

Anti-Aliasing Framework for Time-Domain Speech Enhancement

Joonhyeon Bae¹, Jaepil Ko^{2*}

Abstract

Since the introduction of deep learning, research on speech enhancement has been developing rapidly. Until a few years ago, time-domain models performed poorly compared to time-frequency models. However, recent time-domain models such as DEMUCS, MANNER, and SE-Conformer have achieved competitive performance by adopting U-Net encoder-decoder architectures with skip connections. Despite this progress, these U-Net-based models have an overlooked weakness: aliasing artifacts from down-sampling and up-sampling operations. Prior work has shown that max pooling and strided convolutions introduce aliasing, which appears as network-generated high-frequency noise in the output. This contradicts the fundamental goal of speech enhancement noise reduction. Therefore, applying anti-aliasing methods can be useful for further improving time-domain models. We systematically apply anti-aliasing methods (BlurPooling, sub-pixel convolution with ICNR initialization) to MANNER, a time-domain speech enhancement model, on the VoiceBank+DEMAND dataset. Our results show improvements in perceptual quality metrics. Our contributions are threefold. First, we investigate aliasing artifacts caused by down-sampling and up-sampling operations in U-Net-based time-domain speech enhancement models. Second, we formulate a unified 1D anti-aliasing framework that integrates existing anti-aliasing operations, including BlurPooling, sub-pixel convolution, and ICNR initialization, into waveform-domain speech enhancement architectures. Third, we provide empirical validation on the framework through objective metrics, ablation studies, an additional Demucs experiment, and spectrogram/error-spectrogram visualizations.

Key Words: Anti-Aliasing, Deep Learning, Speech Enhancement, Time-Domain Models.

I. INTRODUCTION

Speech enhancement models can be categorized by the domain in which they process signals: time-domain models operate directly on waveforms, while time-frequency domain models process spectral representations obtained via short-time Fourier transform (STFT). Until recently, time-domain models showed inferior performance compared to time-frequency domain models. However, recent time-domain models such as DEMUCS [1], SE-Conformer [2], and MANNER [3] have outperformed previous works. All three models achieved strong results by utilizing the U-Net [4] structure, which consists of down and up-sampling of data. However, the encoder-decoder structures used in these models suffer from a vulnerability to aliasing during the pooling and up-sampling processes. Aliasing in time-domain speech enhancement appears as network-generated noise in high frequencies.

This contradicts one of the fundamental purposes of speech enhancement: noise reduction. Therefore, applying anti-aliasing methods to time-domain speech enhancement

models is important for improving model performance. MANNER has additional Conformer [5] and attention module between the up and down convolution. We adopted MANNER as our baseline to show those additional modules barely remove aliasing, demonstrating the necessity of explicit anti-aliasing modules. We applied anti-aliasing methods to MANNER. First, BlurPooling proposed by Zhang et al. [6] was applied to remove aliasing in the pooling process. Then, in the up-sampling process, sub-pixel convolution proposed by Shi et al. [7] with ICNR initialization [8] was applied instead of transposed convolution. It should be emphasized that BlurPooling, sub-pixel convolution, and ICNR initialization are existing techniques. Therefore, this paper does not claim novelty in the individual operations themselves. Instead, the novelty of this work lies in formulating these operations as a unified 1D anti-aliasing framework for time-domain speech enhancement, where both down-sampling and up-sampling stages of waveform-domain encoder-decoder models are explicitly modified to reduce aliasing-related artifacts. While prior work [6-7] demonstrated anti-aliasing benefits in computer vision, its application to time-domain speech

Manuscript received April 26, 2026; Revised June 16, 2026; Accepted July 23, 2026. (ID No. JMIS- 26M-04-020)

Corresponding Author (*): Jaepil Ko, +82-54-478-7529, nonezero@kumoh.ac.kr

¹Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea, outersky@snu.ac.kr

²Department of Computer Engineering, Kumoh National Institute of Technology, Gumi, Korea, nonezero@kumoh.ac.kr

enhancement has not been explored.

Our experiments on VoiceBank+DEMAND show that the proposed framework improves perceptual quality on MANNER. To further examine whether the framework is specific to MANNER, we additionally apply the same anti-aliasing modifications to Demucs, another waveform-domain encoder-decoder speech enhancement model. The additional Demucs experiment improves PESQ, CSIG, CBAK, and COVL, supporting that the proposed framework is not limited to a single architecture. Our results demonstrate that anti-aliasing eliminates intrinsic network-generated noise, achieving competitive performance among time-domain models.

