-stage counterparts and existing strong baselines on benchmark SR datasets. Furthermore, we introduce a repainting technique during the inference of the diffusion process, enabling the proposed model to regenerate high-frequency components even in mismatched conditions. An additional contribution is the collection of and evaluation on real SR recordings, using the same microphone at different native sampling rates. We make this dataset freely accessible, to accelerate progress towards real-world speech super-resolution.

1. Introduction

Speech super-resolution (SR), also known as speech bandwidth extension (BWE), aims to increase the sampling rate of low-resolution speech by generating high-frequency components. It can improve the quality and intelligibility of the speech signal. In addition, speech super-resolution has many potential applications in fields such as automatic speech recognition (Li and Lee, 2015; Albahri et al., 2016), hearing aids (Füllgrabe et al., 2010; Van Eeckhoutte et al., 2020), and text-to-speech synthesis (Nakamura et al., 2014). Speech SR has been addressed through various methods such as signal processing and deep learning methods.

Signal processing methods typically adopt the source-filter model (Milner and Shao, 2002), which assumes that speech is produced by an excitation signal followed by a time-variant filter that simulates the vocal tract. The bandwidth extension is then divided into two separate tasks: spectral envelope extension and excitation signal extension. Early studies employ codebook mapping techniques for spectral envelope extension, using two codebooks representing narrowband and wideband spectra (Unno and McCree, 2005; Sadasivan et al., 2016). Statistical methods such as the Gaussian Mixture Model (GMM) (Nour-Eldin and Kabal, 2008, 2011) are also used. Unlike the codebook methods that rely on vector quantization and operate in a discrete space, GMM allows a continuous mapping of wideband coefficients from narrowband parameters. Excitation signal extension is an easier task. The most straightforward method is to perform spectral folding (Makhoul and Berouti, 1979; De Cheveigné and Kawahara, 2002)

* Corresponding author.
E-mail addresses: wang.11401@osu.edu (H. Wang), healy.66@osu.edu (E.W. Healy), dwang@cse.ohio-state.edu (D.L. Wang).

https://doi.org/10.1016/j.csl.2025.101808
Received 8 April 2024; Received in revised form 24 January 2025; Accepted 15 April 2025
Available online 8 May 2025

to extend the excitation signal from narrowband to wideband. To preserve the harmonic structure of the excitation signal, one can employ sinusoidal synthesis (Chan and Hui, 1996; Abel and Fingscheidt, 2019) by utilizing a bank of oscillators in the narrowband.

Deep neural networks (DNNs) have been demonstrated to be superior to traditional statistical approaches for speech SR. Early DNN studies focus on speech magnitude, predicting wideband magnitude from narrowband magnitude using features such as log-power magnitudes (Li and Lee, 2015; Liu et al., 2015; Gu et al., 2016), vocal tract filter parameters (Botinhao et al., 2006; Kontio et al., 2007; Pulakka and Alku, 2011), and cepstral coefficients (Abel and Fingscheidt, 2019). The phase of the upperband is produced by flipping or copying the narrowband phase. Later studies address SR in the time domain Gu and Ling (2017), Kuleshov et al. (2017), which naturally incorporates signal phase and yields better reconstruction quality. For example, Lim et al. (2018) proposed a time–frequency network (TFNet) that incorporates both time-domain and frequency-domain networks. We proposed a time-domain convolutional network trained with a cross-domain loss function (Wang and Wang, 2021). Lin et al. (2021) proposed a two-stage approach that combines the advantages of time- and frequency-domain methods.

Another line of studies treats the SR task as a conditional generation problem. Rather than directly mapping LR (low-resolution) to HR (high-resolution) signals, these studies aim to learn a data distribution conditioned on LR features and then generate HR from the learned distribution. Specifically, mel-spectrogram-based speech synthesizers are employed to generate HR speech by reconstructing mel-spectrograms from LR speech inputs (Prenger et al., 2019; Ling et al., 2018; Gupta et al., 2019; Liu et al., 2022a; Kim et al., 2024). Recent studies have also adopted conditional generative adversarial networks (GANs) (Li et al., 2018; Haws and Cui, 2019; Eskimez et al., 2019; Kataria et al., 2024) or diffusion models that are conditioned on low-resolution signals (Lee and Han, 2021; Han and Lee, 2022; Yu et al., 2022) to extend the bandwidth.

A major challenge in DNN-based super-resolution (SR) is training data acquisition. Most studies use simulation techniques to generate HR/LR pairs by downsampling HR signals. Although this approach is straightforward and can efficiently generate a large amount of data, the simulation process may introduce certain characteristics specific to the downsampling settings, limiting the robustness of the learned models in real-world application (Sulun and Davies, 2020; Wang and Wang, 2021). Another challenge is rooted in the ill-posed nature of the SR problem, as the relationship of LR to HR signals is one-to-many, i.e. an LR signal may correspond to many valid HR signals. Predictive learning methods learn a one-to-one mapping from LR to HR, which can lead to overfitting to a specific simulation setup or produce overly smoothed speech spectrograms (Ling et al., 2015; Wang and Wang, 2021). This phenomenon is manifested in predictive models that perform well for certain types of downsampled signals but poorly for other downsampled signals, let alone challenging real-world scenarios. Generative learning produces a data distribution rather than a one-to-one mapping for the SR task, making it potentially more robust in realistic situations. However, its current SR performance is not competitive with mapping-based models.

To address these challenges, we propose a novel two-stage approach that combines the strengths of predictive and generative learning. Some recent studies have explored hybrid approaches combining predictive and generative models, although they focus on speech separation or universal speech restoration (Andreev et al., 2022; Wang and Wang, 2023; Lemercier et al., 2023; Lutati et al., 2023) rather than speech super-resolution. For instance, Lutati et al. (2023) present a novel approach to speech separation that combines a deterministic separation model with a pretrained diffusion model, using a learned weighting scheme in the frequency domain, and report superior performance in various multi-speaker scenarios. Lemercier et al. (2023) introduce a hybrid approach combining predictive and diffusion models for speech restoration. It employs the output of the predictive model to guide score-based diffusion, leading to reduced artifacts and computational costs.

Our two-stage approach is designed for the SR task, with architectural choices optimized for bandwidth extension. More specifically, the first stage employs predictive learning to generate coarsely enhanced speech, which is fed as the conditioner to the second stage that employs a diffusion learning model. In addition, we introduce novel modifications to the diffusion process and inference, tailored for speech SR. Furthermore, we employ a repainting technique that enhances model generalization across different simulation methods. As a result, the proposed approach achieves better SR performance with robustness to recorded data. In addition, we record and publicize speech corpora by employing the same microphone at multiple native sampling rates.

The contributions of this study are three-fold. First, we propose a novel DNN architecture for speech SR that integrates predictive learning and generative learning. Second, the proposed approach is demonstrated to achieve better SR performance and robustness than other strong baselines. Third, we record and publicize multi-resolution datasets to facilitate future speech SR research in the community.

The rest of the paper is organized as follows. Section 2 describes the background of diffusion models and the SR task. Section 3 presents the proposed approach. Section 4 describes the experimental setup and multi-resolution data recording. Evaluation results and comparisons are presented in Section 5. Finally, concluding remarks are given in Section 6.

2. Background

2.1. Denoising diffusion probabilistic model

A standard denoising diffusion probabilistic model (DDPM) (Ho et al., 2020), consists of a forward process and a reverse process, as depicted in Fig. 1. The DDPM with T time steps contains latent variables $x_0, x_1, \ldots, x_T$, all with the same dimensionality.

