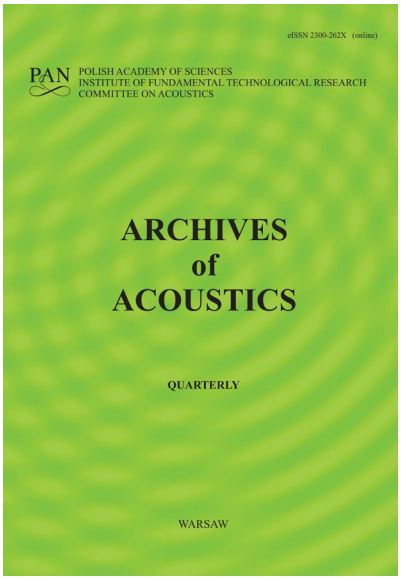
JOURNAL PRE-PROOF

This is an early version of the article, published prior to copyediting, typesetting, and editorial correction. The manuscript has been accepted for publication and is now available online to ensure early dissemination, author visibility, and citation tracking prior to the formal issue publication.

It has not undergone final language verification, formatting, or technical editing by the journal's editorial team. Content is subject to change in the final Version of Record.

To differentiate this version, it is marked as "PRE-PROOF PUBLICATION" and should be cited with the provided DOI. A visible watermark on each page indicates its preliminary status.

The final version will appear in a regular issue of *Archives of Acoustics*, with final metadata, layout, and pagination.



Title: Adversarial Audio Inpainting with Selective State Space Model, Efficient Attention and Large-Scale Pre-Trained Model

Author(s): Junkang Yang, Hongqing Liu, Liming Shi, Lu Gan, Hiromitsu Nishizaki, Chee Siang Leow

DOI: https://doi.org/10.24423/archacoust.2026.4269

Journal: *Archives of Acoustics*

ISSN: 0137-5075, e-ISSN: 2300-262X

Publication status: In press

Received: 2025-06-25

Revised: 2026-03-14

Accepted: 2026-04-12

Published pre-proof: 2026-04-15

Please cite this article as:

Yang J., Liu H., Shi L., Gan L., Nishizaki H., Leow C.S. (2026), Adversarial Audio Inpainting with Selective State Space Model, Efficient Attention and Large-Scale Pre-Trained Model, *Archives of Acoustics*, https://doi.org/10.24423/archacoust.2026.4269

Adversarial Audio Inpainting with Selective State Space Model, Efficient Attention and Large-Scale Pre-Trained Model

Junkang YANG[1], Hongqing LIU[2*], Liming SHI[3], Lu GAN[4], Hiromitsu NISHIZAKI[1], and Chee Siang LEOW[1]

[1]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Kofu, Japan

[2]Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

[3]School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

[4]College of Engineering, Design and Physical Science, Brunel University, London, U.K.

[*]Corresponding Author e-mail: hongqingliu@cqupt.edu.cn

Abstract

Recent deep-learning based speech enhancement algorithms have many applications in the areas of noise reduction, de-reverberation, bandwidth extension, echo cancellation, to name a few. Packet loss is also one of the main causes of voice quality degradation in VoIP calls. Currently, generative adversarial networks (GANs) have shown a strong ability in image generation, and many of those models also work well in speech tasks. In this work, we propose a light-weight model based on GAN to handle the task of audio packet loss concealment. Specifically, we use a U-shaped network operating in the time-frequency domain as a generator, which is trained by a Mel-GAN discriminator with multi-loss. In addition, to enhance the model's performance under unfavorable channels, we introduce noise and bandwidth loss in the training data. The experiments show that our method outperforms the baseline in both objective and subjective metrics under an ideal channel with no other distortions, and it still largely maintains its performance in the presence of noise and bandwidth loss.

Keywords: adversarial learning; audio inpainting; selective state space model; attention mechanism.

1 Introduction

Speech signals are usually broken down into frames for transmission in the channel, and these frames can be acknowledged as data packets during transmission. These packets are often transmitted in a communication system in a disordered fashion and are reassembled into an ordered sequence at the receiving end. When there is interference in the channel or the communication link experiences interuptions, it will produce missing packets or high packet jitter, causing the degradation of speech quality. In applications such as mobile digital communications, videoconferencing systems, and voice over Internet protocol (VoIP) calls, packet loss can dramatically affect the quality of speech to the extent that people cannot understand semantic information from intermittent speech. At the same time, due to the variety of distortion types in the channel and the high real-time nature of these scenarios, it is a challenging task to consider both robustness and lightweight when repairingral networks.