II. RELATED WORK

2.1. Time-Domain Speech Enhancement Models

SEGAN [9] adopts a GAN structure. SEGAN's Generator network directly infers clean speech from noisy waveforms. The Generator adopts U-Net [4] like skip connections. After inference, (clean, noisy) and (output, noisy) pairs are fed to SEGAN's Discriminator. The Discriminator classifies whether an input pair contains clean utterance or Generator output. Conv-TasNet [10] introduced a paradigm shift in time-domain audio processing. Unlike traditional methods using fixed STFT bases, Conv-TasNet learns adaptive basis functions through 1D convolutions, converting waveforms into a learnable latent representation. The model consists of three stages: encoder (time-domain to latent), separator (mask estimation in latent space), and decoder (latent to time-domain). While originally designed for speech separation, this learnable-basis paradigm influenced subsequent time-domain enhancement models [3,8,11], which adopted similar encoder-decoder structures. DEMUCS [1] is also a U-Net structure model and uses LSTM [12] at the bottleneck. Each block in the model's encoder and decoder consists of 1D convolution with GLU activation. TSTNN [13] utilizes Transformer [14] blocks for speech enhancement. TSTNN [13] uses transformers with different dimensional representations to capture both global and local relationships simultaneously. SE-Conformer [2] adopts a convolution-transformer mixed structure called Conformer block [5]. SE-Conformer also uses DEMUCS-like blocks for encoder and decoder but replaces LSTM [12] with Conformer. MANNER [3] adopts a U-Net structure like [1,9]. Unlike [1,9] which have weaknesses in capturing long distance relations, MANNER exploits attention methods in the middle of each encoder and decoder block to capture such relations well.

2.2. Anti-Aliasing in Deep Learning

Traditionally, anti-aliasing has been the domain of computer graphics (CG). Popular methods like multi-sample anti-aliasing (MSAA) [16] and sub-pixel morphological anti-aliasing (SMAA) [17] are currently being actively applied in real-time 3D rendering. However, it has also been actively studied in deep learning. Various anti-aliasing methods have been proposed for deep learning, including BlurPooling [6], sub-pixel convolution [7]. Among these, BlurPooling addresses down-sampling aliasing, while sub-pixel convolution tackles up-sampling artifacts. We focus on these two methods as they directly target the pooling and up-sampling operations in U-Net architectures. Below, we describe each method in detail.

2.2.1. BlurPooling

In classical signal processing, applying a low-pass filter before down-sampling has long been standard practice to prevent aliasing (Nyquist-Shannon theorem). However, deep learning practitioners often overlooked this principle, using max pooling and strided convolutions without proper filtering. Zhang et al. [6] revealed that this causes severe aliasing in CNNs and proposed BlurPooling, which reintroduces the traditional "blur-before subsample" approach into neural networks. Specifically, Zhang applies a Gaussian low-pass filter before subsampling. The BlurPooling operation can be formulated as:

$$\text{BlurPool}(x) = \text{Subsample}(G_\sigma * x), \quad (1)$$

where G_σ is a Gaussian kernel and $*$ denotes convolution. For 1D signals (time-domain audio), the Gaussian kernel is:

$$G_\sigma(i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{i^2}{2\sigma^2}\right). \quad (2)$$

In practice, a binomial filter approximates the Gaussian:

$$G = \frac{1}{4} [1, 2, 1]. \quad (3)$$

This lowpass filtering removes frequencies above the new Nyquist frequency before subsampling, preventing aliasing according to the sampling theorem:

$$f_s \geq 2f_{\max}, \quad (4)$$

where f_s is the sampling frequency and $f_{\max}$ is the maximum frequency in the signal. Zhang et al. [6] also stated that this blurring process makes the model shift-invariant. Zhang et al. also stated that such blurring makes the model robust when dealing with rotation, scaling,

blurring, and noise in input data. Subsequent work by Park et al. [18] showed that blur inside the model has the same effect as an ensemble, and that blurring of features flattens the loss landscape in the network. Both studies [6,18] reveal that anti-aliasing not only reduces aliasing of features but also helps generalization performance.

2.2.2. Sub-Pixel Convolution

Studies like Aitken et al. [8] focused on reducing aliasing in up-sampling of neural networks. Shi et al. [7] proposed sub-pixel convolution. Unlike transposed convolution, sub-pixel convolution increases the number of channels in the output feature, then shuffles the channels to get the desired output dimension. Sub-pixel convolution performs up-sampling in two steps:

Step 1: Increase channel dimension through convolution:

$$f = \text{Conv1D}(x), f \in \mathbb{R}^{B \times T \times (r \cdot C)}. \quad (5)$$

Step 2: Rearrange (shuffle) channels into temporal dimension:

$$y = \text{PixelShuffle}(f), y \in \mathbb{R}^{B \times rT \times C}, \quad (6)$$

where B is a batch size, T is time steps, C is channels, and r is up-sampling factor (typically 2). The PixelShuffle operation is defined as:

$$\text{PixelShuffle}(f)[b, rt + i, c] = f[b, t, rc + i], \quad (7)$$

for $i \in [0, r - 1], t \in [0, T - 1], c \in [0, C - 1]$.