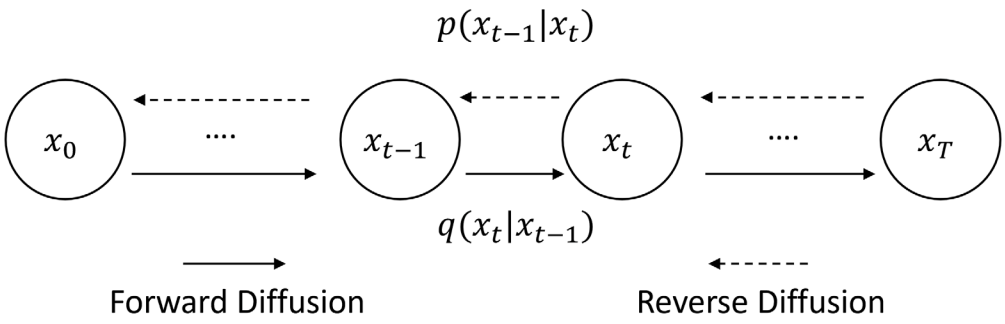

Fig. 1. Illustration of the diffusion processes in a denoising diffusion probabilistic model.

2.1.1. Forward process

During the forward process, the joint distribution is defined as a Markov chain from x_0 to x_T,

$$q(x_1, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{1}$$

At each step a Gaussian noise is added, and the noise is controlled by the noise schedule parameter $\sigma(t)$ and γ, where

$$\sigma(t)^2 = \frac{\sigma_{min}^2((\sigma_{max}/\sigma_{min})^{2t} - e^{-2\gamma t})log(\sigma_{max}/\sigma_{min})}{\gamma + log(\sigma_{max}/\sigma_{min})} \tag{2}$$

is a fixed noise schedule that manipulates the Gaussian distribution in the forward process. A Gaussian noise is added to the previous sample x_{t-1} during each step, gradually converting x_0 to an isotropic Gaussian distribution $p(x_T) = \mathcal{N}(0, I)$. Since each transition step follows a Gaussian distribution, x_t can be directly computed from the distribution of x_0 by marginalizing $x_1, \ldots, x_{t-1}$, i.e.,

$$q(x_t|x_0) = \mathcal{N}(x_t; \mu(x_0, y, t), \sigma(t)^2). \tag{3}$$

As derived in Richter et al. (2023), the mean μ has a closed form solution,

$$\mu(x_0, y, t) = e^{-\gamma t}x_0 + (1 - e^{-\gamma t})y. \tag{4}$$

2.1.2. Reverse process

The reverse process in contrast is to progressively remove the added noise from the latent variable x_T to obtain x_0. In the standard DDPM, this is done by subtracting an estimated noise from a DNN model. During training, the model is trained to predict the added noise given the diffused signal and the noise embedding information. Other methods, such as the score-based diffusion, model the forward process using a stochastic differential equation (SDE). Then a DNN model is used to estimate the score, which is the gradient of the log probability density. Numerical differential equation solvers are utilized to solve a reverse-time SDE using the estimated scores. Repeating these steps gradually derives the target data x_0 from the whitened Gaussian noise x_T.

2.1.3. Conditional reverse process

Inspired by Kawar et al. (2022), Salimans and Ho (2022), Croitoru et al. (2023), instead of estimating noises or scores, we employ a DNN model to directly estimate the target data x_{0t} at each time step t. This approach has two major advantages. First, unlike the standard DDPM which requires either $\mathcal{L}_1$ or $\mathcal{L}_2$ norm on the noise estimation, we can employ an established training objective in the target domain and directly compute a loss on the target data x_0. Second, after the training is done, we have a model that effectively estimates $\hat{x}_{0t}$ given the time step t, which achieves faster inference compared with the standard DDPM process.

2.2. Speech SR problem formulation

Given a low-resolution speech segment s^{lr} at a sampling rate of fs^{lr}, the objective of speech super-resolution is to reconstruct a high-resolution speech signal s^{hr} at a higher sampling rate of fs^{hr} such that $fs^{hr} > fs^{lr}$, thereby restoring high-frequency components. The number fs^{hr}/fs^{lr} is denoted as the *upsampling ratio*. To reconstruct the high-resolution signal, we first employ a predictive learning model f that takes the upsampled low-resolution signal s^{inp} as input, which is obtained by applying interpolation on the LR signal s^{lr}. With the model parameters denoted as ϕ, the model produces the corresponding reconstruction s^{pred}:

$$s^{pred} = f(\phi, s^{inp}), \tag{5}$$

where s^{pred} is the estimated SR signal obtained from the first stage. The predicted SR output is then utilized along with the original LR input as the conditioner for the second stage of diffusion-based learning. We denote the parameters of the diffusion-based model as θ, and the reverse diffusion process as g. During the inference stage, the diffusion model aims to reconstruct the final output from the whitened Gaussian noises given the conditioners, which is formulated as

$$\hat{s} = g(\theta, s^{pred}, s^{inp}). \tag{6}$$

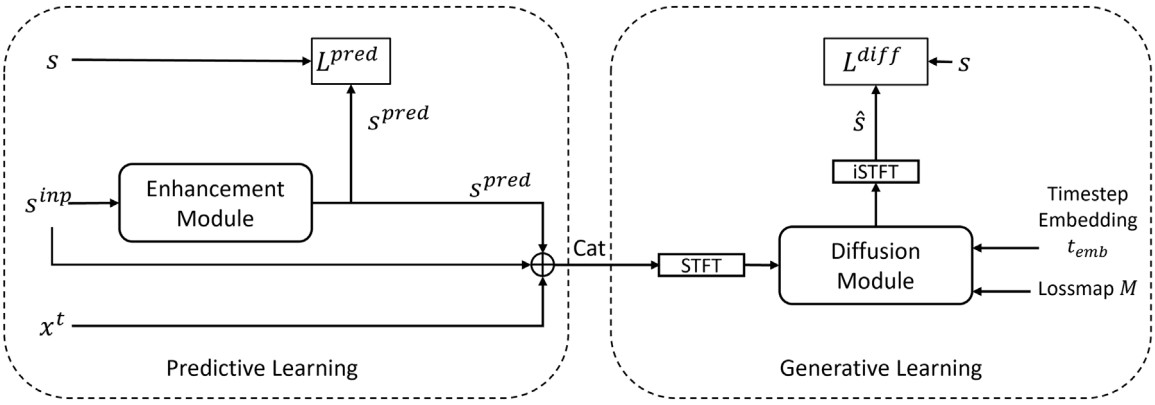

Fig. 2. Two-stage model diagram that depicts the training procedure of the predictive learning stage and the generative learning stage.

