Audio Super-Resolution With Robust Speech Representation Learning of Masked Autoencoder

Seung-Bin Kim, Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*—This paper proposes Fre-Painter, a high-fidelity audio super-resolution system that utilizes robust speech representation learning with various masking strategies. Recently, masked autoencoders have been found to be beneficial in learning robust representations of audio for speech classification tasks. Following these studies, we leverage these representations and investigate several masking strategies for neural audio super-resolution. In this paper, we propose an upper-band masking strategy with the initialization of the mask token, which is simple but efficient for audio super-resolution. Furthermore, we propose a mix-ratio masking strategy that makes the model robust for input speech with various sampling rates. For practical applicability, we extend Fre-Painter to a text-to-speech system, which synthesizes high-resolution speech using low-resolution speech data. The experimental results demonstrate that Fre-Painter outperforms other neural audio super-resolution models.**

***Index Terms*—Audio super-resolution, bandwidth extension, self-supervised learning, masked autoencoder, audio synthesis.**

I. Introduction

AUDIO super-resolution is a process that aims to reconstruct the high-frequency information of low-resolution audio, thereby generating high-resolution audio. This process, also known as bandwidth extension, enhances the audio quality. In the early stages of this field, the primary research focus was the extension of bandwidth from narrowband to wideband by leveraging deep neural networks [1], [2], [3]. As the demand for high-quality audio has increased across various applications, research on audio super-resolution has been actively conducted, leading to significant improvements in the reconstruction of high-quality, high-resolution audio from low-resolution audio. This progress has resulted in the development of models capable of achieving higher resolution in audio super-resolution such as [4], [5].

However, many of these studies have limitations related to the fixed sampling rate of the input audio signal. In practice, audio to be upsampled is not guaranteed to have a consistent sampling rate. Consequently, multiple models need to be trained to handle audio at various sampling rates and perform upsampling. To address these problems, [6] and [7] proposed methods that use a series of the processed audio as input during the training process. This series of steps typically involves sequentially applying a low-pass filter with random ripples and orders to all data at each training step, followed by downsampling the audio signal and upsampling it. However, these operations can slightly decrease the training process efficiency. Moreover, a gap remains between high-resolution and upsampled audio, particularly in the high-frequency range. Therefore, we introduce various masking strategies to efficiently train the system to perform audio super-resolution for a variety of sampling rates and to enhance its ability to reconstruct high-frequency information.

Recently, a masked autoencoder (MAE) with Vision Transformers (ViT) [8] has exhibited its superiority in robust representation learning in computer vision [9]. Moreover, several studies [10], [11] have introduced MAE to speech classification tasks. Although MAE excels in various downstream tasks, it has not yet garnered attention in audio generation tasks such as neural upsampling, neural vocoding, and other speech synthesis tasks.

In this paper, we propose Fre-Painter, a robust neural audio super-resolution system that utilizes robust speech representation learning using MAE and several masking strategies. We utilize a large-scale dataset to pre-train the MAE to generate more robust representations. At the fine-tuning stage, we jointly train the neural vocoder as a generator with the pre-trained MAE encoder. We also introduced upper-band masking strategy, which masks the high-frequency components of the input Mel-spectrogram, to enable efficient training for audio super-resolution. In addition, to ensure the robustness of the model across various sampling rates, we introduced a mix-ratio masking strategy, which randomly determines the masking ratio during training. Consequently, our model can synthesize high-quality, high-resolution audio from low-resolution audio, irrespective of its sampling rate.

Our model, Fre-Painter, derives its name from its ability to reconstruct the missing high-frequency components of a Mel-spectrogram converted from low-resolution audio, similar to the process of painting. The experimental results demonstrate that Fre-Painter excels in synthesizing high-quality, high-resolution audio waveform from a Mel-spectrogram converted from low-resolution audio.

Manuscript received 31 July 2023; revised 30 November 2023; accepted 21 December 2023. Date of publication 5 January 2024; date of current version 10 January 2024. This work was supported in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) under Grant 2019-0-00079, in part by Artificial Intelligence Graduate School Program (Korea University) under Grant 2021-0-02068, and in part by Artificial Intelligence Innovation Hub. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marc Delcroix. *(Corresponding author: Seong-Whan Lee.)*

The authors are with the Department of Artificial Intelligence, Korea University, Seoul 02841, South Korea (e-mail: sb-kim@korea.ac.kr; sh_lee@korea.ac.kr; hayeong@korea.ac.kr; sw.lee@korea.ac.kr).

Digital Object Identifier 10.1109/TASLP.2023.3349053

Furthermore, we extended Fre-Painter to a two-stage text-to-speech (TTS) system. Specifically, by adopting neural audio super-resolution as a vocoder, we developed a TTS system that can synthesize high-resolution speech from text using an acoustic model which is trained on a low-resolution speech dataset. The contributions of this study can be summarized as follows:

- We propose Fre-Painter, a robust neural audio super-resolution system that can synthesize high-quality, high-resolution audio from low-resolution audio of various sampling rates from 2 kHz to 24 kHz.
- We successfully integrate an MAE into audio super-resolution and investigate various masking strategies for speech synthesis tasks.
- We extend Fre-Painter to serve as a vocoder in a two-stage TTS system, enabling the synthesis of high-resolution speech from text with an acoustic model trained on a low-resolution speech dataset.

The audio samples and their corresponding codes are available at https://frepainter.github.io/demo. We encourage the reader to experience the demo audio samples.

II. Related Work

A. Audio Super-Resolution

Audio super-resolution systems have undergone substantial advancements in recent years. With the advancements in deep learning, it has been incorporated into audio super-resolution [12], [13], [14], [15] superseding conventional statistical approaches. TFNet [16] has demonstrated that modeling both the time and frequency domains can significantly enhance audio super-resolution. Temporal FiLM [17] introduces feature-wise modulation into audio super-resolution. [18] proposes a lightweight model based on SEANet [19], and TUNet [20], a low-complexity transformer-aided U-Net, also effectively reduces inference time due to its lightweight architecture.

However, the majority of early research has focused on bandwidth extension from narrowband to wideband, with upsampling to higher resolutions yet to be realized. WSRGlow [21] integrates Oord [22] with Glow, and [23] combines Oord [22] with generative adversarial network (GAN). Both models generate high-resolution audio at 48 kHz. VoiceFixer [6] propose a two-stage super-resolution approach, comprising an analysis step with U-Net [24] and a synthesis step with a neural vocoder [25]. NVSR [26] introduces post-processing through the replacement of lower frequencies.

To enhance the quality of audio, GANs have been explored in audio processing tasks and have also been introduced in audio super-resolution [27], [28], [29]. NU-GAN [30], [31], and AERO [32] propose phase and magnitude modeling based on GANs. BEHM-GAN [33] proposes a method that combines a time-frequency domain generator with multiple time domain discriminators. mdctGAN [34] introduces the modified discrete cosine transform domain in audio super-resolution.

Given the impressive performance of diffusion probabilistic models in audio generation tasks [35], [36], NU-Wave [5] incorporates the diffusion probabilistic model into the audio super-resolution task. NU-Wave 2 [7] enhances the generation of harmonics by introducing short-time Fourier convolution. UDM+ [37] proposes a sampling algorithm that improves the reconstruction of lower frequencies. AudioSR [38] proposes a model capable of robustly performing audio super-resolution across versatile audio types. However, these models require iterative steps to generate high-quality audio, which inevitably leads to a reduction in inference speed. Furthermore, a discernible gap persists between the upsampled audio and high-resolution original audio, particularly in the high-frequency components.