Sub-pixel convolution performs all operations in low resolution space before shuffling. However, Odena et al. [19] and Aitken et al. [8] identified three sources of checkerboard artifacts in up-sampling layers: (1) deconvolution overlap when kernel size is not divisible by stride, (2) random initialization causing each sub-kernel to be initialized independently, and (3) inhomogeneous gradient updates from loss functions with down-sampling operations. While sub-pixel convolution avoids deconvolution overlap by design, it still suffers from checkerboard artifacts due to random initialization. This occurs because the r^2 sub-kernels that generate neighboring high-resolution features are initialized independently but applied to the same low-resolution input, creating discontinuities in the output. ICNR Initialization: To address the random initialization problem, Aitken et al. [8] proposed initialize to convolution NN resize (ICNR), an initialization scheme that makes sub-pixel convolution equivalent to nearest neighbor resize followed by convolution at initialization time. The key insight is to initialize all r^2 sub-kernels identically, so that at initialization the network behaves like convolution followed by nearest-neighbor up-sampling.

Specifically, we first initialize a single kernel W_0 using standard orthogonal initialization, then copy its weights to all r^2 sub-kernels: $W'_n = W_0$ for all $n \in \{0, \dots, r^2 - 1\}$. This ensures that neighboring high-resolution features depend on identical kernels at initialization, eliminating checkerboard patterns while preserving the model's capacity to learn diverse up-sampling kernels during training. In our implementation, we employ ICNR initialization for all sub-pixel convolution layers, ensuring artifact-free up-sampling from the start of training.

III. UNDERSTANDING ALIASING IN TIME-DOMAIN SPEECH ENHANCEMENT

Before describing our methods, we provide a systematic analysis of how aliasing occurs in time-domain speech enhancement models and why it degrades performance.

3.1. Aliasing in Down-Sampling Operations

In U-Net encoder blocks, temporal resolution is reduced through strided convolution or pooling. Consider a 1D time-domain signal $x[n]$ sampled at rate f_s . When we downsample by factor s without proper filtering:

$$y[m] = x[sm]. \quad (8)$$

According to the Nyquist-Shannon sampling theorem, if $x[n]$ contains frequencies above $f_s/(2s)$, these high frequencies will alias into lower frequencies in $y[n]$. Strided convolution performs:

$$y[m] = \sum_{k=0}^{K-1} w[k] \cdot x[sm - k], \quad (9)$$

where $w[k]$ is the convolution kernel. Standard convolution kernels are not explicitly constrained to be low-pass filters, so strided convolution does not guarantee anti-aliased down-sampling. We can formulate max pooling as following:

$$y[m] = \max_{0 \leq k < K} x[sm + k]. \quad (10)$$

Max pooling does not include an explicit anti-aliasing low-pass filter before subsampling, and its subsampling stage can break shift equivariance, making the resulting feature maps sensitive to small input shifts [6].

3.2. Aliasing in Up-Sampling Operations

U-Net decoder blocks increase resolution through up-sampling. Transposed convolution (also called deconvolution) is commonly used:

$$y[n] = \sum_k x[k] \cdot w[n - k \cdot r]. \quad (11)$$

This operation can create checkerboard artifacts due to uneven kernel overlap [8], which results in aliasing artifacts in the frequency domain.

3.3. Why Aliasing Matters for Speech Enhancement

Aliasing in time-domain speech enhancement produces spurious high-frequency components that degrade perceptual quality, contradict the goal, and harm metrics. Human hearing is sensitive to high-frequency artifacts, which can sound like metallic or whistling noise. Speech enhancement aims to reduce noise, but aliasing introduces network-generated noise. Additionally, PESQ and other metrics penalize these artifacts. Key Insight: By applying anti-aliasing methods, we eliminate a source of network-induced degradation, allowing the model to focus on genuine speech enhancement.

IV. PROPOSED METHOD

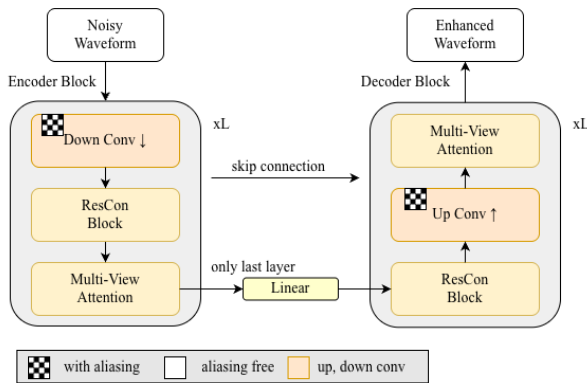
4.1. Baseline: MANNER

MANNER [3] is a time-domain speech enhancement model. The structure is shown in Fig. 1(A). The model adopts a U-Net encoder-decoder architecture with $L=4$ layers. The model begins with a 1D convolution layer that expands the input from 1 channel to $N=60$ channels, followed by batch normalization and ReLU [15] activation. Unlike previous models [1,9] that use simple LSTM or attention at the bottleneck, MANNER introduces Multi-View Attention within each encoder/decoder block, enabling efficient capture of long-distance temporal relationships.