Packet loss concealment (PLC) is the technology to reconstruct the lost part of speech using existing information (Mohamed and Schuller 2020). Some techniques refer to a similar task with the terms audio inpainting (Miotello et al. 2024), waveform interpolation (Lagrange et al. 2005), or extrapolation (Maher 1994; Aironi et al. 2023). Early PLC methods were mainly based on statistical modeling, such as hidden Markov models (HMMs) (Rodbro et al. 2006) or different coding approaches (Janicki and Ksiundefinedżak 2008). As learning-based methods continue to improve, the performance of these methods gradually becomes obsolete. Using deep learning techniques, Lee and Chang 2015 first proposed a DNN-based model to conceal the degradation caused by lost packets, generating the estimated log-power spectra and phases of missing frames and fixing the waveform by putting these features into a decoder. In the subsequent works, recurrent network structures and convolutional modules have been widely used for this task (Mohamed and Schuller 2020; Lotfidereshgi and Gournay 2018b; Lin et al. 2021; Davy et al. 2023). These methods can better capture all sorts of dependencies between samples and improve the overall performance. In recent years, many generative models such as GAN and diffusion models have achieved good performance in the field of speech enhancement (Fu et al. 2021; Yen et al. 2023). In the PLC task, GAN based model can allow the generator to achieve better performance, compared to autoregressive algorithms, with less parameters through adversarial training (Ebner and Eltelt 2020). For example, Zhao 2023 uses a GAN network to repair the magnitude together with a phase reconstruction algorithm, implementing the speech inpainting for audio editing software. Diffusion models are often applied to audio joint task including PLC or inpainting. Like Moliner et al. 2023, which solves bandwidth extension, de-clipping, and inpainting problems with a general diffusion model and achieves convincing results, unfortunately, it seems not suitable for speech and real-time communication systems. In INTERSPEECH 2022, Microsoft organized the 1st Audio Deep Packet Loss Concealment Challenge (Diener, Sootla, et al. 2022) and it received many effective models (Li et al. 2022; Liu et al. 2022; Westhausen and Meyer 2022; Valin, Mustafa, et al. 2022; Guan et al. 2022), including GAN, RNN, and the deep learning methods mixed with traditional ways. This challenge complements this research field, and it is a great reference for

later works.

Currently, the difficulties in the PLC task still exist. When expecting a stronger capability in loss packet reconstruction, more parameters are required in most instances. However, the deployment of the model on the resource-limited devices is so sensitive to this aspect. Furthermore, signals are facing various types of loss in real-world channels. When other distortions exist together with packet loss, most of the systems will have a large degradation in performance, and most of the related work does not evaluate the performance of the model in real-world scenarios. Besides, after front-end speech enhancement algorithms come various downstream tasks such as automatic speech recognition (ASR), speech emotion recognition (SER), and multi-model large language models. Improving effectiveness in front-end processing for downstream tasks is also an issue that needs attention.

To handle these problems, this work proposes a GAN-based model to conceal the packet loss. Our main contributions are as follows.

- By introducing a state space model and efficient local attention in the generator, our method achieves a better performance than the baselines while remaining lightweight.
- By using ASR and speech quality evaluation pretrained models to generate the loss function at the training stage, we improve the model's performance in downstream ASR tasks and subjective evaluation experiments.
- We constructed training data containing bandwidth loss and noise, which is closer to the real-world situation. Experiments show that our model trained on these data presents a higher robustness in the real-world data.

2 Proposed method

Figure 1 illustrates the GAN proposed in this paper, which is primarily composed of a generator and a discriminator. The generator is responsible for transforming input spectrograms or waveforms into enhanced outputs, while the discriminator evaluates the quality of the generator's outputs and generates multi-loss to optimize the entire network.

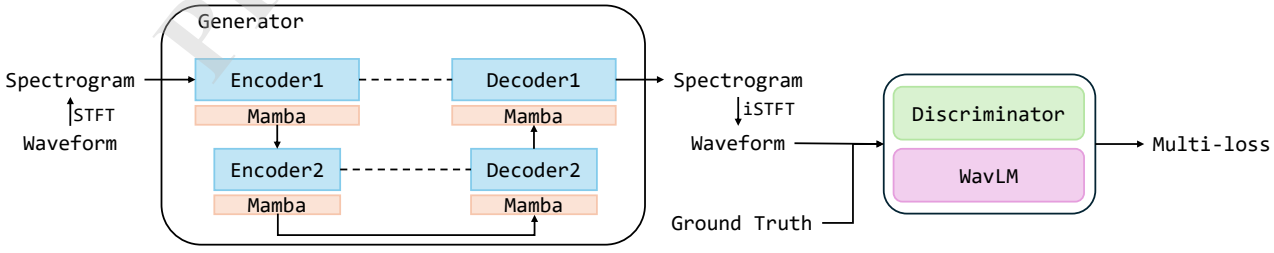


Figure 1: Structure of proposed generative adversarial network.

The generator is the core component of the network, accepting spectrograms as input, where it is first converted into spectrograms via the short-time Fourier transform (STFT). The

spectrograms are then fed into two separate encoder-decoder paths within the generator. Each encoder and decoder incorporates a Mamba block for feature modeling and fusion. At the end, the output spectrogram is converted back into a waveform via inverse short-time Fourier transform (iSTFT). It is worth mentioning that there is a skip connection between each encoder and its corresponding decoder.

Discriminator is also a crucial component of this network, whose primary function is to distinguish between fake waveforms and real waveforms. It takes the output waveforms from the generator and real waveforms as input, analyzes and evaluates the input data through a multi-layer neural network structure, and outputs a probability value or feature representation indicating the authenticity of the input data. Discriminator provides the generator with an adversarial loss function, encouraging the generator to continuously optimize the quality of the generated waveforms, making them closer to the distribution and characteristics of real waveforms.