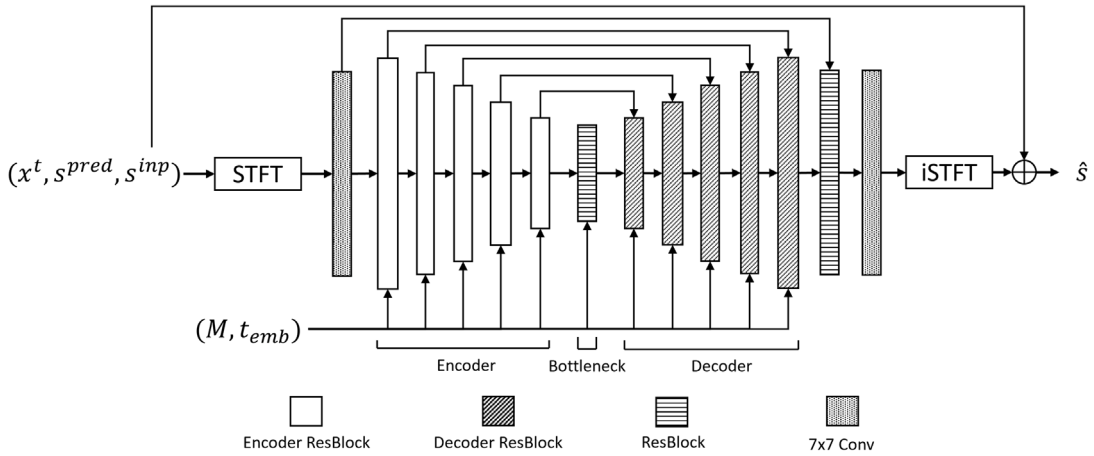

Fig. 3. Diagram of the proposed attentive residual convolutional network (ARCN) as the diffusion module, and "ResBlock" denotes an attentional residual block.

3. Model description

We introduce a two-stage DNN model for speech super-resolution, as shown in Fig. 2. In the first stage, we utilize a time-domain dual-path attentional recurrent network (DPARN) (Pandey and Wang, 2020) for predictive learning. Its output conditions the second-stage diffusion learning. For generative learning, we propose an attention-based residual convolutional network (ARCN) and estimate the added noise during the diffusion process. During inference, we use the enhanced speech from the first stage as the conditioner and progressively generate the missing bandwidth through the reverse diffusion process. Detailed descriptions are given in the following subsections.

3.1. DNN architectures

3.1.1. Predictive learning

For predictive learning, we employ the dual-path attentional recurrent neural network (DPARN) (Pandey and Wang, 2020) as the enhancement module. DPARN is an enhanced version of the dual-path recurrent neural network, which was originally introduced for time-domain speaker separation (Luo et al., 2020). In a dual-path network, utterances are divided into overlapping chunks and processed sequentially using intra-chunk and inter-chunk recurrent neural networks (RNNs) to efficiently handle the given time series. This approach reduces the sequence length for RNN modeling, improving training efficiency. Furthermore, it allows for a relatively small frame shift in time-domain speech processing, resulting in a significant performance gain (see Table 1 of Luo et al., 2020, and Section VI.E of Pandey and Wang, 2022).

DPARN further incorporates inter-chunk and intra-chunk attention, resulting in notable improvements in performance and training efficiency. To reduce the computational burden, we make two modifications: (i) adopting residual connections instead of dense connections between RNN modules and (ii) limiting the number of DPARN blocks to two. More details of DPARN design can be found in Pandey and Wang (2020).

3.1.2. Diffusion module

We implement the diffusion module using an ARCN, which is based on UNet. The UNet architecture has been extended in recent years to incorporate attention and residual blocks in diffusion-related models (Ho et al., 2020; Richter et al., 2023). However, directly adopting the standard UNet is unsuitable for audio signals, as it is designed to process inputs of fixed resolution, such as fixed-sized images or fixed-length audio segments. This limitation arises because the receptive field is constrained by the fixed size of its convolutional kernels. In contrast, ARCN incorporates an RNN as its bottleneck, which encodes sequences of varying lengths by maintaining a fixed-length hidden state vector.

For diffusion training, ARCN requires five inputs: the diffused signal x_t, the enhanced speech signal s^{pred} from the enhancement module, the LR speech input s^{inp}, a lossmap M, and the time step embedding t_{emb} (see Fig. 3). Prior to feeding these inputs to ARCN, we apply Short-Time Fourier Transform (STFT) to x_t, s^{pred}, and s^{inp}, concatenating their real and imaginary parts as separate convolutional channels. The time step embedding and lossmap serve as local conditioners for the residual block. The time step embedding is generated using Fourier embedding in Vaswani et al. (2017) followed by two linear layers. The lossmap M consists of binary indicators (0 and 1), where time–frequency (T–F) units marked as ones are those to be super-resolved.

Our ARCN adopts an encoder–decoder architecture with a bottleneck block, as illustrated in Fig. 3. The encoder and decoder have symmetric designs, incorporating skip connections to enhance feature reusability and combat the vanishing gradient issue. Instead of standard convolution layers, we build the encoder, decoder and bottleneck with attentional residual blocks to improve learning of temporal dependencies between frames. Residual connections within each block facilitate feature reuse and accelerate training convergence. Operating on complex spectrogram vectors obtained from STFT, the model concatenates real and imaginary parts into a 3-dimensional representation ($C \times T \times F$), where C represents the number of convolutional channels (6 channels in this study), T the number of time frames and F the size of the feature dimension. The input is initially transformed into 64 convolutional channels using the 7×7 input convolution layer. The 64-channel input undergoes processing through 5 attentional residual blocks in the encoder, 1 bottleneck block, and then 5 attentional residual blocks in the decoder. The decoder output undergoes another residual block, and is then processed with a 7×7 convolutional layer. Finally, the speech SR estimate, $\hat{s}$, is derived by adding the LR input s^{inp} with the inverse STFT (iSTFT) applied to the output.

Fig. 4 shows the details of an attention residual block in the encoder. It comprises two residual layers, a lightweight attention layer, and a downsampling operation. The output of the attention layer is added to its input using a residual connection. To prevent the checkerboard artifact (Odena et al., 2016), we implement the downsampling operation using finite impulse response filters (Zhang, 2019). Each residual layer in our model comprises two convolution layers, and utilizes the time step embedding and lossmap as local conditioners. Each convolution layer consists of a 2-dimensional (2-D) convolution, group normalization (Wu and He, 2018) and Sigmoid Linear Unit (SiLU) nonlinearity (Elfwing et al., 2018). A convolution layer has a kernel size of 1×3 (time $\times$ frequency), zero-padding of size 1 applied to each side of the feature maps along the frequency dimension, and 64 output channels. The downsampled lossmap is incorporated through its multiplication with the output of the first convolution layer after a pointwise convolution. A residual block in the decoder is similar, but uses an upsampling operation instead of downsampling. As illustrated in Fig. 4, the bottleneck block has a different arrangement of residual layers.

Fig. 5 shows the architecture of an attention layer, which operates on an input of shape $C \times T \times F$. The attention layer transforms the input using three pointwise convolution layers to derive the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) tensors of $E \times T \times F$, $E \times T \times F$, and $C \times T \times F$ respectively, where E is set to 5 to reduce the computational burden. They are subsequently rearranged into 2-D matrices $T \times E \cdot F$, $T \times E \cdot F$, and $T \times C \cdot F$, respectively. The correlation scores between the query and key matrices are calculated through matrix multiplication $\mathbf{W} = \mathbf{Q}\mathbf{K}^T$, where superscript T denotes matrix transpose. This produces $\mathbf{W}$ with dimensions $T \times T$. The attention score is then computed by applying a Softmax function to $\mathbf{W}$, and multiplied with the value matrix $\mathbf{V}$:

$$\mathbf{A} = \text{Softmax}(\mathbf{W})\mathbf{V}. \tag{7}$$

The result, denoted as the attentive output ($\mathbf{A}$), emphasizes important time frames. This mechanism enables ARCN to capture long-term temporal dependencies. To preserve information from the original input, a residual connection is employed to merge the attentive output with the input as shown in Fig. 5.