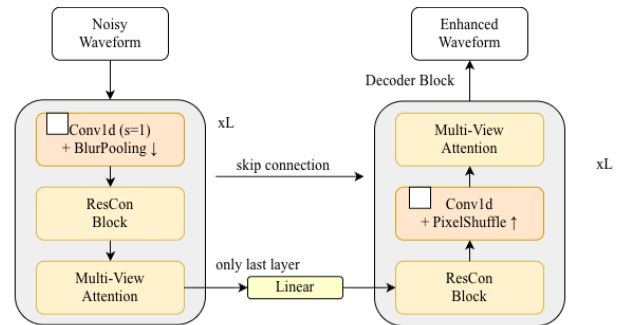
Encoder-Decoder Architecture: Each encoder block consists of three components applied sequentially: (1) Down Conv a strided convolution that reduces temporal resolution while maintaining channel dimension, followed

by batch normalization and ReLU; (2) ResCon (Residual Conformer) block inspired by the Conformer architecture [5], this block enriches channel representations through pointwise convolution (expanding channels by factor $G_0=2$), depth-wise convolution with Swish activation [20], and another pointwise convolution, with residual connections; (3) Multi-View Attention block extracts three complementary representations (channel, global, local) from the signal and fuses them to emphasize important features. The decoder mirrors this structure, replacing Down Conv with Up Conv (transposed convolution with the same kernel size and stride) to restore temporal resolution, and using $G_1=1/2$ in ResCon to reduce channels back. Skip connections transfer features from encoder to decoder via elementwise summation. The final output is obtained through a mask gate. The mask is element-wise multiplied with the output of the first convolution layer to produce denoised features, which are then reduced to 1 channel via a final convolution to produce the enhanced speech.

Multi-View Attention: The Multi-View Attention block is the key innovation of MANNER, extracting three complementary views of the input signal. The input $x \in \mathbb{R}^{N \times T_l}$ is split into three paths via 1×1 convolutions, each reducing channels from N to $N/3$. The Channel Attention path applies both average and max pooling across time, processes them through shared linear layers, and produces channel-wise weights α_C via sigmoid activation, emphasizing important channel features. The Global Attention path chunks the signal (chunk size $C=64$, 50% overlap) and applies self-attention across chunks to capture long-range dependencies. The Local Attention path applies depth wise convolution (kernel size $C/2-1$) within each chunk, then uses concatenated average and max pooling to produce local weights α_L . These three attention outputs are concatenated and fused via convolution, then processed through a residual gate (mask gate structure) to control information flow before being added back to the input via



(A) MANNER architecture



(B) MANNER architecture applied with proposed framework

Fig. 1. Architectural difference after applying proposed method.

residual connection. Vulnerability to aliasing: The baseline MANNER uses strided convolutions (down-sampling) and transposed convolutions (up-sampling), both of which are prone to aliasing.

4.2. Anti-Aliasing Framework for Speech Enhancement

We propose a systematic anti-aliasing framework applicable to U-Net-based speech enhancement models. Fig. 1(B) illustrates MANNER with proposed framework. The framework addresses aliasing at two stages: Down-sampling anti-aliasing (BlurPooling): After each strided convolution in encoder blocks, we insert a Gaussian low-pass filter (kernel size 3, $\sigma=1.0$) followed by subsampling. This implements the BlurPooling approach [6]. In practice, we replace each stride-2 1D convolution with a stride-1 1D convolution followed by a BlurPooling layer.

Up-sampling anti-aliasing (Sub-pixel Convolution): We replace all transposed convolutions in decoder blocks with sub-pixel convolutions [7]. In practice, we replace transposed convolution with one-dimensional convolution with augmented channels with following PixelShuffle. In addition to the architectural anti-aliasing modules, We also apply PCS [21] for constructing a perceptually enhanced target. PCS applies frequency-dependent gamma correction to STFT magnitudes to enhance perceptual quality:

$$S_{PCS}(t, f) = \text{sign}(S(t, f)) \cdot |S(t, f)|^{\gamma(f)}, \quad (12)$$

where $\gamma(f)$ is a frequency-dependent exponent designed based on human auditory perception. This is a domain-agnostic technique that adds no parameters and minimal computation.