B. Self-Supervised Learning of Audio Representation

Learning general representations through self-supervised learning with large-scale unlabeled data and fine-tuning the model with labeled data has yielded impressive results in various fields of natural language processing and computer vision.

In the field of speech processing, a vast amount of unlabeled data is available for utilization. Initially, self-supervised learning was primarily implemented for tasks such as emotion recognition [39] and speaker recognition [40]. However, it has gradually expanded to other tasks as well. [41], [42], [43] propose mask prediction-based self-supervised learning methods and extend them to speech recognition.

In addition to speech recognition, self-supervised learning has been utilized in other tasks. [44] utilizes representations from self-supervised learning as additional linguistic representations to bridge the information gap between text and speech for TTS systems. [45] and [46] successfully enhance the performance by incorporating self-supervised learning into speaker recognition and verification tasks, respectively. [47] leverages self-supervised learning to extract discrete units from unlabeled language pair data instead of scarce labeled language pair data for speech-to-speech translation. [48] has demonstrated that the utilization of scaled-up pre-training data to learn universal speech representations results in improvements across various speech processing tasks.

Audio-MAE [10] extends the masked autoencoder to self-supervised learning from audio spectrograms and introduces it in audio and speech classification tasks. Based on the observation that the high-frequency components of the Mel-spectrogram converted from low-resolution audio are missing, similar to the masked inputs in MAE, we aim to incorporate a Masked Autoencoder into our audio super-resolution system.

C. Neural Audio Synthesis

In recent years, significant progress has been achieved in audio synthesis using neural networks. It began with WaveNet [49], which is an autoregressive neural network. This model outperform traditional concatenative and parametric methods. However, it had a drawback of being slow during the inference stage because it require predicting a tremendous number of audio samples autoregressively.

To address these issues, more efficient approaches to audio synthesis models, such as WaveRNN [50] and Parallel WaveNet [51], have been introduced. With the remarkable achievements of GAN-based models in the field of computer

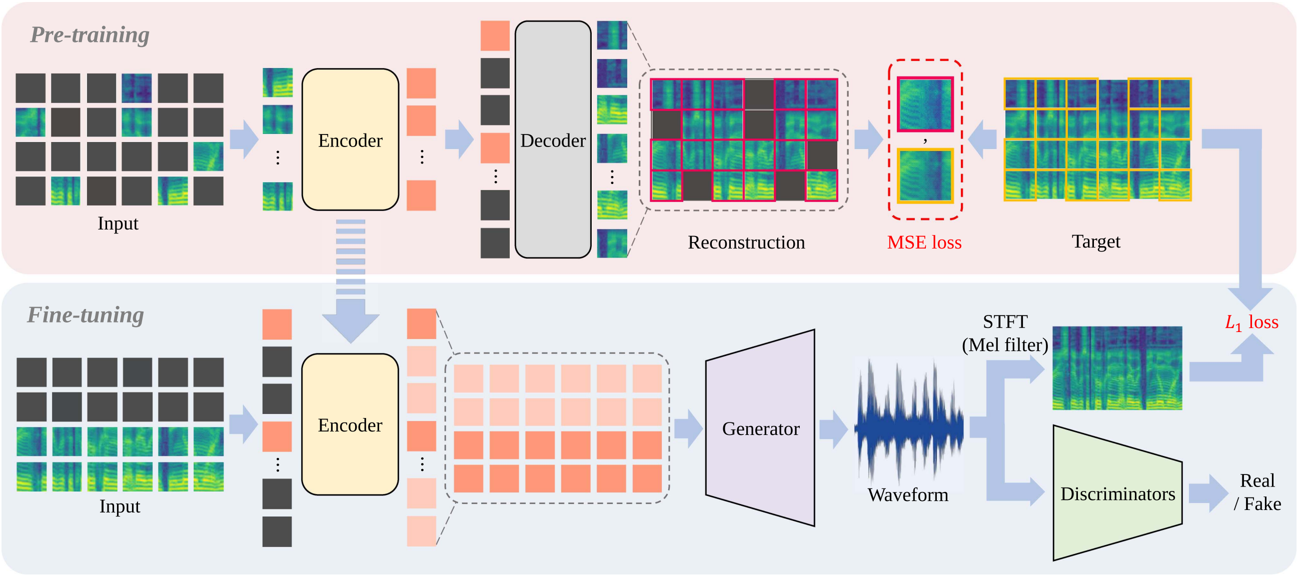


Fig. 1. Overall framework of Fre-Painter. Initially, we pre-train the masked autoencoder using a random masking strategy. Subsequently, the generator is jointly trained with the pre-trained encoder of masked autoencoder. For audio super-resolution, we adopt an upper-band masking strategy during the fine-tuning.

vision, research on GAN-based models has been conducted in the audio synthesis domain. MelGAN [52] propose a multi-scale architecture with discriminators that operate on different audio scales to learn features for different audio frequency ranges. [53] propose a multi-resolution Short-time Fourier transform (STFT) loss that can effectively capture the time-frequency distribution of a realistic audio signal. HiFi-GAN [54] has demonstrated the importance of periodic pattern modeling in audio for improving sound quality and propose multi-period discriminators that handles periodic audio signals by incorporating several sub-discriminators. UnivNet [55] introduce multi-resolution spectrogram discriminators to address the over-smoothing problem, and EnCodec [56] extend it by incorporating a complex-valued STFT. [57], [58], [59] propose generators that consider the audio properties. For a fair comparison, we employ HiFi-GAN as the baseline model for the audio super-resolution to evaluate the effectiveness of our methods.

III. Method

In this paper, we propose a robust neural audio super-resolution system called Fre-Painter. For robust speech latent representation, we utilize a self-supervised learning model based on the MAE. Following the pre-training of the MAE with a substantial volume of data, we jointly train the encoder of pre-trained and the neural vocoder, incorporating frequency-domain masking strategy for audio super-resolution. Fig. 1 illustrates the framework of Fre-Painter. Further details are presented in the following subsections.

A. Masked Autoencoder

The high-frequency component of the Mel-spectrogram, which is converted from low-resolution audio, is devoid of information. In an audio super-resolution task, this vacant component must be reconstructed based on information from the low-frequency component. This process is similar to the manner in which MAE reconstructs masked patches to their original values. Therefore, we adopt a MAE framework. Furthermore, by pre-training with a substantial volume of data, our model can extract a robust speech representation from the audio.

1) Patch Embeddings: The input audio waveform is converted into a Mel-spectrogram and thereafter divided into regular non-overlapping patches as in [9]. These patches are embedded by two-dimensional (2D) convolution and flattened to patch embeddings. Subsequently, 2D sinusoidal positional embeddings are added to the patch embeddings to enable the model to capture the time- and frequency-wise structure of the Mel-spectrogram.

2) Encoder: For self-supervised learning, we partition the patches into unmasked and masked subsets, and then apply masking specifically to the masked subset. Detailed masking strategies are described in Subsection III-C. We construct encoder following the ViT [8], [60]. During the pre-training, encoding is performed solely on the patches of the unmasked subset, thereby reducing the computational cost and training time.

3) Decoder: The decoder is constructed similarly to the encoder. However, we aim to use only the encoder in downstream task and discard the decoder. Therefore, to empower the encoder to extract a more robust representation from the Mel-spectrogram, we purposefully design the decoder with fewer layers and a reduced hidden dimension compared with the encoder, making it shallower and narrower.