3.2. Loss functions

We simultaneously optimize two modules through joint training. The predictive learning module utilizes a complex-domain loss (denoted as $\mathcal{L}^{pred}$) proposed in Wang et al. (2020) to predict SR speech. This loss function is defined in terms of $\mathcal{L}_1$ differences in magnitude, and real and imaginary spectrograms:

$$\mathcal{L}^{pred} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\Big[\,\|S^{pred}(t,f)| - |S(t,f)\|+ \\ (|S_r^{pred}(t,f) - S_r(t,f)| + |S_i^{pred}(t,f) - S_i(t,f)|)\,\Big]. \tag{8}$$

Here, T and F denote the total number of time frames and frequency bins, indexed by t and f, respectively. S^{pred} and S represent the STFTs of the predicted and original speech signals, respectively. Subscripts r and i refer to the real and imaginary parts of the complex vectors, respectively, and $|\cdot|$ measures the magnitude. To compute the STFT vectors S^{pred} and S, the waveforms are first divided into 32-ms frames with a frame shift of 8 ms and then multiplied by a Hanning window.

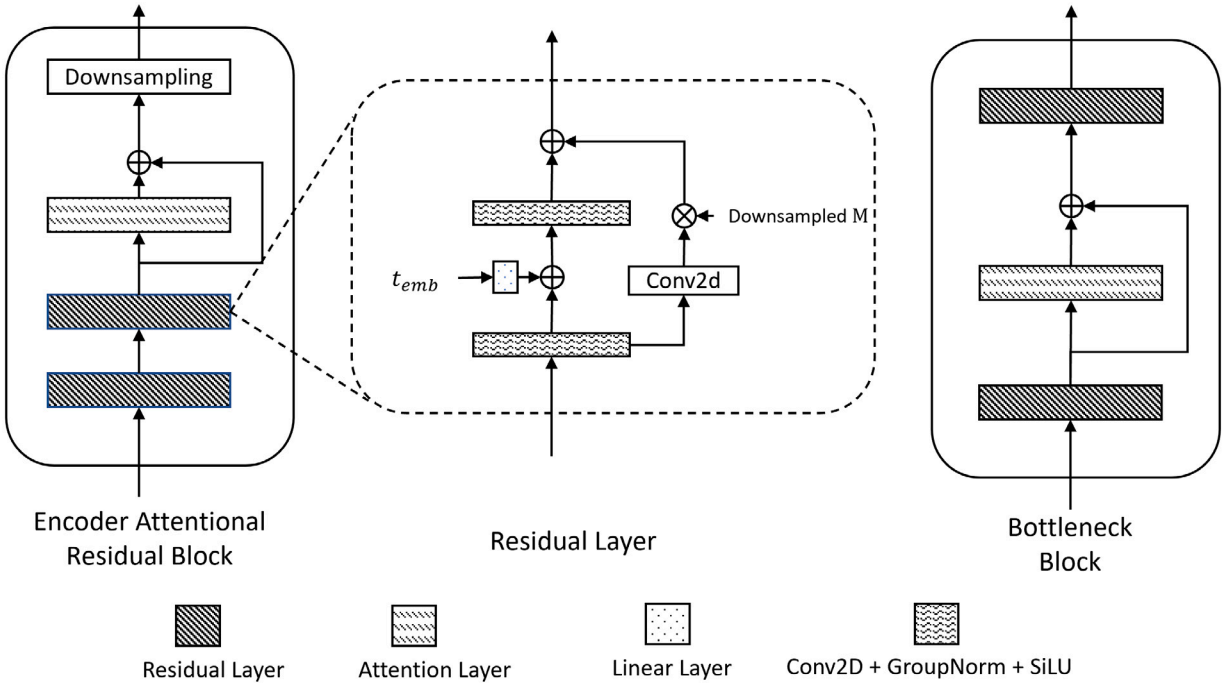

Fig. 4. Diagrams showing the detailed design of an attentional residual block within the ARCN encoder.

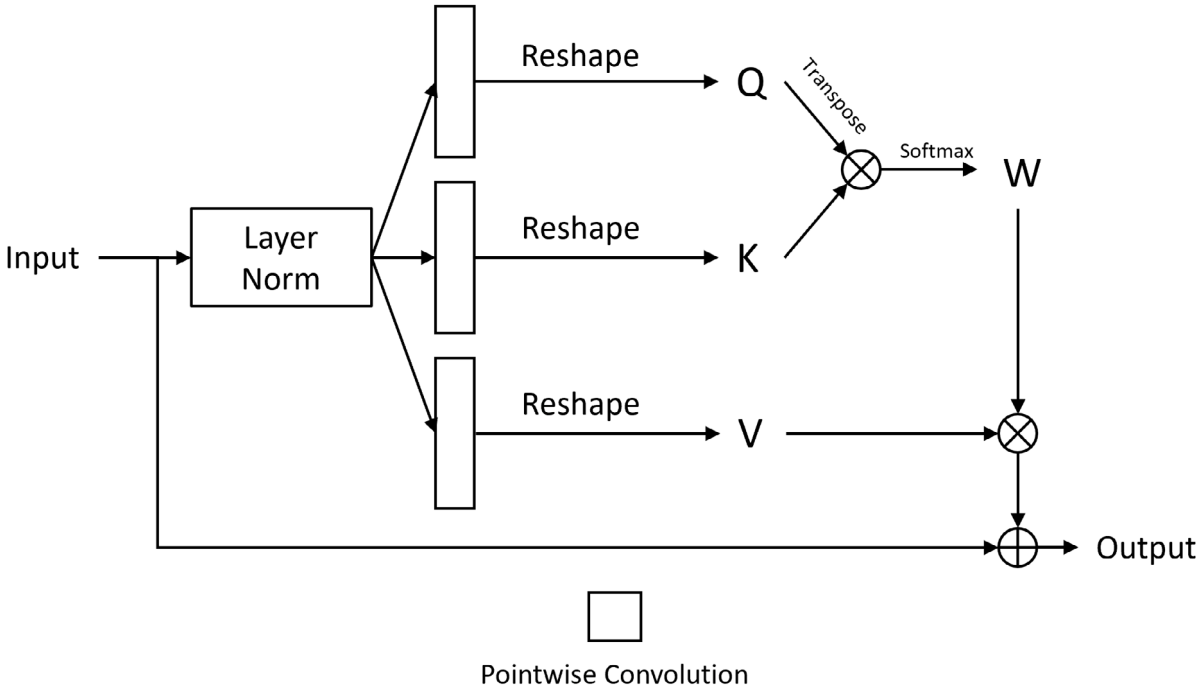

Fig. 5. The architecture of an attention layer.

For the diffusion module, we adopt a T–F loss proposed in Wang and Wang (2020):

$$\mathcal{L}^{T}(\hat{s}, s) = \frac{1}{N}\sum_{n=1}^{N}|\hat{s}(n) - s(n)|$$

$$\mathcal{L}^F(\hat{S}, S) = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F} \| \hat{S}(t,f)| - |S(t,f)|| \\ \mathcal{L}^{diff} = \alpha\mathcal{L}^T + (1-\alpha)\mathcal{L}^F. \tag{9}$$

The time domain loss $\mathcal{L}^T$ is the $\mathcal{L}_1$ difference between SR and HR signals ($\hat{s}$ and s), where n to indexes time samples. The frequency domain loss $\mathcal{L}^F$ compares the magnitudes of the STFTs of SR and HR signals. The T–F loss combines $\mathcal{L}^T$ and $\mathcal{L}^F$, and we empirically set α to 0.85.