Furthermore, WavLM (Chen et al. 2022) is a transformer (Vaswani et al. 2017) based pre-trained audio language model that serves as an auxiliary discriminative component in this system, working alongside the discriminator as part of a multi-loss calculation framework. WavLM conducts in-depth semantic and feature-level analysis of input waveforms, leveraging its pre-trained knowledge and feature extraction capabilities acquired from large-scale audio data to evaluate the differences and similarities between generated and real waveforms from multiple perspectives. This provides richer supervisory signals during model training, helping the generator produce waveforms that better align with the semantic and acoustic characteristics of real audio. By enhancing both the quality and naturalness of the generated audio, WavLM enables the entire model to holistically consider various factors such as waveform authenticity and semantic features in audio generation tasks.

The entire workflow is as follows. The input waveform is converted into a spectrogram via STFT, and the spectrogram is fed into the generator. After feature processing by the encoder, the decoders reconstruct the spectrogram. Then the output spectrogram is converted back into a waveform via iSTFT. The generated waveform and the ground-truth waveform are then served as inputs into the discriminator and pre-trained model to compute the adversarial loss and audio quality loss, respectively. The total loss is used to update the parameters of the generator through backpropagation and gradient descent, progressively improving the generator's output quality.

We will introduce each component of the network in detail next.

2.1 Generator

The generator mainly contains encoders, decoders and state space model blocks (Mamba).

2.1.1 Encoder and decoder

The encoder–decoder in Figure 2 is designed under 3 practical constraints of PLC. First, PLC requires restoring locally time-frequency patterns such as harmonics. These features are

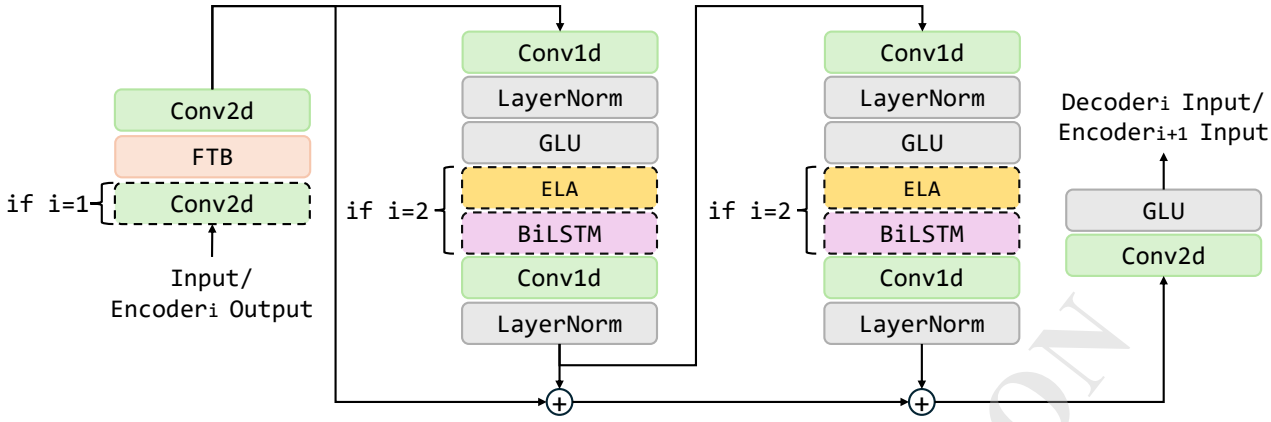


Figure 2: Structure of encoder.

typically disrupted by burst losses. So we use lightweight convolutions as the main building modules to efficiently capture local structures with low latency. Besides, accurate inpainting also needs contextual cues beyond a short receptive field, especially when consecutive packets are missing; therefore, we need to enhance cross-frequency interactions and long-context modeling at the bottleneck rather than stacking heavy global attention everywhere. Additionally, the model is intended for real-time or resource-limited deployment, so each additional module must provide clear gains per computational cost.

Based on these considerations, we adopt a U-Net style encoder–decoder with skip connections to preserve fine-grained spectral details and to reduce over-smoothing during reconstruction. As illustrated in Figure 2, the encoder first applies 2D convolution to extract local time-frequency features, followed by a frequency transformer block (FTB) (Dai et al. 2024) to strengthen frequency-wise dependency modeling, since packet loss often breaks spectral continuity across frequency bins. When operating in the first encoder layer ($i = 1$), an additional 2D convolution and GLU activation are used to selectively filter informative channels and suppress artifacts introduced by lossy segments. Subsequently, 1D convolution and layer normalization are employed to capture short-term temporal correlations and stabilize training. In the second encoder layer ($i = 2$), we introduce efficient local attention (ELA) (Xu and Wan 2024) together with a bidirectional LSTM (BiLSTM) near the bottleneck, where the temporal resolution is already reduced; this design allows the model to capture salient regions and longer-range dependencies with minimal computational overhead. The processed features are then combined with earlier-stage representations through skip connections, mitigating gradient vanishing and facilitating multi-level feature fusion.