Given that we encode only the unmasked subset, it is necessary to fill the sections corresponding to the masked subset with mask tokens. Hence, we combine the encoder output with trainable mask tokens. Subsequently, we restore the order of the patches to their original sequence and add positional embedding for the decoder input. Finally, we add a linear projection layer above the decoder blocks to reconstruct the Mel-spectrogram.

4) Objective: In the pre-training, the MAE is optimized with the objective of minimizing the mean square error (MSE) loss between the reconstructed and ground truth Mel-spectrogram.

$$z_u = Enc(x_u), \tag{1}$$

$$\hat{x} = Dec(z_u, t_m), \tag{2}$$

$$\mathcal{L}_{pre} = \mathbb{E}_{(\hat{x},x)}[||\hat{x} - x||_2], \tag{3}$$

where x_u, t_m, and x denote the unmasked subset, trainable mask token, and Mel-spectrogram of the ground truth, respectively. $\mathcal{L}_{pre}$ is calculated only for the masked subset.

B. Audio Super-Resolution

For the downstream task, audio super-resolution, we discard the decoder from the pre-trained MAE and retain only the encoder. In contrast to pre-training, which uses only unmasked patches for an input, the encoder for downstream task takes the entire patches x, including the masked patches, to reduce the mismatch with the inference scenario, which uses a degraded Mel-spectrogram.

To generate the upsampled audio waveform, Fre-Painter uses the generator G of HiFi-GAN [54]. The generator synthesizes a waveform from the output of encoder. For the reconstruction loss, we calculate L_1 distance between Mel-spectrograms as follows:

$$z = Enc(x), \tag{4}$$

$$\hat{y} = G(z), \tag{5}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{(\hat{y},x)}[||\phi(\hat{y}) - x||_1], \tag{6}$$

where ϕ denotes the STFT with applying Mel filter.

Discriminator D, which distinguishes between the ground truth waveform and waveform synthesized via a generator, is used for adversarial training. We use the multi-period discriminator (MPD) [54] and multi-scale discriminator (MSD) [52]. In addition, we add feature matching loss [52] that measures the L_1 distance between the intermediate features of the discriminator as follows:

$$\mathcal{L}_{adv}(D;G) = \mathbb{E}_{(y,z)}\left[(D(y) - 1)^2 + (D(G(z)))^2\right], \tag{7}$$

$$\mathcal{L}_{adv}(G;D) = \mathbb{E}_{z}\left[(D(G(z)) - 1)^2\right], \tag{8}$$

$$\mathcal{L}_{fm}(G;D) = \mathbb{E}_{(y,z)}\left[\sum_{i=1}^{T} \frac{1}{N_i}||D^i(y) - D^i(G(z))||_1\right], \tag{9}$$

where D^i, N_i, and T denote the i-th layer feature map, number of units, and number of layers in the discriminator, respectively. The total loss for Fre-Painter can be expressed as follows:

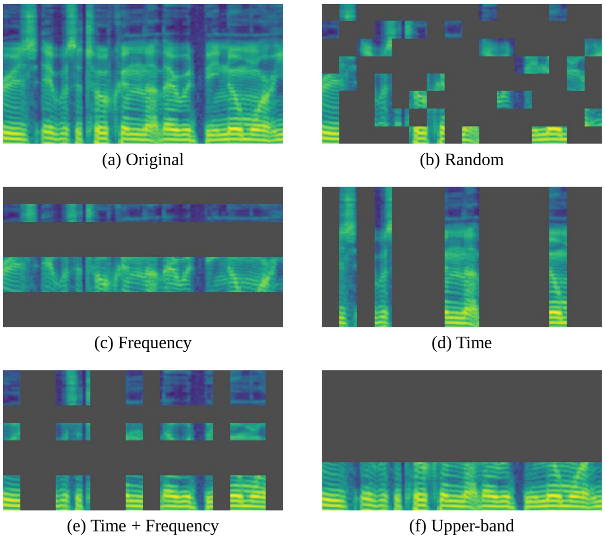
$$\mathcal{L}_G = \mathcal{L}_{adv}(G;D) + \lambda_{fm}\mathcal{L}_{fm}(G;D) + \lambda_{rec}\mathcal{L}_{rec}, \tag{10}$$

$$\mathcal{L}_D = \mathcal{L}_{adv}(D;G), \tag{11}$$

where λ_{fm} and λ_{rec} are loss weights.

C. Masking Strategies

We introduce efficient and effective masking strategies for training an audio super-resolution system. Illustrative examples of masking can be found in Fig. 2.



(a) Original

(b) Random

(c) Frequency

(d) Time

(e) Time + Frequency

(f) Upper-band

Fig. 2. Illustration of masking strategies on Mel-spectrogram.

1) Masking Strategies for Pre-Training: The masking strategies during the pre-training follows the strategies in [10]. A random masking strategy involves masking some of the patches randomly. Given that a Mel-spectrogram carries information according to time and frequency of the audio, frequency and time masking strategies mask frequency-wise and time-wise patches, respectively. A time + frequency masking strategy mask both directions simultaneously.

2) Upper-Band Masking: We propose an upper-band masking strategy that masks above the target frequency bands of the Mel-spectrogram during training. In this way, these masks can delude the model to regard the masked Mel-spectrogram as the degraded Mel-spectrogram, which is transformed from audio with a low resolution.

3) Mix-Ratio Masking: To robustly take the input of various sampling rates, we propose mix-ratio upper-band masking. We hypothesize that masking with a certain ratio results in trade-off between upsampling and reconstruction ability according to the sampling rate. To consider both abilities, we utilize mix-ratio masking for the input with various sampling rates. A detailed analysis is presented in Section V-A2.

4) Mask Token Initialization: In the downstream task, we initialize the mask token with a certain value to make the masked Mel-spectrogram similar to the Mel-spectrogram converted from low-resolution audio. We set the initial value of the mask token as the value averaged over the upper-band of the Mel-spectrogram converted from low-resolution audio.

IV. Experiments

A. Experimental Setup

1) Datasets: We pre-train the MAE on the LibriTTS dataset [61], which is a large multi-speaker speech synthesis dataset containing 586 hours of audio for 2,456 speakers. For the downstream task, experiments are conducted on the VCTK dataset [62], which contains 41 hours of audio for 108 speakers. We divide the 108 speakers into 100 speakers for the training dataset and the remaining eight speakers for the test dataset. For accurate evaluation, audio samples are trimmed of silence using a 20 dB threshold. To generate low-resolution input audio for the test dataset used in evaluating audio super-resolution, the audio samples were processed using an order eight Chebyshev Type I low-pass filter with 0.05 dB of ripple. They were then downsampled to the input sampling rate and subsequently upsampled to the input sampling rate of the models. Audio super-resolution experiments are conducted at various input sampling rates, including an extremely low sampling rate of 2 kHz, and multiple rates of 4, 8, 12, 16, and 24 kHz.

2) Preprocessing: For the purpose of using the Mel-spectrogram converted from 24 kHz audio as input, we downsample all the training data to 24 kHz, excluding the LibriTTS dataset which originally has a sampling rate of 24 kHz. Then the audio is transformed into a linear spectrogram using a STFT. The Fourier transform, window size, and hop size are set as 2048, 1200, and 300 frames, respectively. Finally, a Mel filter with 128 bins is applied to the linear spectrogram. For loss calculation, we used audio with a sampling rate of 24 kHz and 48 kHz as the target audio for the model that upsamples to 24 kHz and 48 kHz, respectively.