The overall loss for joint training of the two modules is given by:

$$\mathcal{L} = \mathcal{L}^{pred} + \lambda(\tau)\mathcal{L}^{diff}. \tag{10}$$

We use a time varying $\lambda(\tau) = 1/(e^{\tau} - 1)$ to control the weight of the diffusion loss $\mathcal{L}^{diff}$ at different diffusion steps τ (Lu et al., 2023).

3.3. Training and inference

Algorithm 1. Training Procedure

(1) $s^{inp} = Upsample(Downsample(Filtering(s^{hr})))$
(2) $s^{pred} = f(\phi, s^{inp})$
(3) Sample $t \sim Uniform(1, 2, ..., T)$
Sample $z \sim \mathcal{N}(0, I)$
(4) $x_t \leftarrow \mu(s^{hr}, s^{inp}, t) + \sigma(t)^2 z$
$\hat{s} \leftarrow g(\theta, x_t, s^{pred}, s^{inp}, t)$
(5) Take gradient descent on
$\mathcal{L}^{pred} + \lambda(t)\mathcal{L}^{diff}$

The training procedure of the proposed approach is summarized in Algorithm 1. Firstly we simulate the upsampled LR signal as the input speech to the model. Then, the coarsely enhanced speech s^{pred} obtained by the prediction network f is employed as the conditioner for the diffusion model. In steps 3–5, we perform diffusion-based training. We randomly sample a time step t, and use it to generate the time step embedding. The model g will estimate the target SR speech at the given diffusion step. Gradient descent serves to optimize the predictive model f and the generative model g jointly.

Algorithm 2. Inference Procedure

procedure RESAMPLE(x)
Return Upsample(Downsample(Filtering(x)))
end procedure
(1) $s^{inp} \leftarrow Upsample(s^{lr})$
(2) $s^{pred} \leftarrow f(\phi, s^{inp})$
(3) Initialize $x_{0T} \leftarrow s^{pred}$
for $t = T-1, T-2, ..., 0$ **do**
Sample $z \sim \mathcal{N}(0, I)$
$x_t \leftarrow \mu(x_{0t+1}, s^{inp}, t) + \sigma(t)^2 z$
$x_{0t} \leftarrow g(\theta, x_t, s^{pred}, s^{inp})$
$x_{0t} \leftarrow s^{inp} + (x_{0t} - Resample(x_{0t}))$ ▷ Repainting
end for
return $\hat{s} \leftarrow x_{00}$

Algorithm 2 outlines the inference process, featuring two key modifications from the standard DDPM reverse process. First, we initialize the diffused input signal x_{0T} using the predicted speech s^{pred} obtained in the first stage, instead of white Gaussian noise. Second, we leverage the low-frequency components of s^{inp} to guide the reverse diffusion process. Utilizing given narrowband components, we inject LR information into the sampling process by replacing the low-frequency region with the ground truth. Specifically, we apply the same set of downsampling and upsampling operations employed during training (the *Resample* operation defined in the algorithm), i.e.,

$$x_{0t} = s^{inp} + (x_{0t} - Resample(x_{0t})). \tag{11}$$

Here we essentially combine the high-frequency part of the SR prediction x_{0t} at time step t, with the low-frequency part of the input signal s^{inp}. This process is similar to the repainting technique in the image domain (Lugmayr et al., 2022), and applied in DiffWave based SR (Yu et al., 2022). Different from repainting, our approach operates in the time domain. Compared to Yu et al. (2022), we do not subtract a gradient term in each reverse diffusion step. In addition, we initialize the reverse diffusion process with the output of the prediction module s^{pred}, which is referred to as the shallow diffusion mechanism in Liu et al. (2022b). Our inference algorithm not only yields improved objective scores but also expedites the diffusion process.

4. Experimental setup

4.1. Simulated dataset

We conduct our experiments on the VCTK dataset (Veaux et al., 2017), which contains 44 h of speech recordings from 108 speakers, and follow the task design in Han and Lee (2022). We exclusively utilize the *mic1* recordings from the VCTK dataset for our experiments, while omitting the recordings of speakers p280 and p315. Our model is trained on the first 98 speakers, validated on the excluded two speakers, and tested on the remaining 8 speakers. We preprocess the utterances as follows. First, we resample them to 16 kHz if they have a higher sampling rate. Second, we normalize them to have zero mean and unit variance; as suggested in Wang and Wang (2021) this helps with generalization to untrained data. To generate LR speech, we first convolve the HR speech with an eighth-order Chebyshev type I lowpass filter, and then subsample it to a desired sampling frequency. We then use cubic spline interpolation to upsample generated LR signals to match the lengths of the corresponding HR signals before feeding them to the proposed model.

4.2. Recorded datasets

Although experiments on simulated data show superior performance for DNN-based models, speech SR performance in actual applications may be significantly worse due to the mismatch between synthetic and recorded data. In a previous study (Wang and Wang, 2021), we found that trained SR models are very sensitive to recording channels and downsampling schemes.

To overcome the limitations of simulated data, we created recordings of two datasets, Device and Produced Speech (DAPS) (Mysore, 2014) and VCTK (Veaux et al., 2017), each at different sampling rates. The DAPS dataset comprises carefully aligned speech recordings, including studio recordings and the corresponding versions captured on common consumer devices such as tablets and smartphones, in real-world environments. We choose the clean-room subset, which includes approximately 4 and 1/2 h of high-quality data, corresponding to an average of 14 min from each of the 20 speakers included in the dataset. We carefully select 200 segments (approximately corresponding to utterances) from both male and female speakers, ensuring that each segment is separated by periods of silence. To ensure compatibility with the D/A converter, we resample the original recordings to 48 kHz. For the VCTK dataset, we utilize the 96 kHz version[1] and choose 200 utterances from a diverse group of 24 speakers.

The recordings took place in the anechoic chamber of The Ohio State University Department of Speech and Hearing Science, with the dimensions of 13 × 7.5 × 8 ft (length × width × height), measured from wedge tips. The depth of wedges is 2 ft. Fig. 6 shows the recording setup and the anechoic chamber. We conducted three separate recording sessions at the three lowest sampling rates of 8, 16, and 32 kHz of a HyperX SoloCast microphone. During the recording, we played back the speech utterances from a Windows PC. The D/A converter employed was an M-Audio Mobile Pre DAC operating at a 48 kHz sampling rate with 24-bit precision. The analog signal was then delivered to a Mackie HR824 powered loudspeaker, renowned for its ‘‘ruler-flat’’ frequency response (+/−1.5 dB, 37 Hz – 20 kHz). The speech signal maintained an approximate level of 65 dBA. We recorded speech signals using the HyperX SoloCast microphone placed 1 m directly in front of the loudspeaker (see Fig. 6). The microphone featured onboard A/D conversion, allowing us to capture the digital signal using another Windows PC. Inside the anechoic chamber, only the loudspeaker and recording microphone were present. All other equipment was located outside the chamber. We had remote control of the apparatus, enabling us to carry out all three recording sessions without re-entering the chamber. Throughout the recording sessions, the lights and fans in the chamber remained off to minimize any external interference. The recorded datasets are publicly accessible at https://web.cse.ohio-state.edu/$\protect$\relax\svsim$$wang.77/pnl/corpus/Heming/RecordedSR.html.