ELA is an efficient local attention mechanism for deep convolutional neural networks. It improves upon coordinate attention by using 1D convolution and group normalization to enhance feature representation, avoiding the generalization issues of batch normalization in coordinate attention and the negative effects of channel dimension reduction on attention generation. Specifically, as Figure 3 depicts, ELA first employs strip pooling to extract one-dimensional feature vectors in the horizontal and vertical directions from the input feature map. Given an

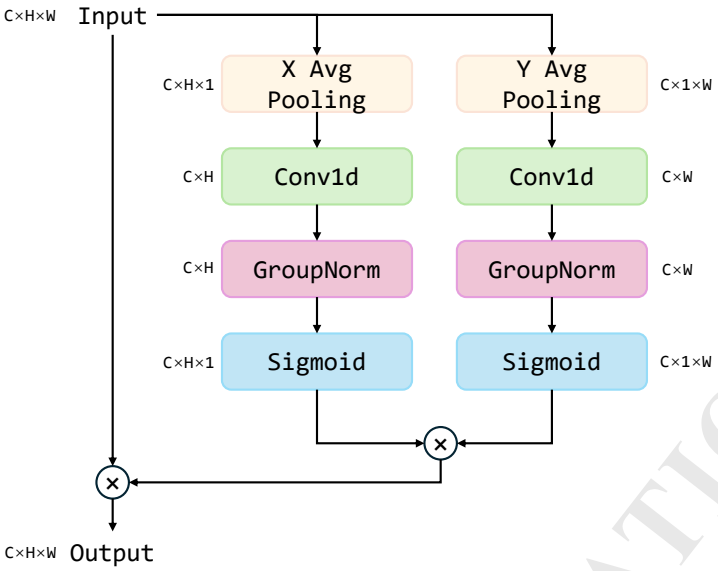


Figure 3: Efficient local attention.

input tensor $X \in \mathbb{R}^{C\times H\times W}$, the operation is

$$X_h(c,w) = \frac{1}{H}\sum_{i=1}^{H} X(c,i,w), \tag{1}$$

$$X_w(c,h) = \frac{1}{W}\sum_{j=1}^{W} X(c,h,j). \tag{2}$$

where $X_h \in \mathbb{R}^{C\times W}$ and $X_w \in \mathbb{R}^{C\times H}$ are 2D feature maps (matrices) obtained by average pooling along the height and width dimensions, respectively. Here, $c \in \{1,\ldots,C\}$ denotes the channel index, $h \in \{1,\ldots,H\}$ and $w \in \{1,\ldots,W\}$ denote the spatial indices along height and width, respectively.

After pooling, a 1D convolution is applied along the remaining spatial dimension to capture local contextual interactions. Specifically, the convolution operates along the width dimension for X_h and along the height dimension for X_w, while being applied independently for each channel. After that, group normalization and Sigmoid activation functions are used to process the feature maps, generating positional attention predictions for both directions, given by

$$Y_h = \sigma(\mathrm{GN}(\mathrm{Conv}(X_h))), \tag{3}$$

$$Y_w = \sigma(\mathrm{GN}(\mathrm{Conv}(X_w))), \tag{4}$$

where $\sigma(\cdot)$, $\mathrm{GN}(\cdot)$, $\mathrm{Conv}(\cdot)$ represent Sigmoid fuction, group normalization, and 1D convolution, respectively. Finally, the predictions from both directions are combined via a product operation to obtain the final positional attention map. That is,

$$Y = X \cdot Y_h \cdot Y_w, \tag{5}$$

where $\cdot$ denotes elementwise multiplication. The attention maps Y_h and Y_w are broadcast along the height and width dimensions, respectively, before being applied to the input feature map X.

ELA can accurately locate regions of interest without reducing the channel dimension, thus avoiding the loss of feature information caused by channel dimension reduction. Additionally, ELA has a lightweight structure with low computational complexity. For example, in the medical image detection task (Sun et al. 2025; Hao et al. 2024), the model using ELA as the core component effectively improved the accuracy while having fewer parameters and computational complexity than the baselines. ELA also shows good generalization ability, stably enhancing the performance of both large and small networks. For instance, it significantly improves the accuracy of ResNet (He et al. 2016) series networks with minimal impact on model parameters (Xu and Wan 2024). These conclusions can also be verified in our experiments.

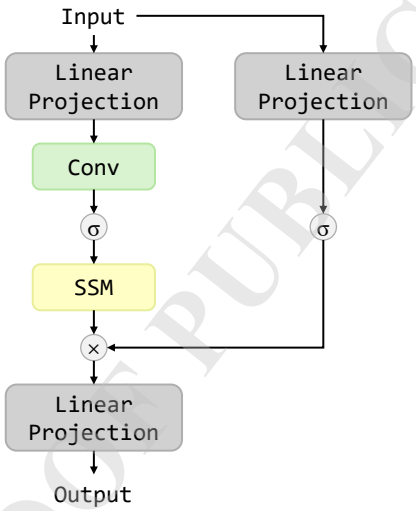


Figure 4: Mamba block.

2.1.2 State space model block

Packet loss concealment requires the model to infer missing speech segments from the surrounding context, particularly under burst loss conditions where consecutive frames are unavailable. While convolutional layers are effective at modeling local structures, their receptive field grows slowly with depth, and recurrent or attention-based modules often incur high computational cost. To address this challenge, we introduce a selective state space model (SSM) based on the Mamba architecture (Gu and Dao 2023) as a core temporal modeling component in the generator. Its detailed description is shown in Figure 4.