3) Training: We train Fre-Painter using the AdamW optimizer [63] with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and weight decay $\lambda = 0.01$ with two NVIDIA RTX A6000 GPUs. For the learning rate scheduling, we use the exponential decay with factor $\gamma = 0.999^{1/8}$ and an initial learning rate of 2×10^{-4}. We pre-train and fine-tune the model for 100 k steps and 200 k steps, respectively. The models designed for upsampling to 24 kHz and 48 kHz consist of 22.91 M and 22.95 M parameters, respectively. To reduce the training time and memory usage during fine-tuning, we adopt the windowed generator training [64], which generates a waveform from only a segment of latent representation z. Segmentation is performed by randomly extracting 32 frames from the latent representations. For discriminator and loss calculation, audio corresponding to the segmented latent representations is also extracted.

4) Implementation Details: We use a vanilla ViT with 12 layers and 256 dimensions as the encoder and eight layers and 196 dimensions as the decoder. We use a patch size of 16×16. The generator and discriminators follow the same architecture as HiFi-GAN v1 [54] except for the upsample rates and kernel sizes. For the 24 kHz upsampling model, these are [5,5,4,3] and [11,11,8,7] respectively, whereas for the 48 kHz upsampling model, they are [5,5,4,3,2] and [11,11,8,7,4] respectively. We use only the first sub-discriminator of MSD.

B. Evaluation Metrics

1) Log-Spectral Distance: For objective evaluation, we measure the log-spectral distance (LSD) between the upsampled and ground truth audio. To evaluate the distortion of high-frequency bands, we calculate a high-frequency LSD (LSD-HF), which is the spectral distance between the high-frequency bands. The range of the high-frequency bands is set to more than half of the input audio sampling rate, according to the Nyquist theorem. A low-frequency LSD (LSD-LF) is also used to measure the reconstruction of low-frequency bands. The formulation for calculating the LSD is as follows:

$$LSD(\hat{S}, S) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K}\sum\nolimits_{k=1}^{K} Dis_{t,k}^2}, \tag{12}$$

$$LSD_{HF}(\hat{S}, S) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{K - b_{nf} + 1}\sum\nolimits_{k=b_{nf}}^{K} Dis_{t,k}^2}, \tag{13}$$

$$Dis_{t,k} = \log_{10}\left(\frac{\hat{S}(t,k)}{S(t,k)}\right), \tag{14}$$

where T, K, S, $\hat{S}$, and b_{nf} denote the time frames, total number of bins in the spectrogram, ground truth audio spectrogram, upsampled audio spectrogram, and bin that corresponds to the Nyquist frequency, respectively.

2) Virtual Speech Quality Objective Listener: We also employ the virtual Speech quality objective listener (ViSQOL) [32], [65], [66] for objective evaluation. In order to compare the upsampled audio with the original audio, we adopt the audio mode, which enables the analysis up to the higher bands of 24 kHz frequency.

3) Mean Opinion Score: For subjective evaluation, we conduct a mean opinion score (MOS) test with 20 listeners using the crowd-sourced method via Amazon Mechanical Turk. The listeners listen to 100 random audio samples and rated their quality. To affirm statistical validity, we conduct an one-way analysis of variance test on the MOS results. By confirming a p-value of less than 0.05, we validate the statistical significance of our experimental results.

4) ABX Preference Test: We conduct an ABX preference test to directly compare the two models. Listeners listen to audio from each model and choose which one they prefer. The order of the models is randomized, and if there is no preference, they can choose X. The test is carried out with 10 listeners on 30 random audio samples.

5) Pronunciation Accuracy: To evaluate the pronunciation accuracy of TTS results, we use an automatic speech recognition (ASR) models to measure the phoneme error rate (PER) and word error rate (WER). For predicting phonemes and words, the ASR models used are wav2vec 2.0 [41] and Whisper [67], respectively.

6) Inference Speed: We calculate the inference speeds to enable comparisons across models. The results are presented in two ways: speed denotes the number of audio waveform samples that can be generated per second, and real-time denotes how many seconds of audio can be generated in one second.

V. Results

A. Analysis on Masking Strategies

1) Pre-Training: To investigate speech representation with respect to masking strategies during pre-training, we first train the MAE with various masking strategies. Subsequently, we fine-tune the model to reconstruct the Mel-spectrogram without masking. To evaluate the performance of the pre-trained model,

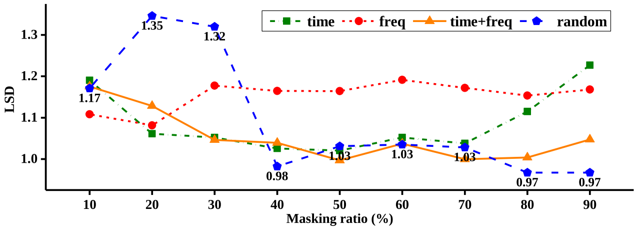
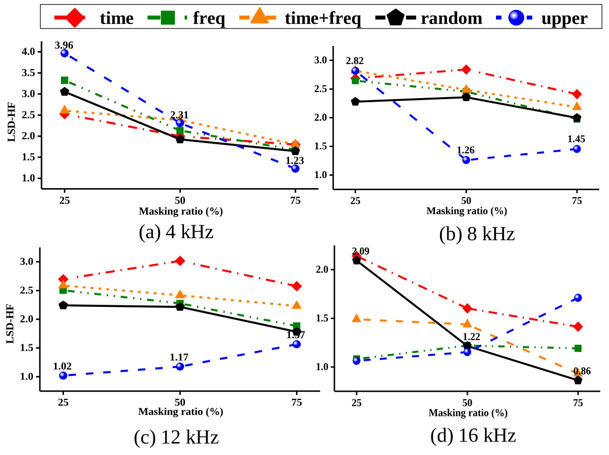


Fig. 3. LSD results by masking strategies in the pre-training.



(a) 4 kHz

(b) 8 kHz

(c) 12 kHz

(d) 16 kHz

Fig. 4. LSD-HF results by masking strategies in the downstream task.

we measure the LSD between the reconstructed and ground truth Mel-spectrograms. Fig. 3 shows that pre-training the MAE with a high ratio (80%) of random masking results in a lower reconstruction error than other masking strategies, as in the classification task [10]. Hence, we use the pre-trained MAE with a high ratio (80%) of random masking for the downstream task.

2) Fine-Tuning: For audio super-resolution, we explore various masking strategies, as shown in Fig. 4. The results show that upper-band masking is an appropriate method for reconstructing the missing part of a Mel-spectrogram. We observe that the model fine-tuned with upper-band masking can only fill in the Mel-spectrogram above a certain band. We also observe that model fine-tune with a high ratio of upper-band masking show a performance decline for audio with a 16 kHz sampling rate. The reason is that the encoder in Fre-Painter, which is a ViT, receives Mel-spectrograms split into patches as input. During training, the network corresponding to the masked areas is continuously trained to encode only the masked values. However, in the inference process, the 16 kHz Mel-spectrogram has real values in the areas that were previously masked. Consequently, the encoder receives new values as input that it has not seen during training. This leads to improper encoding and a decline in the model's performance. Therefore, we propose a mix-ratio masking strategy, which can more accurately reconstruct audio across any sampling rates audio. We compare our mix-ratio masking strategy with two different methods. The first method, mix-ratio v1, randomly selects the pre-defined ratios of 25%, 50%, and 75%. The second method, mix-ratio v2, randomly selects ratios between 0% and 75%. From results shown in Fig. 5, we can observe that mix-ratio v2 outperforms the other methods on average. Therefore, we ultimately adopt mix-ratio v2.