4.3. Baseline methods

We compare with several baseline methods for speech SR. The first baseline is a signal processing method that upsamples LR signals to the desired sampling rate using cubic spline interpolation (McKinley and Levine, 1998). In addition, we compare with three strong DNN-based SR methods, which include a predictive learning based model (our previous study, denoted as SR-AECNN Wang and Wang, 2021), a diffusion-based generative model (NuWave2 Han and Lee, 2022) and a vocoder based model (NVSR Liu et al., 2022a). SR-AECNN utilizes a convolutional autoencoder to perform cross-domain speech SR, and leverages cross-domain training objectives. NuWave2, is an extended version of NuWave, which utilizes a vanilla diffusion model to convert time-domain LR speech to SR speech. NuWave2 is trained with multiple resolutions and performs reasonably well. NVSR comprises a mel-bandwidth extension stage, and a vocoder waveform synthesis and post-processing stage. A pretrained neural vocoder upsamples a low-dimensional mel-spectrogram to a high-resolution waveform for speech SR.

4.4. Evaluation metrics

To measure the SR performance, we use three objective metrics: scale-invariant signal-to-noise ratio (SISNR), log-spectral distance (LSD) (Gray and Markel, 1976), and PESQ for wideband speech (Beerends et al., 2002). SISNR is a time-domain metric that measures

[1] available at https://datashare.ed.ac.uk/handle/10283/2774.

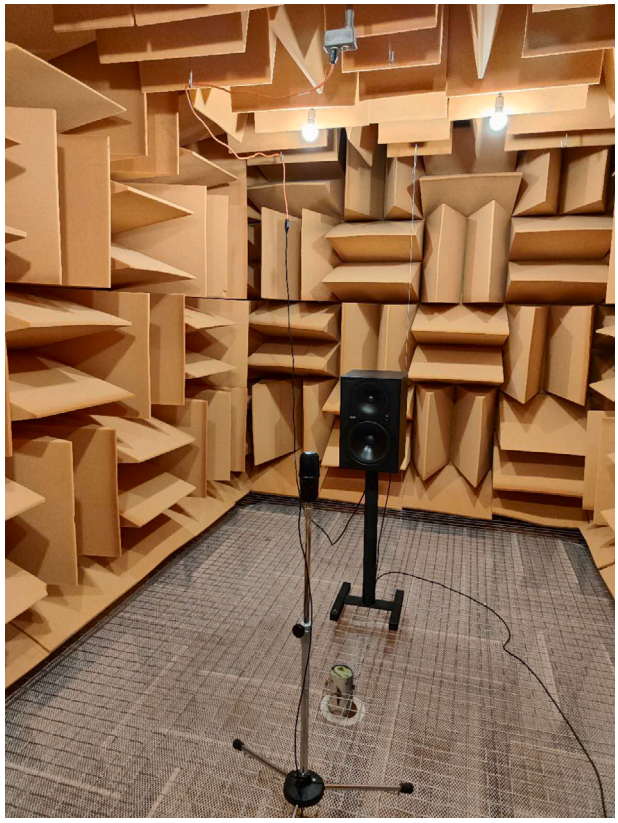

Fig. 6. Photo of the anechoic chamber and recording setup used for real SR data acquisition.

signal power relative to noise power in dB. LSD is a frequency-domain metric that calculates the logarithmic distance between two magnitude spectra in dB, as follows:

$$\mathrm{LSD}(S,\hat{S}) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{F}\sum_{f=1}^{F}[\log_{10}\frac{|S(t,f)|^2}{|\hat{S}(t,f)|^2}]^2} \tag{12}$$

LSD will be 0 dB when two spectra are identical, which is the minimum possible distance. PESQ for wideband speech is a standard metric of perceptual speech quality, and higher PESQ means better listening quality.

4.5. Training setup

We use an Adam optimizer (Kingma and Ba, 2015) to train our model with a batch size of 32 utterances for 100 epochs. The initial learning rate is 0.0006, and is halved if the validation loss does not improve for three consecutive epochs. Gradient clipping is employed with a maximum value of 1.0 to prevent gradient explosion. We also apply the exponential moving average to stabilize the training process. Within each batch, 4 s of each utterance are randomly selected, and shorter utterances are padded with zeros to ensure that all input features are of the same size. Zero-padded parts are discarded in loss calculations. All models are trained on two NVIDIA Volta V100 32 GB GPUs, and the DataParallel module from PyTorch (Paszke et al., 2019) is used to evenly distribute the batch to the two GPUs during each training step. In diffusion training, we configure noise scheduling with $\sigma_{min} = 0.05$, $\sigma_{max} = 0.5$, and $\gamma = 1.5$ following the setup in Richter et al. (2023). Training involves 1000 steps, and inference is capped at 10 steps.

5. Evaluations and comparisons

5.1. Evaluation results and comparisons on VCTK

Table 1 reports the SR performance of the proposed model from 8 kHz to 16 kHz on the VCTK test set, as well as those of four comparison baselines. As shown in the table, our approach achieves the best SISNR of 21.15 dB, and PESQ of 4.05. Its LSD performance also surpasses those of SR-AECNN and NuWave2, demonstrating the advantages of combined predictive and generative modeling. In addition, all DNN-based approaches show significant LSD improvement over the signal processing baseline. Although

Table 1
Speech SR performance from 8 to 16 kHz for proposed and baseline methods on the VCTK dataset.

	SISNR↑	LSD↓	PESQ↑
Cubic Upsampling	18.99	2.72	3.44
SR-AECNN	20.18	0.88	3.72
NuWave2	19.17	1.17	2.53
NVSR	16.06	**0.78**	2.71
Proposed	**21.15**	0.81	**4.05**

Table 2
Single-stage versus two-stage models.

	SISNR↑	LSD↓	PESQ↑
Cubic Upsampling	18.99	2.72	3.44
Stage 1 only	20.89	0.82	3.87
Stage 2 only	19.12	0.85	3.71
Stage 1 + Stage 2	**21.15**	**0.81**	**4.05**

Table 3
Different inference strategies on the VCTK dataset.

	SISNR↑	LSD↓	PESQ↑
Cubic Upsampling	18.99	2.72	3.44
Initialize with White Noise	19.87	0.87	3.73
Shallow Diffusion	19.89	0.86	3.78
Repainting (proposed)	**21.15**	**0.81**	**4.05**

Table 4
Evaluation on LR speech generated by different filters.

	Chebyshev			Bessel		
	SISNR↑	LSD↓	PESQ↑	SISNR↑	LSD↓	PESQ↑
Cubic Upsampling	18.99	2.72	3.44	17.21	2.89	3.33
SR-AECNN	20.18	0.88	3.72	−8.84	2.43	1.07
NuWave2	19.17	1.17	2.53	10.83	1.86	1.77
NVSR	16.06	**0.78**	2.71	13.64	**1.05**	2.77
Proposed	**21.15**	0.81	**4.05**	**17.43**	1.23	**3.52**

LSD is commonly used for SR evaluations, it does not strongly correlate with the perceptual quality of SR speech, and its use should be complemented by other speech quality metrics such as PESQ. For NVSR, its SISNR is 2.93 dB worse than the signal processing baseline, and its PESQ drops from 3.44 to 2.71. NVSR obtains superior performance in LSD, but the worst PESQ results. This outcome may be attributed to the misalignment of vocoder-synthesized speech and ground truth signal. In summary, the proposed model achieves the overall best SR performance.

To further understand the effects of two-stage training, we conduct an ablation study to compare the two-stage model to single-stage models on the SR task. The findings of this analysis are presented in Table 2. Two single-stage models correspond to predictive learning only with DPARN (stage 1 only) and diffusion-based learning only with ARCN (stage 2 only). The ablation results indicate that the predictive learning model outperforms the generative model, especially in SISNR and PESQ. The proposed two-stage approach yields consistently better results than the single-stage models.