4.3. Implementation Details

For BlurPooling, we use the Triangle-3 kernel $[1, 2, 1]/4$, which approximates a 1D Gaussian filter. We apply reflection padding to avoid boundary artifacts. Subpixel convolution is applied to all 4 decoder blocks. We make no architectural changes to the baseline MANNER hidden dimensions and attention mechanisms remain identical. Because the framework modifies generic down-sampling and up-sampling operations, it can be applied to other U-Net-based time-domain enhancement models. In Section 6.2, we further examine this applicability by applying the same anti-aliasing modifications to Demucs.

V. EXPERIMENTAL SETUP

5.1. Dataset

We used the commonly used VoiceBank+DEMAND

dataset [22] to validate the proposed method. VoiceBank+DEMAND dataset is obtained by mixing VoiceBank and DEMAND. The original sampling rate was 48 kHz, but we down-sampled it to 16 kHz. VoiceBank+DEMAND's training set consists of 11,572 utterances, in which 28 different voices (equal ratio of male to female) are mixed with noise at SNR levels of 15, 10, 5, and 0 dB. The test set consists of 824 utterances mixed with noise at SNR levels (17.5, 12.5, 7.5, 2.5 dB) using a single male and female voice.

5.2. Evaluation Metrics

We used PESQ, STOI, CSIG, CBAK, and COVL, metrics commonly used in many speech enhancement research studies. PESQ (Perceptual Evaluation of Speech Quality) [23] is an indicator for measuring speech quality and ranges from -0.5 to 4.5. Short-time objective intelligibility (STOI) [24] is an indicator for measuring speech intelligibility and ranges from 0 to 100. CSIG: mean opinion score (MOS) prediction of signal distortion attending only to speech signal [11], CBAK: MOS prediction of the intrusiveness of background noise [11], and COVL: MOS prediction of the overall effect [11].

5.3. Training Configuration

Optimizer: Adam with $\beta_1=0.9$, $\beta_2=0.999$, $\epsilon=10^{-8}$. Learning rate: Initial LR= 1×10^{-5} , OneCycleLR scheduler [25] with max LR= 5×10^{-4} . Loss function:

$$\mathcal{L} = \mathcal{L}_{L1} + 0.1 \times \mathcal{L}_{MR-STFT}, \quad (13)$$

where $\mathcal{L}_{MR-STFT}$ is Multi-Resolution STFT loss with FFT sizes [512, 1024, 2048]. Data augmentation: Tempo perturbation [26] with factors [0.9, 1.0, 1.1]. Hardware: NVIDIA A6000 GPU, training time approximately 80 hours. Implementation: PyTorch 1.13.1, torchaudio 0.13.1.

VI. RESULTS AND ANALYSIS

6.1. Comparison with Baseline

We directly compare the performance of the baseline MANNER [3] and our anti-aliased version to verify the efficacy of our framework. As shown in Table 1, the proposed method outperforms the baseline by a significant margin in perceptual metrics, achieving a PESQ of 3.37 (vs. 3.21) and COVL of 4.00 (vs. 3.91). The substantial gain in PESQ (+0.16), which is highly sensitive to signal distortion and noise, indicates that the anti-aliasing modules effectively mitigated the distinct "synthetic noise" often

Table 1. Comparison between baseline MANNER and proposed model.

Method	PESQ	STOI	CSIG	CBAK	COVL
MANNER	3.21	95	4.53	3.65	3.91
Proposed	3.37	94	4.55	3.38	4.00

introduced by time-domain models. While the baseline MANNER is already a strong model for intelligibility (STOI 95), it suffers from perceptual degradation due to aliasing. Our method improves perceptual quality, as reflected by PESQ and COVL, while showing a small decrease in STOI. This indicates a trade-off between reducing aliasing-related perceptual artifacts and preserving fine intelligibility-related temporal cues.

6.2. Additional Experiment on Demucs

To examine whether the proposed framework is specific to MANNER, we additionally evaluate it on Demucs, a widely used waveform-domain encoder-decoder model with explicit down-sampling and up-sampling operations. This makes Demucs suitable for validating the proposed anti-aliasing framework on another representative time-domain architecture. Demucs and Demucs+Proposed Framework were trained under the same local pipeline, and the only differences were BlurPooling, 1D sub-pixel convolution, and ICNR initialization. As shown in Table 2, Demucs+Proposed improves PESQ, CSIG, CBAK, and COVL over the reproduced Demucs baseline. STOI shows only marginal decreases. These results suggest that the proposed framework is not limited to MANNER and can also improve perceptual quality in another time-domain architecture.