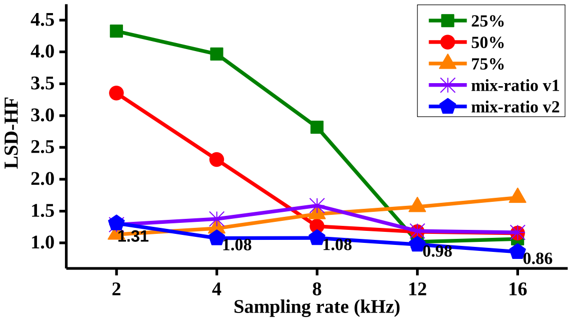


Fig. 5. LSD-HF results by masking ratio in the downstream task.

B. Evaluation of Audio Super-Resolution

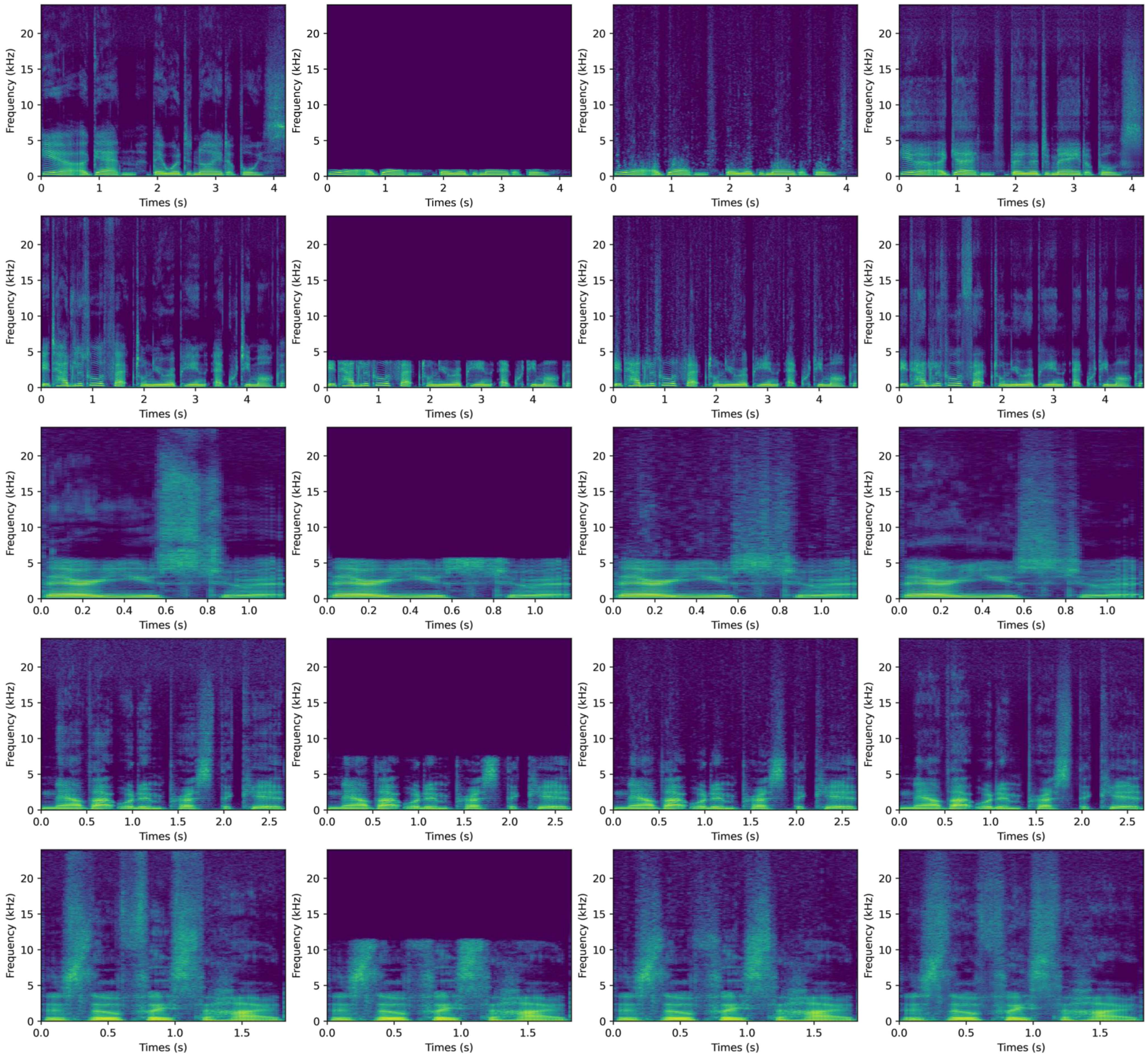
Our evaluation was conducted in two stages, depending on the target sampling rate. In the initial experiment, with the target sampling rate designated at 24 kHz, we validated the capability of Fre-Painter to correctly execute audio super-resolution. As a baseline, we used HiFi-GAN [54], which is a generator of Fre-Painter, and NU-Wave 2 [7]. NU-Wave 2 [7] was trained by modifying its original target sampling rate from 48 kHz to 24 kHz, based on the official implementation. Note that NU-Wave [7] can upsample from 6 kHz. For a fair comparison, MOS is conducted only for input sampling rate above 8 kHz

In the second experiment, we extended Fre-Painter to enable upsampling up to 48 kHz and conducted a performance comparison with other models. For this phase, we used the official implementations and checkpoints provided. VoiceFixer [6], NVSR [26], and AudioSR [38], which are trained with a target sampling rate of 44.1 kHz, are evaluated with 44.1 kHz audio as target. This information is provided for reference purposes. MOS is conducted on audio suer-resolution results from 12 kHz to 48 kHz.

Experiments are conducted at various sampling rates. Tables I and II present that Fre-Painter can successfully perform audio super-resolution from input audio of various sampling rates. Furthermore, Tabels III and IV exhibit that Fre-Painter outperforms the other models in terms of the LSD, LSD-HF, ViSQOL, and MOS. Because Fre-Painter uses a Mel-spectrogram as an input for the downstream task, it shows a slightly higher error rate in terms of the LSD-LF compared to NU-Wave 2 [7] and UDM+ [37], which use a waveform as an input. Hence, there is room for improvement in the performance by utilizing a low-resolution waveform additionally. The results of audio super-resolution can be seen in Fig. 6. For comparison, we selected NU-Wave 2 [7], which has the best LSD-HF metric among models with a target sampling rate of 48 kHz.

C. Comparison of Inference Speeds

Diffusion-based generative models require a iterative process to produce high-quality audio waveform, which inherently results in slower inference speeds. In contrast, as shown in



(a) Ground truth (b) Input audio (c) NU-Wave 2 (d) Fre-Painter

Fig. 6. Spectrograms of the ground truth, low-resolution input, and upsampled audio that has been upsampled to 48 kHz from its respective sampling rates. The sampling rates of the input audio are arranged sequentially as 2, 8, 12, 16, and 24 kHz, per row.

Table IV, Fre-Painter has an inference speed that is 28 to 243 times faster than diffusion-based models.