We also conduct an ablation study on different inference strategies, with results presented in Table 3. We test three strategies: (1) Vanilla diffusion, which starts from Gaussian noise and gradually recovers SR speech during the reverse diffusion process; (2) Shallow diffusion, starts with the diffused LR signal for easier processing; and (3) the proposed repainting technique, which treats LR speech as ground truth and iteratively refines the diffused output during each inference step. Shallow diffusion shows a slight improvement over vanilla diffusion. Our repainting technique incorporates LR speech information in each inference step and significantly improves SR performance.

5.2. Robustness evaluation

Our previous study (Wang and Wang, 2021) reveals that predictive learning DNNs are sensitive to specific downsampling filters used in training. Now, we assess the filter robustness of the proposed model and other baselines for the SR task from 8 kHz to 16 kHz on the VCTK test set. Specifically, we train the models on Chebyshev filter-simulated data and test them on LR speech generated by a Chebyshev filter (matched) and a fifth-order Bessel filter (mismatched). As shown in Table 4, our proposed model exhibits good performance in both matched and mismatched conditions. Compared to cubic spline upsampling, our model demonstrates consistent improvements across SISNR, LSD, and PESQ metrics. SR-AECNN shows severe performance degradation in the mismatched condition.

Table 5
Speech SR performance from 8 to 16 kHz and from 8 to 32 kHz on recorded data.

	8 kHz → 16 kHz						8 kHz → 32 kHz					
	DAPS			VCTK			DAPS			VCTK		
	SISNR↑	LSD↓	PESQ↑	SISNR↑	LSD↓	PESQ↑	SISNR↑	LSD↓	PESQ↑	SISNR↑	LSD↓	PESQ↑
Cubic Upsampling	12.68	3.34	3.23	12.53	3.37	3.06	12.40	4.39	2.62	10.97	4.93	2.33
SR-AECNN	−10.51	2.09	1.29	10.45	1.37	1.98	−5.06	1.41	1.32	6.61	1.50	1.52
NuWave2	12.17	1.17	2.84	11.89	1.06	2.41	10.43	1.38	1.45	8.83	1.48	1.35
NVSR	10.87	1.15	1.92	10.79	1.07	2.02	10.69	1.15	2.04	8.98	1.06	2.08
Proposed	**12.34**	**1.09**	**3.68**	**13.97**	**1.02**	**3.64**	**12.45**	**1.19**	**3.07**	**11.50**	**1.23**	**2.93**

NuWave2 also shows a considerable performance drop. NVSR, on the other hand, has a relatively stable performance across the two conditions as it leverages a vocoder to synthesize the SR speech. It has the best LSD scores, but its SISNR and PESQ scores are even worse than the cubic spline baseline.

5.3. Evaluation results and comparisons on recorded data

Finally, we evaluate on the two recorded datasets described in Section 4.2. We present the results in Table 5 for two SR tasks: from 8 kHz to 16 kHz, and from 8 kHz to 32 kHz. All the models are trained on the VCTK dataset with the corresponding upsampling ratios. For 32 Hz audio evaluation, we employ wideband PESQ and STOI for objective metric calculation. Comparing to the results in Table 1, all SR methods show reduced performance on the recorded VCTK for the task of 8 kHz → 16 kHz, reflecting the mismatch between simulated and recorded data. For both SR tasks, our model shows a consistent improvement compared to the cubic upsampling method. In addition, the proposed model outperforms other baselines in all metrics except for NVSR which produces better LSD scores for the 8 kHz → 32 kHz task. The SR-AECNN model yields large performance deterioration, particularly on the DAPS dataset. In contrast, generative and vocoder based models deliver relatively stable SR performance. It is worth noting that, while NuWave2 and NVSR produce strong LSD scores, their SISNR and PESQ scores are not only substantially worse than our model but also than the signal processing baseline. This evaluation suggests that our proposed approach overcomes the limitation of training with simulated data, and can be applied in real-world environments.

Finally, we provide audio demos of various speech SR techniques at https://whmrtm.github.io/uploads/GPSR_demo.html.

6. Conclusion

In this study, we have introduced a two-stage model that combines predictive and generative approaches for speech SR. Our model consistently outperforms other strong baselines in both simulated and realistic datasets. The model is capable of generating high-frequency speech in mismatched conditions, making it suitable for real-world applications. An additional contribution is the collection and publication of recorded data using the same microphone at multiple native sampling rates, to facilitate SR progress in real-world environments.

CRediT authorship contribution statement

Heming Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Eric W. Healy:** Writing – review & editing, Supervision, Data curation. **DeLiang Wang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by an NIDCD, United States (R01 DC012048) grant, the Ohio Supercomputer Center, United States, and the Pittsburgh Supercomputing Center, United States (under NSF grant ACI-1928147).

Data availability

Data will be made available on request.

References

Abel, J., Fingscheidt, T., 2019. Sinusoidal-based lowband synthesis for artificial speech bandwidth extension. IEEE/ACM Trans. Audio Speech Lang. Process. 27, 765–776.

Albahri, A., Rodriguez, C.S., Lech, M., 2016. Artificial bandwidth extension to improve automatic emotion recognition from narrow-band coded speech. In: Procedings of ICSPCS. pp. 1–7.

Andreev, P., Alanov, A., Ivanov, O., Vetrov, D., 2022. Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement. 1, (2), arXiv:2203.13086.

Beerends, J.G., Hekstra, A.P., Rix, A.W., Hollier, M.P., 2002. Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part II: psychoacoustic model. J. Audio Eng. Soc. 50, 765–778.

Botinhao, C.V., Carlos, B., Caloba, L.P., Petraglia, M.R., 2006. Frequency extension of telephone narrowband speech signal using neural networks. In: Proceedings of CESA. pp. 1576–1579.

Chan, C.F., Hui, W.K., 1996. Wideband re-synthesis of narrowband CELP-coded speech using multiband excitation model. In: Proceeding of ICSLP, vol. 1, pp. 322–325.

Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M., 2023. Diffusion models in vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell..

De Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111, 1917–1930.

Elfwing, S., Uchibe, E., Doya, K., 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw. 107, 3–11.

Eskimez, S.E., Koishida, K., Duan, Z., 2019. Adversarial training for speech super-resolution. IEEE J. Sel. Top. Signal Process. 13, 347–358.

Füllgrabe, C., Baer, T., Stone, M.A., Moore, B.C., 2010. Preliminary evaluation of a method for fitting hearing aids with extended bandwidth. Int. J. Audiol. 49, 741–753.

Gray, A., Markel, J., 1976. Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. 24, 380–391.

Gu, Y., Ling, Z.H., 2017. Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension. In: Proceedings of INTERSPEECH. pp. 1123–1127.

Gu, Y., Ling, Z.H., Dai, L.R., 2016. Speech bandwidth extension using bottleneck features and deep recurrent neural networks.. In: Proceedings of INTERSPEECH. pp. 297–301.

Gupta, A., Shillingford, B., Assael, Y., Walters, T.C., 2019. Speech bandwidth extension with WaveNet. In: Proceedings of WASPAA. pp. 205–208.

Han, S., Lee, J., 2022. NU-Wave 2: A general neural audio upsampling model for various sampling rates. In: Proceedings of INTERSPEECH. pp. 4401–4405.