6.3. Comparison with State-of-the-Art Models

The highest performance was achieved when BlurPooling,

sub-pixel convolution, and PCS were all applied together. Table 2 shows the speech enhancement models on VoiceBank+DEMAND, which includes time-domain models like SEGAN [9], Conv-TasNet [10], TSTNN [13], DEMUCS [1], SE-Conformer [2], and MANNER [3], and time-frequency models like MetricGAN [27], DB-AIAT [28], and CMGAN [29]. Our model achieves the highest PESQ (3.37) among time-domain models and second-best overall COVL (4.00).

6.4. Ablation Study: Cumulative Analysis

We investigate the role of each component in the proposed framework using a cumulative ablation study. As shown in Table 3, BlurPooling provides the largest improvement, increasing PESQ from 3.21 to 3.37. This is expected because BlurPooling directly suppresses high-frequency components before temporal down-sampling, which is the main source of aliasing in U-Net-based waveform models. In contrast, sub-pixel convolution with ICNR initialization mainly affects the decoder-side up-sampling stage. Its individual metric gain is smaller, but it provides a structured alternative to transposed convolution and helps reduce periodic reconstruction artifacts. PCS provides a marginal additional gain in COVL, as it modifies the perceptual target rather than the aliasing behavior of the network architecture. Overall, the ablation results indicate that down-sampling anti-aliasing is the dominant factor, while decoder-side anti-aliasing and perceptual target construction provide complementary benefits.

6.5. Spectrogram and Error-Spectrogram Analysis

To provide visual evidence for the effect of the proposed anti-aliasing framework, we compare the spectrograms of clean speech, noisy speech, baseline enhanced speech, and proposed enhanced speech. We also visualize error spectrograms computed as the absolute difference between the enhanced and clean log-magnitude spectrograms. As

Table 2. Comparison with speech enhancement models on VoiceBank+DEMAND (bold: best performance, underline: second-best performance).

Method	Domain	PESQ	STOI	CSIG	CBAK	COVL
SEGAN [9]	T	2.16	92	3.48	2.94	2.80
DEMUCS [1] (reproduced)	T	2.52	93	4.05	3.39	3.34
Conv-TasNet [10]	T	2.89	94	3.87	3.31	3.33
MetricGAN [27]	T-F	2.86	-	3.99	3.18	3.42
TSTNN [13]	T	2.96	95	4.10	3.77	3.52
SE-Conformer [2]	T	3.13	95	4.45	3.55	3.82
MANNER [3]	T	3.21	95	4.53	3.65	3.91
DB-AIAT [28]	T-F	3.31	96	<u>4.61</u>	<u>3.75</u>	3.96
CMGAN [29]	T-F	3.41	96	4.63	3.94	4.12
DEMUCS+anti-aliasing methods	T	2.57	93	4.10	3.40	3.39
MANNER+anti-aliasing methods	T	<u>3.37</u>	94	4.55	3.38	<u>4.00</u>

Table 3. Cumulative ablation of anti-aliasing and perceptual target components.

Method	PESQ	STOI	CSIG	COVL
MANNER (baseline)	3.21	95	4.53	3.91
+BlurPooling [6]	3.37	94	4.54	3.99
+sub-pixel conv [8]	3.37	94	4.55	3.99
+PCS [21]	3.37	94	4.55	4.00

shown in Fig. 2, the proposed model preserves the main harmonic and formant structures while reducing spurious spectral energy in several low-energy and high-frequency regions. The highlighted regions in Fig. 2 show that the baseline output contains horizontal spectral components in low-energy regions, especially around the high-frequency band before the main speech segment. In the proposed output, these aliasing-related components are substantially attenuated. The error spectrograms further confirm that the proposed framework reduces residual spectral artifacts in the same time-frequency regions compared with the baseline. These visual observations support the improvements

in perceptual and overall quality metrics such as PESQ and COVL.

6.6. Analysis of STOI and CBAK Trade-Offs

Our method improves PESQ (+0.16), CSIG (+0.02), and COVL (+0.09) on MANNER, but STOI and CBAK decrease. This indicates a trade-off rather than a uniform improvement across all objective metrics. Since BlurPooling applies a low-pass filtering effect before temporal decimation, it can suppress spectral artifacts but may also slightly smooth fine temporal or high-frequency cues related to intelligibility and background-noise perception. The additional Demucs experiment provides a more nuanced view of this trade-off. On Demucs, the proposed framework improves PESQ, CSIG, CBAK, and COVL, while STOI decreases only marginally. Therefore, the metric behavior depends on the baseline architecture, but the overall trend suggests that the proposed framework improves perceptual and overall quality while introducing only a small trade-off in fine-detail preservation.