D. Ablation Study

We conducted an ablation study to demonstrate the effectiveness of the proposed method, and the results are shown in the Table V. We can observe that if the model is trained from scratch, it generally shows slightly lower quality compared to the model that is fine-tuned with a pre-trained encoder. Without initializing the mask token, the model cannot accurately recognize which position has been masked or not. Consequently, audio super-resolution is inadequately executed in the high-frequency region. In cases where training is conducted without masking, although the low-frequency information can be reconstructed, upsampling is not performed correctly.

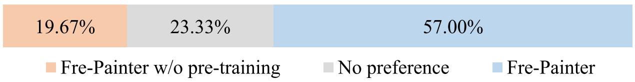


Fig. 7. Result of ABX preference test for the ablation study of pre-training.

To demonstrate the effectiveness of pre-training, we conducted an ABX preference test to directly compare models with and without pre-training. Fig. 7 exhibits that audio generated by Fre-Painter with pre-training is more preferred. In our model, the encoder uses a Mel-spectrogram converted from 24 kHz audio as input. Therefore, we can pre-train the encoder using a large

TABLE I
OBJECTIVE EVALUATION RESULTS AND COMPARISONS FOR AUDIO SUPER-RESOLUTION TO 24 KHZ

Model	LSD (↓)	LSD-HF (↓)	LSD-LF (↓)	ViSQOL (↑)
		2 kHz → 24 kHz		
Input	4.81	5.01	0.62	2.04
HiFi-GAN	3.18	3.30	0.93	1.99
NU-Wave 2	1.91	1.98	**0.68**	2.47
Fre-Painter	**1.27**	**1.30**	0.77	**3.32**
		4 kHz → 24 kHz		
Input	4.47	4.88	0.85	2.07
HiFi-GAN	3.05	3.32	0.80	2.05
NU-Wave 2	1.70	1.84	**0.69**	2.64
Fre-Painter	**1.14**	**1.20**	0.70	**3.53**
		8 kHz → 24 kHz		
Input	3.77	4.56	0.92	2.09
HiFi-GAN	2.59	3.10	0.86	2.07
NU-Wave 2	1.39	1.62	**0.68**	2.92
Fre-Painter	**1.06**	**1.18**	0.75	**3.64**
		12 kHz → 24 kHz		
Input	3.15	4.37	0.79	2.12
HiFi-GAN	1.99	2.67	0.85	2.11
NU-Wave 2	1.18	1.55	**0.56**	3.12
Fre-Painter	**0.97**	**1.14**	0.76	**3.72**
		16 kHz → 24 kHz		
Input	2.56	4.27	0.84	2.37
HiFi-GAN	1.53	2.34	0.86	2.35
NU-Wave 2	1.04	1.58	**0.59**	3.20
Fre-Painter	**0.91**	**1.10**	0.78	**3.92**

TABLE II
COMPARISON OF THE MOS WITH 95% CONFIDENCE INTERVALS FOR AUDIO SUPER-RESOLUTION TO 24 KHZ

Metric	GT	Input	HiFi-GAN	NU-Wave 2	**Fre-Painter**
MOS (↑)	3.84 ± 0.07	3.44 ± 0.07	3.51 ± 0.07	3.65 ± 0.08	**3.76 ± 0.08**

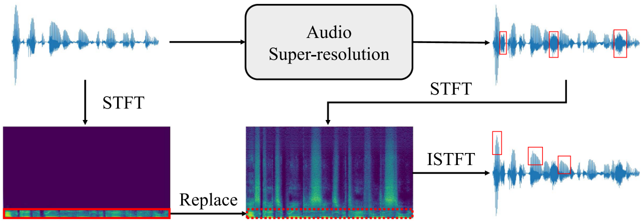


Fig. 8. Overview of post-processing through the replacement of lower frequencies.

TABLE III
OBJECTIVE EVALUATION RESULTS AND COMPARISONS FOR AUDIO SUPER-RESOLUTION TO 48 KHZ

Model	LSD (↓)	LSD-HF (↓)	LSD-LF (↓)	ViSQOL (↑)
		2 kHz → 48 kHz		
Input	4.92	5.02	0.37	1.57
VoiceFixer	1.31	1.33	0.82	2.26
NU-Wave 2	1.56	1.59	0.40	1.72
NVSR*	1.30	1.33	**0.32**	2.45
Fre-Painter	**1.09**	**1.10**	0.71	**2.84**
		4 kHz → 48 kHz		
Input	4.78	4.98	0.61	1.58
VoiceFixer	1.25	1.28	0.88	2.44
NU-Wave 2	1.48	1.54	0.42	1.81
NVSR*	1.22	1.27	**0.36**	2.61
AudioSR*	1.52	1.59	0.45	2.60
Fre-Painter	**1.01**	**1.03**	0.69	**3.03**
		8 kHz → 48 kHz		
Input	4.47	4.88	0.74	1.56
VoiceFixer	1.21	1.26	0.92	2.56
NU-Wave 2	1.26	1.36	0.46	2.07
NVSR*	1.12	1.22	0.43	2.79
UDM+	1.34	1.46	**0.11**	2.29
AudioSR*	1.42	1.55	0.45	2.79
Fre-Painter	**0.98**	**1.02**	0.72	**3.10**
		12 kHz → 48 kHz		
Input	4.20	4.83	0.75	1.58
VoiceFixer*	1.17	1.24	0.95	2.61
NU-Wave 2	1.11	1.26	0.44	2.34
NVSR*	1.06	1.20	0.46	2.85
UDM+	1.16	1.33	**0.14**	2.59
AERO	**0.93**	**0.95**	0.82	3.07
AudioSR*	1.30	1.48	0.56	2.87
Fre-Painter	0.95	0.99	0.78	**3.16**
		16 kHz → 48 kHz		
Input	3.90	4.74	0.81	1.80
VoiceFixer*	1.15	1.23	0.96	2.73
NU-Wave 2	1.03	1.21	0.49	2.62
NVSR*	1.00	1.20	0.46	3.01
UDM+	1.03	1.26	**0.13**	2.88
AudioSR*	1.22	1.46	0.56	2.98
Fre-Painter	**0.90**	**0.95**	0.79	**3.36**
		24 kHz → 48 kHz		
Input	2.77	3.85	0.69	2.97
WSRGlow	0.91	1.20	0.43	3.43
VoiceFixer*	1.10	1.25	0.95	3.07
NU-Wave 2	0.84	1.08	0.45	3.45
NVSR*	0.93	1.28	0.44	3.50
UDM+	**0.83**	1.16	**0.14**	3.65
AudioSR*	1.07	1.45	0.54	3.45
Fre-Painter	0.86	**0.89**	0.79	**3.77**

Models marked with * indicate that their target sampling rate is 44.1 khz.

amount of 24 kHz audio data, which is easier to obtain than 48 kHz audio data. This approach enables more robust speech representation learning.

E. Post-Processing Through Replacement of Lower Frequencies

To improve the weak performance of Fre-Painter in lower frequencies, we adopt post-processing through lower frequencies replacement [26], [38]. This method involves replacing the lower frequencies of the generated audio with the lower frequencies of the input audio using STFT and inverse STFT (ISTFT), as the input audio contains accurate low-frequency information. Fig. 8 illustrates the overview of the post-processing. Table VI exhibits significant improvements in terms of LSD-LF and ViSQOL metrics when post-processing is applied to Fre-Painter's output.