The key motivation for using Mamba lies in its ability to model long-range temporal dependencies with linear computational complexity. Unlike Transformer-based attention, whose cost scales quadratically with sequence length, the state space formulation allows efficient propagation of contextual information across long time spans, which is particularly important for reconstructing speech segments affected by continuous packet losses. Moreover, the selective

mechanism in Mamba enables the model to dynamically control information flow, allowing it to emphasize informative regions while attenuating less relevant context.

In our architecture, the Mamba block is placed within the encoder–decoder pipeline to complement convolutional feature extraction. While convolutional layers and ELA focus on local time-frequency structures, the Mamba block aggregates long-range temporal information across frames, thereby improving the continuity and naturalness of reconstructed speech. This design allows the proposed model to effectively handle both short and long packet losses while maintaining a favorable trade-off between performance and computational efficiency.

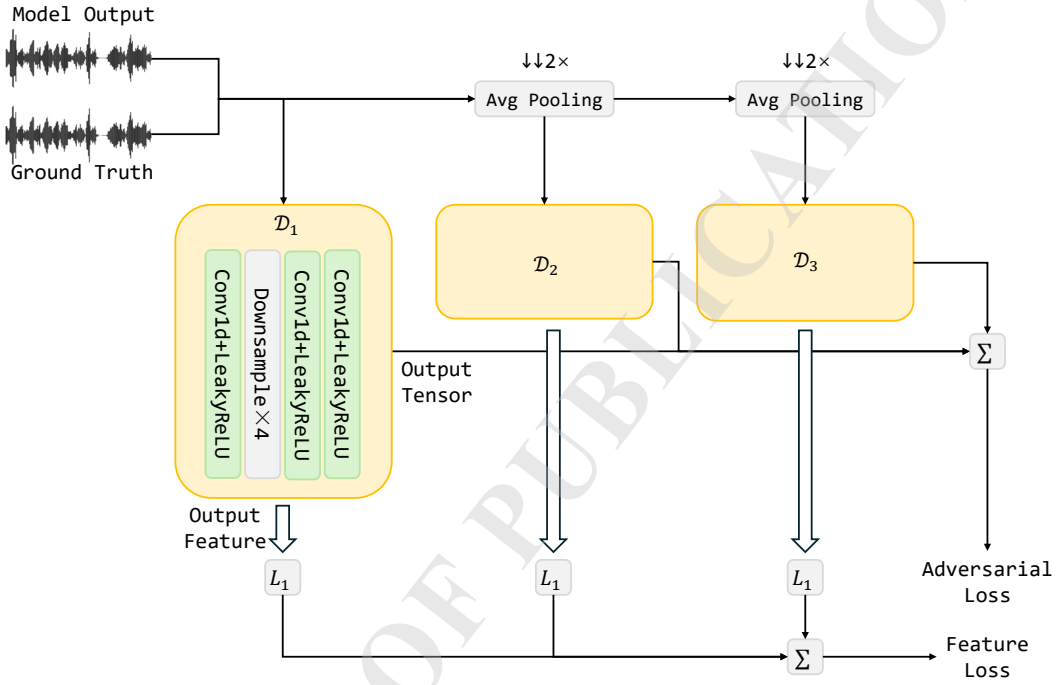


Figure 5: Structure of discriminator.

2.2 Discriminator

Inspired by J. Yang et al. 2024, we employed a multi-scale discriminator architecture which is shown in 5 as the core component of the adversarial training framework to optimize the generator's performance. This discriminator consists of three independent sub-discriminators $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$, each designed with inspiration from MelGAN (Kumar et al. 2019). Specifically, each sub-discriminator comprises seven convolutional layers, with the first four layers featuring downsampling capabilities to progressively compress the spatiotemporal resolution of the input signal and capture multi-scale features. The input to the discriminator undergoes differentiated processing: $\mathcal{D}_1$ receives the raw waveform, $\mathcal{D}_2$ processes a 2 times downsampled waveform, and $\mathcal{D}_3$ analyzes a 4 times downsampled waveform. This design enables the generator to maintain consistency across different temporal scales, thereby enhancing the naturalness of the output speech and its spectral coherence. Furthermore, the multi-level features output by the discriminator are used to compute the adversarial loss and feature loss. The former optimizes

the generator by minimizing the difference between generated and real samples, while the latter constrains the generator to align its intermediate-layer feature distribution with that of real data.

2.3 WavLM

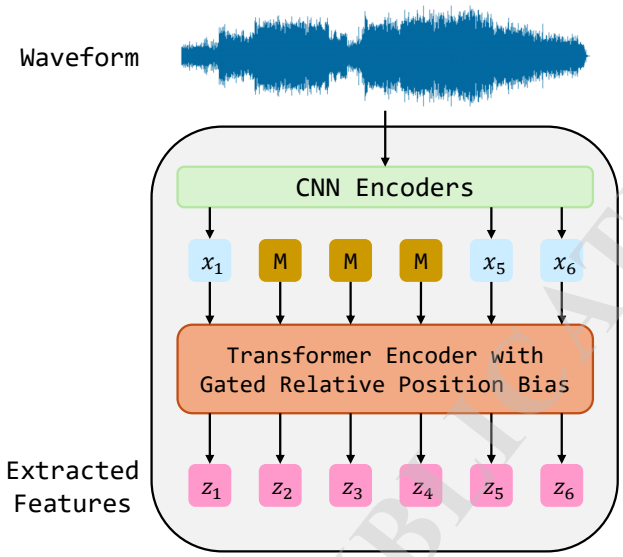


Figure 6: Structure of WavLM.