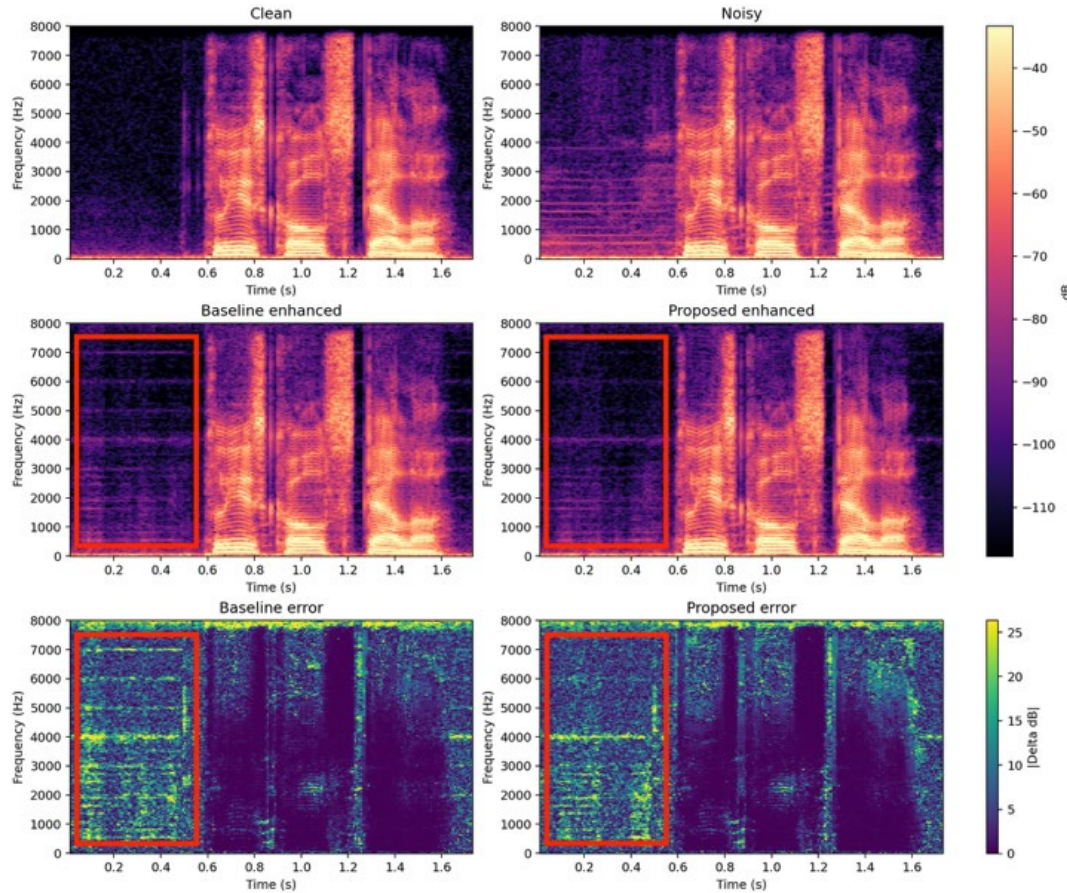


Fig. 2. Spectrogram and error-spectrogram comparison. The first two rows show clean speech, noisy speech, baseline enhanced speech, and proposed enhanced speech. The last row shows the error spectrograms between the enhanced outputs and the clean reference. Compared with the baseline, the proposed framework reduces spurious spectral energy in several time-frequency regions while preserving the main speech structures. The red boxes highlight low-energy regions where the baseline output contains horizontal spectral components, which are substantially reduced in the proposed output.

VII. DISCUSSION

7.1. Why Anti-Aliasing Works

Our results demonstrate that anti-aliasing improves time-domain speech enhancement. From a signal processing perspective, down-sampling and up-sampling in neural networks violate the Nyquist-Shannon theorem when proper low-pass filtering is absent, introducing aliasing, and anti-aliasing methods enforce proper sampling practices. From a perceptual perspective, aliasing-related artifacts can appear as spurious spectral components, especially in high-frequency or low-energy regions. The spectrogram and error-spectrogram analyses in Section 6.5 show that the proposed framework reduces such residual spectral artifacts in several time-frequency regions, which is consistent with the improvements in PESQ and COVL. From a learning perspective, anti-aliasing removes a source of noise from the training signal, allowing the model to focus on learning genuine speech patterns rather than compensating for network artifacts. From a generalization perspective, BlurPooling improves shift-invariance [6], making the model more robust to input variations (translation, slight timing differences).

7.2. Limitations

Our evaluation is limited in dataset scope: we only evaluate on VoiceBank+DEMAND (single dataset, controlled noise), and future work should test on diverse datasets such as DNS Challenge (realistic noise), VCTK (multi-speaker), and real-world recordings (in-the-wild conditions). Additionally, we rely on objective metrics (PESQ, STOI, etc.) for subjective evaluation, and human listening tests (MOS, MUSHRA) would provide stronger evidence of perceptual improvement and validate our hypothesis about STOI/CBAK trade-offs.