Haws, D., Cui, X., 2019. CycleGAN bandwidth extension acoustic modeling for automatic speech recognition. In: Proceedings of ICASSP. pp. 6780–6784.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851.

Kataria, S., Villalba, J., Moro-Velázquez, L., Żelasko, P., Dehak, N., 2024. Time-domain speech super-resolution with GAN based modeling for telephony speaker verification. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 1736–1749.

Kawar, B., Elad, M., Ermon, S., Song, J., 2022. Denoising diffusion restoration models. Adv. Neural Inf. Process. Syst. 35, 23593–23606.

Kim, S.B., Lee, S.H., Choi, H.Y., Lee, S.W., 2024. Audio super-resolution with robust speech representation learning of masked autoencoder. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 1012–1022.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of ICLR.

Kontio, J., Laaksonen, L., Alku, P., 2007. Neural network-based artificial bandwidth expansion of speech. IEEE Trans. Audio Speech Lang. Process. 15, 873–881.

Kuleshov, V., Enam, S.Z., Ermon, S., 2017. Audio super-resolution using neural nets. In: Proceedings of ICLR (Workshop Track).

Lee, J., Han, S., 2021. NU-Wave: A diffusion probabilistic model for neural audio upsampling. In: Proceedings of INTERSPEECH. pp. 1634–1638.

Lemercier, J.M., Richter, J., Welker, S., Gerkmann, T., 2023. Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration. In: Proceedings of ICASSP. p. 5.

Li, K., Lee, C.H., 2015. A deep neural network approach to speech bandwidth expansion. In: Proceedings of ICASSP. pp. 4395–4399.

Li, S., Villette, S., Ramadas, P., Sinder, D., 2018. Speech bandwidth extension using generative adversarial networks. In: Proceedings of ICASSP. pp. 5029–5033.

Lim, T.Y., Yeh, R.A., Xu, Y., Do, M.N., Hasegawa-Johnson, M., 2018. Time-frequency networks for audio super-resolution. In: Proceedings of ICASSP. pp. 646–650.

Lin, J., Wang, Y., Kalgaonkar, K., Keren, G., Zhang, D., Fuegen, C., 2021. A two-stage approach to speech bandwidth extension.. In: Proceedings of INTERSPEECH. pp. 1689–1693.

Ling, Z.H., Ai, Y., Gu, Y., Dai, L.R., 2018. Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension. IEEE/ACM Trans. Audio Speech Lang. Process. 26, 883–894.

Ling, Z.H., Kang, S.Y., Zen, H., Senior, A., Schuster, M., Qian, X.J., Meng, H.M., Deng, L., 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. IEEE Signal Process. Mag. 32, 35–52.

Liu, H., Choi, W., Liu, X., Kong, Q., Tian, Q., Wang, D., 2022a. Neural vocoder is all you need for speech super-resolution. In: Proceedings of INTERSPEECH. pp. 4227–4231.

Liu, J., Li, C., Ren, Y., Chen, F., Zhao, Z., 2022b. DiffSinger: Singing voice synthesis via shallow diffusion mechanism. In: Proceedings of AAAI, vol. 36, pp. 11020–11028.

Liu, B., Tao, J., Wen, Z., Li, Y., Bukhari, D., 2015. A novel method of artificial bandwidth extension using deep architecture. In: Proceedings of INTERSPEECH. pp. 2598–2602.

Lu, Y., Wang, Z., Bal, G., 2023. Understanding the diffusion models by conditional expectations. arXiv:2301.07882.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L., 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of CVPR. pp. 11461–11471.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: Proceedings of ICASSP. pp. 46–50.

Lutati, S., Nachmani, E., Wolf, L., 2023. Separate and diffuse: Using a pretrained diffusion model for better source separation. In: Proceedings of ICLR.

Makhoul, J., Berouti, M., 1979. High-frequency regeneration in speech coding systems. In: Proceedings of ICASSP, vol. 4, pp. 428–431.

McKinley, S., Levine, M., 1998. Cubic Spline Interpolation, vol. 45, College of the Redwoods, pp. 1049–1060.

Milner, B., Shao, X., 2002. Speech reconstruction from Mel-frequency cepstral coefficients using a source-filter model. In: Proceedings of ICSLP. pp. 2421–2424.

Mysore, G.J., 2014. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. IEEE Signal Process. Lett. 22, 1006–1010.

Nakamura, K., Hashimoto, K., Oura, K., Nankaku, Y., Tokuda, K., 2014. A Mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech. In: Proceedings of INTERSPEECH. pp. 2494–2498.

Nour-Eldin, A.H., Kabal, P., 2008. Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech. In: Proceedings of INTERSPEECH. pp. 1501–1504.

Nour-Eldin, A.H., Kabal, P., 2011. Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrowband speech. In: Proceedings of INTERSPEECH. p. r1188.

Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. Distill 1, e3.

Pandey, A., Wang, D.L., 2020. Dual-path self-attention RNN for real-time speech enhancement. arXiv:2010.12713.
Pandey, A., Wang, D.L., 2022. Self-attending RNN for speech enhancement to improve cross-corpus generalization. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 1374–1385.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of NeuralIPS. pp. 8026–8047.
Prenger, R., Valle, R., Catanzaro, B., 2019. WaveGlow: A flow-based generative network for speech synthesis. In: Proceeding of ICASSP. pp. 3617–3621.
Pulakka, H., Alku, P., 2011. Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband Mel spectrum. IEEE Trans. Audio Speech Lang. Process. 19, 2170–2183.
Richter, J., Welker, S., Lemercier, J.M., Lay, B., Gerkmann, T., 2023. Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Trans. Audio Speech Lang. Process. 2351–2364.
Sadasivan, J., Mukherjee, S., Seelamantula, C., 2016. Joint dictionary training for bandwidth extension of speech signals. In: Proceedings of ICASSP. pp. 5925–5929.
Salimans, T., Ho, J., 2022. Progressive distillation for fast sampling of diffusion models. arXiv:2202.00512.
Sulun, S., Davies, M.E., 2020. On filter generalization for music bandwidth extension using deep neural networks. IEEE J. Sel. Top. Signal Process. 15, 132–142.
Unno, T., McCree, A., 2005. A robust narrowband to wideband extension system featuring enhanced codebook mapping. In: Proceedings of ICASSP. pp. I–805.
Van Eeckhoutte, M., Folkeard, P., Glista, D., Scollie, S., 2020. Speech recognition, loudness, and preference with extended bandwidth hearing aids for adult hearing aid users. Int. J. Audiol. 59, 780–791.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5998–6008.
Veaux, C., Yamagishi, J., MacDonald, K., 2017. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR).
Wang, H., Wang, D.L., 2020. Time-frequency loss for CNN based speech super-resolution. In: Proceedings of ICASSP. pp. 861–865.
Wang, H., Wang, D.L., 2021. Towards robust speech super-resolution. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 2058–2066.
Wang, H., Wang, D.L., 2023. Cross-domain diffusion based speech enhancement for very noisy speech. In: Proceedings of ICASSP. p. 5.
Wang, Z.Q., Wang, P., Wang, D.L., 2020. Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 1778–1787.
Wu, Y., He, K., 2018. Group normalization. In: Proceedings of ECCV. pp. 3–19.
Yu, C.Y., Yeh, S.L., Fazekas, G., Tang, H., 2022. Conditioning and sampling in variational diffusion models for speech super-resolution. arXiv:2210.15793.
Zhang, R., 2019. Making convolutional networks shift-invariant again. In: Proceedings of ICML. pp. 7324–7334.