The architecture of WavLM (Chen et al. 2022) is shown in Figure 6, and it is composed of a CNN encoder and a transformer encoder, with the latter incorporating gated relative position bias (Chi et al. 2022) to better capture the sequence order of input speech. The CNN encoder is tasked with extracting features from the input waveform. It consists of 7 temporal convolutions, each followed by layer normalization and a GELU activation layer. The convolutional kernels have 512 channels, with strides of (5,2,2,2,2,2,2) and widths of (10,3,3,3,3,2,2), ensuring each output represents approximately 25 ms of audio with a stride of 20 ms. The features x extracted by the CNN are masked and then fed into the transformer encoder as input. At the bottom of the transformer encoder is a convolution-based relative position embedding layer (Raffel et al. 2020) with a kernel size of 128 and 16 groups. To enhance model performance, WavLM introduces gated relative position bias, which is encoded based on the offset between key and query in self-attention. Specifically, for the input hidden states $\{\mathbf{h}_i\}_{i=1}^{T}$, each $\mathbf{h}_i$ is linearly projected into a query, key, and value $(\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i)$. Here, $\mathbf{h}_i \in \mathbb{R}^D$ denotes the hidden representation at time step i output by the WavLM transformer encoder, where D is the hidden dimension of the pre-trained model. In our implementation, we use WavLM Base, for which $D = 768$. The self-attention outputs are calculated by incorporating attention logits with gated relative position bias r_{i-j}. The calculations are

$$g_i^{(update)}, g_i^{(reset)} = \sigma(\mathbf{q}_i \cdot \mathbf{u}), \sigma(\mathbf{q}_i \cdot \mathbf{w}), \tag{6}$$

$$\tilde{r}_{i-j} = \mathbf{w} \cdot g_i^{(reset)} \cdot d_{i-j}, \tag{7}$$

$$r_{i-j} = d_{i-j} + g_i^{(update)} \cdot d_{i-j} + (1 - g_i^{(update)}) \cdot \tilde{r}_{i-j}, \tag{8}$$

Here, $\mathbf{u}, \mathbf{w}$ are learnable parameter tensors with the same dimensions as the query representation $\mathbf{q}_i$, and the dot product in equation (6) denotes an inner product that produces scalar gating values. $\sigma(\cdot)$ represents sigmoid function. The gating mechanism allows the relative position bias to be dynamically adjusted based on the current speech content. Additionally, d_{i-j} is calculated using bucketed relative position embeddings, with its range varying according to the offset. Consequently, WavLM learns rich speech representation capabilities through its pre-training framework of masked speech denoising and prediction, enabling it to capture multi-level information in speech signals, such as content, speaker characteristics, and environmental noise.

In our task, using the deep features extracted by WavLM to compute the loss function can guide the model to not only focus on waveform-level matching but also optimize higher-level semantic and perceptual characteristics of speech. The WavLM-based loss can measure the distribution difference between enhanced speech and clean speech in the latent space, thereby avoiding the over-smoothing issues that may arise from low-level metrics, resulting in more natural enhanced speech with better-preserved details.

2.4 Loss

The loss function of our model adopts an adversarial training approach, combining multiscale STFT loss, generator adversarial loss, feature loss, and pre-trained model loss. The total loss function can be expressed as

$$\mathcal{L} = \mathcal{L}_{MSTFT} + \mathcal{L}_{adv}^{\mathcal{G}} + \lambda_f \mathcal{L}_f + \mathcal{L}_{pre}, \tag{9}$$

where $\lambda_f = 100$, which is used to balance the weight of feature loss.

The multi-scale STFT loss L_{MSTFT} is

$$\mathcal{L}_{MSTFT} = \mathbb{E}_{(x,y)\sim p_{data}} \left[\sum_{m=1}^{3} \left(\frac{||s(y,\theta_m) - s(\mathcal{G}(x),\theta_m)||_F}{||s(y,\theta_m)||_F} + \frac{1}{N} ||log \frac{s(y,\theta_m)}{s(\mathcal{G}(x),\theta_m)}|| \right) \right], \tag{10}$$

where $s(\mathcal{G}(x),\theta_m)$ represents the STFT magnitude spectrum of the output signal $\mathcal{G}(x)$ under the parameter θ_m, $||\cdot||_F$ and $||\cdot||$ denote the Frobenius norm and ℓ_1 norm, respectively, and N is the number of elements in the magnitude spectrum. The multi-scale parameters θ_m cover different FFT sizes (512, 1024, 2048) and hop length (50, 120, 240). The logarithm in equation (10) is elementwise.