VIII. CONCLUSION

This work addresses a fundamental but overlooked issue in time-domain speech enhancement: aliasing artifacts introduced by pooling and up-sampling operations in U-Net architectures. We provided a systematic analysis of aliasing in time-domain speech enhancement, with mathematical formulations (Section 3) showing how strided operations violate the Nyquist-Shannon sampling theorem. We demonstrated that anti-aliasing methods (BlurPooling for down-sampling, sub-pixel convolution for up-sampling) effectively address this issue, achieving improvements of PESQ +0.16 and COVL +0.09 over the MANNER baseline. We achieved competitive performance among time-domain models on VoiceBank+DEMAND (PESQ: 3.37, COVL: 4.00), with minimal computational overhead. We provided cumulative ablation studies (Table 3) showing that anti-

aliasing components contribute consistently to performance improvements. These results suggest that explicit anti-aliasing should be considered as a practical design choice for time-domain encoder-decoder speech enhancement models.

ACKNOWLEDGEMENT

This research was supported by Kumoh National Institute of Technology (2024–2026).

REFERENCES

- [1] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proceedings of the Interspeech*, 2020, pp. 3291-3295.
- [2] E. Kim and H. Seo, "SE-conformer: Time-domain speech enhancement using conformer," in *Proceedings of the Interspeech*, 2021, pp. 2736-2740.
- [3] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "MANNER: Multi-view attention network for noise erasure," in *Proceedings of the ICASSP*, 2022, pp. 7842-7846.
- [4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the MICCAI*, 2015, pp. 234-241.
- [5] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, and J. Yu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of the Interspeech*, 2020, pp. 5036-5040.
- [6] R. Zhang, "Making convolutional networks shift-invariant again," in *Proceedings of the ICML*, 2019, pp. 7324-7334.
- [7] W. Shi, J. Caballero, F. Huszar, J. Totz, A. Aitken, and R. Bishop, et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the CVPR*, 2016, pp. 1874-1883.
- [8] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *arXiv Preprint arXiv:1707.02937*, 2017.
- [9] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proceedings of the Interspeech*, 2017, pp. 3642-3646.
- [10] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019.
- [11] Y. Hu and P. C. Loizou, "Evaluation of objective

- quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229-238, 2008.
- [12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [13] K. Wang, B. He, and W. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *Proceedings of the ICASSP*, 2021, pp. 7098-7102.
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. Gomez, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [15] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the ICML*, 2010, pp. 807-814.
- [16] K. Akeley, "Reality engine graphics," in *Proceedings of the SIGGRAPH*, 1993, pp. 109-116.
- [17] J. Jimenez, J. Echevarria, T. Sousa, and D. Gutierrez, "SMAA: Enhanced subpixel morphological anti-aliasing," *Computer Graphics Forum*, vol. 31, no. 2, pp. 355-364, 2012.
- [18] N. Park and S. Kim, "Blurs behave like ensembles: Spatial smoothings to improve accuracy, uncertainty, and robustness," in *Proceedings of the ICML*, 2022, pp. 17390-17419.
- [19] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, pp. e3, 2016.
- [20] P. Ramachandran, B. Zoph, and Q. Le, "Swish: a self-gated activation function," *arXiv Preprint arXiv:1710.05941*, 2017.
- [21] R. Chao, C. Yu, S. W. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," in *Proceedings of the Interspeech*, 2022, pp. 3298-3302.
- [22] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proceedings of the SSW*, 2016, pp. 146-152.
- [23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the ICASSP*, 2001, pp. 749-752.
- [24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.
- [25] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Proceedings of the SPIE Defense + Commercial Sensing*, 2019, pp. 369-386.
- [26] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of the Interspeech*, 2015, pp. 3586-3589.
- [27] S. W. Fu, C. F. Liao, Y. Tsao, and S. D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the ICML*, 2019, pp. 2031-2041.
- [28] G. Yu, A. Li, Y. Wang, Y. Guo, H. Wang, and C. Zheng, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proceedings of the ICASSP*, 2022, pp. 7847-7851.
- [29] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proceedings of the ICASSP*, 2022, pp. 1-5.

AUTHORS



Joonhyeon Bae received his B.S. degree in the Department of Computer Engineering from Kumoh National Institute of Technology, Korea, in 2022. He received his M.S. degree in Interdisciplinary Program in Artificial Intelligence from Seoul National University. His research interests include Speech Enhancement, Multimodal Retrieval, Neural Audio Synthesis.



Jaepil Ko received the B.S., M.S., and Ph.D. degrees from Yonsei University, Korea, in 1996, 1998, and 2004, respectively. Since 2004, he has been a Professor with the Computer Engineering Department, Kumoh National Institute of Technology, Korea. His research interests include computer vision, machine learning, and image processing.