F. Text-to-Speech Synthesis With Audio Super-Resolution

For practical applicability, we extend Fre-Painter to a two-stage TTS system. In conventional two-stage TTS system, an acoustic model generates a Mel-spectrogram as an intermediate representation [68], [69], and then a neural vocoder synthesizes an audio waveform from the Mel-spectrogram. Additionally, if audio super-resolution is performed using models that take an audio waveform as input, a total of three stages are involved. On the other hand, Fre-Painter can replace traditional neural vocoders while also performing audio super-resolution.

We trained Glow-TTS [69] as an acoustic model with audio of 16 kHz, which is a common sampling rate in speech

TABLE IV
COMPARISON OF INFERENCE SPEEDS AND MOS WITH 95% CONFIDENCE INTERVALS FOR AUDIO SUPER-RESOLUTION FROM 12 KHZ TO 48 KHZ

Method	MOS (↑)	Speed (kHz) (↑)	Real-time (↑)	# Param. (M) (↓)
GT (48 kHz)	3.81 ± 0.05	-	-	-
GT (12 kHz)	3.72 ± 0.04	-	-	-
VoiceFixer	3.78 ± 0.05	1151.91	×26.12	122.07
NVSR	**3.80 ± 0.05**	1348.16	×30.57	116.85
NU-Wave 2	3.77 ± 0.05	175.53	×3.66	**1.71**
UDM+	3.79 ± 0.05	20.40	×0.43	2.31
AERO	3.76 ± 0.05	4396.72	×91.60	19.43
AudioSR+	3.77 ± 0.05	32.10	×0.67	258.20
Fre-Painter	**3.80 ± 0.05**	**5006.23**	×**104.30**	22.95

TABLE V
ABLATION STUDY RESULTS FOR OBJECTIVE EVALUATION AND COMPARISON OF MOS WITH 95% CONFIDENCE INTERVALS

Method	MOS (↑)	LSD (↓)	LSD-HF (↓)	LSD-LF (↓)
Fre-Painter	**3.87 ± 0.09**	**1.08**	**1.19**	**0.78**
w/o pre-training	3.83 ± 0.08	1.14	1.24	0.85
w/o mask token init.	3.72 ± 0.08	2.12	2.47	0.92
w/o masking	3.61 ± 0.08	2.51	2.99	0.90

TABLE VI
OBJECTIVE EVALUATION RESULTS AND COMPARISONS FOR POST-PROCESSING THROUGH REPLACEMENT

Model	LSD (↓)	LSD-HF (↓)	LSD-LF (↓)	ViSQOL (↑)
Fre-Painter	0.86	**0.89**	0.79	3.77
Fre-Painter + Post-processing	**0.69**	**0.89**	**0.36**	**3.97**

Audio super-resolution is performed from 24 khz to 48 khz.

TABLE VII
COMPARISON OF MOS WITH 95% CONFIDENCE INTERVALS, VISQOL, AND PRONUNCIATION ACCURACY IN TEXT-TO-SPEECH WITH AUDIO SUPER-RESOLUTION

Method	MOS (↑)	ViSQOL (↑)	PER (↓)	WER (↓)
GT (48kHz)	3.75 ± 0.02	-	9.45	1.15
GT (16kHz)	3.73 ± 0.02	1.82	9.66	1.15
Glow-TTS + HiFi-GAN	3.67 ± 0.02	1.61	15.76	3.90
Glow-TTS + HiFi-GAN + NU-Wave 2	3.68 ± 0.02	2.02	15.49	3.61
Glow-TTS + HiFi-GAN + UDM+	3.68 ± 0.02	1.99	15.39	3.24
Glow-TTS + Fre-Painter	**3.71 ± 0.02**	**2.34**	**14.99**	**2.57**

recognition datasets. In the inference procedure, we used the trained Glow-TTS model to input text and generate a Mel-spectrogram. Instead of using a neural vocoder, we utilized Fre-Painter to synthesize a 48 kHz audio waveform directly from the generated Mel-spectrogram. Notably, NU-Wave 2 [7] and UDM+ [37] cannot upsample the Mel-spectrogram; therefore, we utilize these in conjunction with HiFi-GAN. However this three-stage process is ineffective. Table VII shows that Fre-Painter performs better in TTS pipelines. The combination of TTS system and audio super-resolution not only enhances audio quality but also yields a slight improvement in pronunciation accuracy. NU-Wave 2 [7] and UDM+ [37] achieve some enhancement effects through diffusion-based generation. Fre-painter, using the MAE-based generator, demonstrates superior performance compared to HiFi-GAN and the three-stage pipeline. The experimental results exhibits that it is feasible to efficiently train the acoustic model with low-resolution data and synthesize high-quality speech by utilizing an audio super-resolution model as a neural vocoder.

VI. DISCUSSION

A. Broader Impact

By utilizing the Mel-spectrogram as input, Fre-painter can be effectively applied within a two-stage speech synthesis system. Even when high-resolution samples are scarce for low-resource languages or speakers, an acoustic model can be trained using low-resolution speech data. The incorporation of Fre-painter as a neural vocoder can facilitate the synthesis of high-resolution audio waveforms.

In real-world scenarios, there exist audio files that have been upsampled and stored without the restoration of high-frequency information. For instance, in datasets like CommonVoice, all audio files are set to a 48 kHz sampling rate. However, many of these recordings lack high-frequency information. Once an audio file is upsampled and stored, it becomes challenging to discern its original sampling rate. Our model is capable of enhancing audio quality by filling in the missing high-frequency information, irrespective of the audio's original sampling rate.

B. Limitations

Fre-Painter has demonstrated remarkable reconstruction capabilities specifically in high-frequency information, leading to an overall superior performance. However, compared to models that directly use audio waveform as an input, Fre-Painter exhibits a slightly unsatisfactory reconstruction of low-frequency information. Despite the low resolution of the audio, it contains accurate information in the low-frequency range. Therefore, utilizing input audio waveform can efficiently reconstruct information in the low-frequency range. Post-processing through replacement has significantly improves the low-frequencies, but there is still room for improvement.

We have observed that Fre-Painter encounters failure cases when performing audio super-resolution on some real-world audio samples that are mixed with noise. When processing audio with noise, the model also enhances the noise parts. These unintended results indicate that the model performs audio super-resolution on the overall audio without identifying or separating the noise from the audio.

C. Future Work

Our future research will aim to explore strategies that use audio waveforms as supplementary inputs to improve the reconstruction of low-frequency information. Furthermore, we plan to expand the functionality of Fre-Painter to include audio restoration tasks such as denoising. In addition, we will replace the HiFi-GAN generator with BigVGAN [59] and scale up for better generalization on various out-of-distribution scenarios.

VII. CONCLUSION

We propose Fre-Painter, a robust neural audio super-resolution system with MAE and neural vocoder. We successfully integrate MAE into audio synthesis tasks by synthesizing high-resolution audio from low-resolution audio. The experimental results demonstrate the effectiveness of Fre-Painter for both subjective and objective metrics. We also investigate the various masking strategies for pre-training and audio super-resolution. We believe that it will be beneficial to future speech research. Furthermore, we extend Fre-Painter to serve as a vocoder in a two-stage TTS system that can synthesize high-resolution speech from text using an acoustic model trained on a low-resolution speech dataset.

REFERENCES

[1] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 873–881, Mar. 2007.

[2] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.

[3] M. Zöhrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2398–2409, Dec. 2015.

[4] X. Liu and C. Bao, "Audio bandwidth extension based on ensemble echo state networks with temporal evolution," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 594–607, Mar. 2016.

[5] J. Lee and S. Han, "NU-Wave: A diffusion probabilistic model for neural audio upsampling," in *Proc. Interspeech*, 2021, pp. 1634–1638.