The generator adversarial loss $L_{adv}^{\mathcal{G}}$ and feature loss L_f are

$$\mathcal{L}_{adv}^{\mathcal{G}} = \mathbb{E}_{x\sim p_x} \left[\frac{1}{K} \sum_k \max(0, 1 - \mathcal{D}_k(\mathcal{G}(x))) \right], \tag{11}$$

$$\mathcal{L}_f = \mathbb{E}_{(x,y)\sim p_{data}}\left[\frac{1}{KL}\sum_{k,l}||\mathcal{D}_k^l(y) - \mathcal{D}_k^l(\mathcal{G}(x))||_1\right], \tag{12}$$

where p_{x} is the distribution of lossy speech, p_{data} denotes the empirical joint distribution of lossy speech and clean speech from the training dataset, K is the number of discriminators ($K = 3$), L is the number of layers in the discriminator, and $\mathcal{D}_k^l$ denotes the feature output of the l-th layer in the k-th discriminator. This loss design achieves noise-robust bandwidth extension by jointly optimizing time-frequency domain reconstruction accuracy (STFT loss) and adversarial training (adversarial loss and feature loss).

The pre-trained model loss $\mathcal{L}_{pre}$ is the KL divergence between the output features of prediction x and ground truth y from WavLM.

2.5 Datasets

We used clean speech data from 2023 DNS challenge (Dubey et al. 2023) and VCTK (Yamagishi et al. 2019), noise data from 2023 DNS challenge (Dubey et al. 2023) and WHAM! (Wichern et al. 2019) in the training phase of the model. We first added noise to the clean speech data with a probability of 0.5, making the SNR of speech in the range of -10 - 0 dB, and then randomly masked it to simulate the packet loss effect, with a mask window length of $t_{win} \in [20, 120]$ (ms). We synthesized a total of 300,000 samples, each of which was normalized to a length of 5 seconds and resampled to 16 kHz, with a total duration of 416.67 hours. During training, we used 90% of the data as the training set and 10% as the validation set.

In the testing phase, we used the blind test set from the 2022 PLC challenge (Diener, Sootla, et al. 2022) and the last 5 speakers in the VCTK dataset (Yamagishi et al. 2019) for testing. For the VCTK test data, we set the duration of lost packets to a multiple of 20 ms, and the maximum packet loss time length to 120 ms, equivalent to 6 consecutive packets. The distribution of lost parts when the packet loss rate is in the range of 10% - 40% is completely based on the settings in (Aironi et al. 2023). In addition, we used real speech collected from the wireless communication system of marine ships to evaluate the performance of the model in real scenarios. According to the coarse estimation, in this dataset, the packet loss rates approximately range from 0% to 40% with burst durations on the order of 0-100 ms. We made this part of the data public on this link[1].

2.6 Detailed configurations

During model training, we employ the Adam optimizer (with $\beta_1 = 0.8$ and $\beta_2 = 0.999$) to jointly optimize the generator and discriminator under a fixed learning rate of 1×10^{-3}. The training is conducted on 2 NVIDIA RTX3090 GPUs for 100 epochs, and the checkpoint achieving the best performance on the validation set is selected for testing. For the STFT configuration in our network, the FFT size is set to 512 with a hop length of 128. In addition,

[1] https://drive.google.com/file/d/182xMlO768kH3ntOO8C_vNIOqYZOIR4cu/view?usp=sharing

we used the "WavLM Base" pre-trained model provided by the official repository[2] to generate pre-trained model loss.

2.7 Evaluation metrics

In the evaluations of the model, we selected a set of multidimensional metrics to comprehensively measure speech quality, comprehensibility, and recognition performance. Specifically, we conducted tests on the following metrics in the experiments.

DNSMOS (Reddy et al. 2022) is an objective speech quality evaluation index based on neural networks, which simulates human subjective auditory perception using deep networks. PLCMOS (Diener, Purin, et al. 2023), like DNSMOS, is based on neural networks. The difference is that it focuses on evaluating the quality of packet loss concealment in communication scenarios, modeling through perceptual linear prediction coefficients and auditory masking effects. Its calculation process considers frequency band energy distribution and psychoacoustic characteristics. The range of values for these two metrics is 1 to 5, and they are both non-intrusive, which means that no reference signal is required during the evaluation.

Additionally, we employed the PESQ (Rix et al. 2001), a widely used objective metric in the field of speech enhancement. PESQ is an objective speech quality assessment method based on the human auditory model, primarily used to measure the perceptual quality of speech signals after processes such as encoding or decoding, transmission, or noise reduction. Its output scores typically range from -0.5 to 4.5, with higher scores indicating better quality. We also used STOI (Taal et al. 2011), which focuses on the objective evaluation of speech intelligibility. By analyzing the time-frequency characteristics of speech signals, STOI predicts how easily a listener can understand the speech, with scores ranging from 0 to 1, and a score closer to 1 indicates higher intelligibility. It is worth mentioning that both metrics require reference signals for testing.

On the other hand, we randomly selected 25 native English-speaking males and 25 native English-speaking females online and asked them to listen to speech signals processed by different models to obtain mean opinion score (MOS) scores. This subjective evaluation follows a MOS listening test in accordance with the general guidelines of ITU-T Recommendation P.800. MOS is a subjective evaluation metric where human listeners rate speech quality based on auditory perception, typically on a scale from 1 (poor) to 5 (excellent), directly reflecting subjective human perception. The average of all individual scores was taken as the final metric.

In order to measure the impact of different models on downstream tasks, we use speech processed by these models for ASR and use word accuracy (WAcc) as an indicator. We use the same ASR model for all methods.