[6] H. Liu et al., "VoiceFixer: A. unified framework for high-fidelity speech restoration," in *Proc. Interspeech*, 2022, pp. 4232–4236.

[7] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," in *Proc. Interspeech*, 2022, pp. 4401–4405.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[10] P.-Y. Huang et al., "Masked autoencoders that listen," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. Curran Associates, Inc., 2022, vol. 35, pp. 28708–28720.

[11] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked autoencoding audio spectrogram transformer," in *Proc. Interspeech*, 2022, pp. 2438–2442.

[12] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Proc. Interspeech*, 2015, pp. 2578–2582.

[13] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 883–894, May 2018.

[14] H. Wang and D. Wang, "Towards robust speech super-resolution," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2058–2066, 2021.

[15] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[16] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 646–650.

[17] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. W. Koh, and S. Ermon, "Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations.," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[18] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, "Real-time speech frequency bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 691–695.

[19] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEANet: A multi-modal speech enhancement network," in *Proc. Interspeech*, 2020, pp. 1126–1130.

[20] V.-A. Nguyen, A. H. T. Nguyen, and A. W. H. Khong, "Tunet: A block-online bandwidth extension model based on transformers and self-supervised pretraining," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 161–165.

[21] K. Zhang, Y. Ren, C. Xu, and Z. Zhao, "WSRGlow: A. glow-based waveform generative model for audio super-resolution," in *Proc. Interspeech*, 2021, pp. 1649–1653.

[22] A. van den Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Workshop Speech Synth. Workshop*, 2016, p. 125.

[23] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, "Bandwidth extension is all you need," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 696–700.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[25] Q. Tian et al., "TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis," 2020, *arXiv:2011.12206*.

[26] H. Liu, W. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech super-resolution," in *Proc. Interspeech*, 2022 pp. 4227–4231.

[27] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5029–5033.

[28] S. E. Eskimez and K. Koishida, "Speech super resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 3717–3721.

[29] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," 2019, *arXiv:1903.09027*.

[30] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, "Nu-GAN: High resolution neural upsampling with GAN," 2020, *arXiv:2010.11362*.

[31] S. Hu, B. Zhang, B. Liang, E. Zhao, and S. Lui, "Phase-aware music super-resolution using generative adversarial networks," in *Proc. Interspeech*, 2020 pp. 4074–4078.

[32] M. Mandel, O. Tal, and Y. Adi, "AERO: Audio super resolution in the spectral domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[33] E. Moliner and V. Välimäki, "BEHM-GAN: Bandwidth extension of historical music using generative adversarial networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 943–956, 2023.

[34] C. Shuai, C. Shi, L. Gan, and H. Liu, "mdctGAN: Taming transformer-based GAN for speech super-resolution with modified DCT spectra," in *Proc. Interspeech*, 2023, pp. 5112–5116.

[35] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2021.

[36] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. Int. Conf. Learn. Representations*, 2021.

[37] C.-Y. Yu, S.-L. Yeh, G. Fazekas, and H. Tang, "Conditioning and sampling in variational diffusion models for speech super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[38] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "AudioSR: Versatile audio super-resolution at scale," 2023, *arXiv:2309.07314*.

[39] Z. Lian, J. Tao, B. Liu, and J. Huang, "Unsupervised representation learning with future observation prediction for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 3840–3844.

[40] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," in *Proc. Interspeech*, 2019, pp. 1153–1157.

[41] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[42] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[43] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, and L.-R. Dai, "A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1927–1939, 2023.

[44] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, "HierSpeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 16624–16636.

[45] D. Cai, W. Wang, and M. Li, "Incorporating visual information in audio based self-supervised speaker recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1422–1435, 2022.

[46] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-Y. Lee, "Improving the adversarial robustness for speaker verification by self-supervised learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 202–217, 2022.

[47] A. Lee et al., "Direct speech-to-speech translation with discrete units," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 3327–3339.

[48] S. Chen et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[49] A. V. D. Oord et al., “WaveNet: A generative model for raw audio,” 2016, *arXiv:1609.03499*.

[50] N. Kalchbrenner et al., “Efficient neural audio synthesis,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2410–2419.

[51] A. Oord et al., “Parallel wavenet: Fast high-fidelity speech synthesis,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3918–3926.

[52] K. Kumar et al., “MelGAN: Generative adversarial networks for conditional waveform synthesis,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[53] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel waveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *Proc. IEEE Speech Signal Process. Int. Conf. Acoust.*, 2020, pp. 6199–6203.

[54] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–170 33.

[55] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, “UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation,” in *Proc. Interspeech*, 2021, pp. 2207–2211.

[56] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Trans. Mach. Learn. Res.*, 2023.

[57] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, “Fre-GAN: Adversarial frequency-consistent audio synthesis,” in *Proc. Interspeech*, 2021, pp. 2197–2201.

[58] S.-H. Lee, J.-H. Kim, K.-E. Lee, and S.-W. Lee, “FRE-GAN 2: Fast and efficient frequency-consistent audio synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6192–6196.

[59] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “BigVGAN: A universal neural vocoder with large-scale training,” in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

[60] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019 pp. 9992–10002.

[61] H. Zen et al., “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Proc. Interspeech*, 2019, pp. 1526–1530.

[62] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” The Centre for Speech Technology Research (CSTR), Univ. Edinburgh, 2019, doi: 10.7488/ds/2645.

[63] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[64] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, “End-to-end adversarial text-to-speech,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[65] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O’Gorman, and A. Hines, “VISQOL v3: An open source production ready objective speech and audio metric,” in *Proc. IEEE 12th Int. Conf. Qual. Multimedia Experience*., 2020, pp. 1–6.

[66] Z. Borsos et al., “AudioLM: A language modeling approach to audio generation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[67] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, “Whisper: Tracing the spatiotemporal process of information diffusion in real time,” *IEEE Trans. Visual. Comput. Graph.*, vol. 18, no. 12, pp. 2649–2658, 2012.

[68] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, “Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 13198–13206.

[69] J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-TTS: A. generative flow for text-to-speech via monotonic alignment search,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 8067–8077.

Seung-Bin Kim received the B.S. degree in physics from the University of Seoul, Seoul, South Korea, in 2021. He is currently working toward the master’s and Ph.D. degrees with the Department of Artificial Intelligence, Korea University, Seoul. His research interests include artificial intelligence and audio signal processing.

Sang-Hoon Lee received the B.S. degree in life science from Dongguk University, Seoul, South Korea, in 2016, and the Ph.D. degree in brain and cognitive engineering from Korea University, Seoul, South Korea, in 2023. He is currently a Postdoctoral Researcher with AI Research Center, Korea University, Seoul. His research interests include artificial intelligence and audio signal processing.

Ha-Yeong Choi received the B.S. degree in computer science and engineering from Handong Global University, Pohang, South Korea, in 2022. She is currently working toward the master’s degree in artificial intelligence with Korea University, Seoul, South Korea. Her research interests include deep learning-based speech synthesis and generative modeling.

Seong-Whan Lee (Fellow, IEEE) received the B.S. degree in computer science and statistics from Seoul National University, Seoul, South Korea, in 1984, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1986 and 1989, respectively. He is currently the Head of the Department of Artificial Intelligence, Korea University, Seoul. His research interests include artificial intelligence, pattern recognition, and brain engineering. He is a Fellow of the International Association of Pattern Recognition and Korea Academy of Science and Technology.