[2] https://github.com/microsoft/unilm/tree/master/wavlm

2.8 Baseline models

We tested the proposed model and all the models submitted to the PLC Challenge (Diener, Sootla, et al. 2022) on the blind test set provided by the challenge, and all training and testing conditions were kept consistent. In addition, we compared 3 traditional algorithms: Opus (Valin, Maxwell, et al. 2016), WebRTC (Blum et al. 2021), EVS (Lecomte et al. 2015), and several deep learning-based models such as PLAAE (Pascual et al. 2021), TFGAN(Wang et al. 2021), bin2bin (Aironi et al. 2023) based on generative adversarial networks, a model based on recurrent neural networks (Lotfidereshgi and Gournay 2018a), and the MFM pre-training method and the model in (D.-H. Yang et al. 2023) on VCTK and our collected speech datasets. We adopted the settings claimed in the papers of these models and retrained and reimplemented them when necessary. We also referred to the experimental results in (Aironi et al. 2023).

Table 1: Performance comparison on 2022 PLC challenge blind test data.

Team ID*	PLCMOS	DNSMOS	CMOS	Wacc	#Parameters
Clean (reference)	4.51	3.89	0	0.98	-
#12	4.28	3.80	-0.55	0.85	2.36
#1	3.74	3.79	-0.64	0.84	5.9
#9	3.83	3.68	-0.81	0.80	-
#14	3.98	3.69	-0.84	0.79	4.97
#11	3.28	3.51	-1.10	0.75	-
#10	3.48	3.74	-1.04	0.74	-
#6	2.90	3.48	-1.31	0.71	3.77
Proposed model	**4.34**	**3.85**	**-0.42**	**0.89**	**1.71**

* Since the methods and names of most models in this challenge are unknown, team numbers are used to represent the models.

3 Experiment results

Table 1 presents a comparison of the performance of different models on the 2022 PLC challenge blind test data, focusing on key metrics such as PLCMOS, DNSMOS, CMOS, and WAcc, where CMOS is defined by

$$\text{CMOS} = \text{MOS} - \text{MOS}_{clean}. \tag{13}$$

In PLCMOS, which evaluates the quality of inpainting, the score of clean speech is 4.51. The proposed model achieves 4.34, the closest to the reference among all teams, significantly outperforming others like #12 at 4.28 and #1 at 3.74. This highlights the model's ability to maintain signal quality close to the ideal reference. For DNSMOS, measuring overall audio quality, the reference's score is 3.89. The proposed model scores 3.85, again leading other teams such as #9 at 3.68 and #14 at 3.69. This indicates the model's effectiveness in preserving audio quality. As for CMOS, which indicates the subjective feeling of human beings' distortion, the

reference is 0. The proposed model's score of -0.42 is much closer to zero compared to teams like #6 at -1.31 and #11 at -1.10, showing a better experience brought to the listeners. Finally, in WAcc, the ASR performance metric, the reference is 0.98, and the proposed model attains 0.89, surpassing other teams like #12 at 0.85 and #1 at 0.84. This shows the proposed model demonstrates superior performance across all key metrics, effectively balancing signal quality, audio fidelity, and minimal distortion, making it the best solution among these models. With a parameter number of 1.71M, our model achieves a PLCMOS score of 4.34, a DNSMOS score close to 3.85, and an impressive WAcc of 0.89, outperforming all other models across these three dimensions. Besides, Model #1, with a parameter number of 5.90M, which is nearly three times that of our model, only reaches a WAcc of 0.85, while its PLCMOS and DNSMOS scores show no significant advantage. Model #12, with 2.36M parameters, which is still 0.65M more than ours, achieves a WAcc of approximately 0.85, and its PLCMOS and DNSMOS performance is also lower than our model. Model #14, with 4.97M parameters, also fails to match our model's levels in PLCMOS, DNSMOS, and WAcc. Similarly, Model #6, with 3.77M parameters, lags behind our model in all three key metrics.

Table 2: Average performance comparison on VCTK data.

Type	Method	PESQ	STOI
Traditional	Opus	1.77	0.77
	WebRTC	1.70	0.70
	EVS	1.89	0.78
Learning-based	PLAAE	2.04	0.84
	TFGAN	1.97	0.81
	bin2bin	2.72	0.88
	RNN	2.23	0.83
	MFM	2.52	0.85
Learning-based	Proposed model	**3.07***	**0.91***

* Statistically significant improvement over all baseline methods (paired t-test, $p < 0.05$).

Table 2 compares the average performance of different methods on the VCTK dataset, including traditional methods and learning-based approaches, with performance evaluated using PESQ and STOI metrics. Among the traditional methods, EVS achieves the best performance with a PESQ of 1.89 and an STOI of 0.78. Among the learning-based baseline models, bin2bin demonstrates the strongest performance in both metrics. The proposed model achieves a PESQ of 3.07 and an STOI of 0.91. Compared to EVS, the proposed model improves PESQ by 1.18 and STOI by 0.13, and compared to bin2bin, it increases PESQ by 0.35 and STOI by 0.03. This indicates that the proposed model surpasses all baseline models in both key metrics, fully demonstrating its significant advantages in enhancing speech quality and intelligibility.

Figure 7 shows the spectrograms output by the bin2bin model and the proposed model. There are a large number of obvious blank stripes in the lossy speech spectrogram, which represent the packet loss caused by compression or transmission. Although the bin2bin model

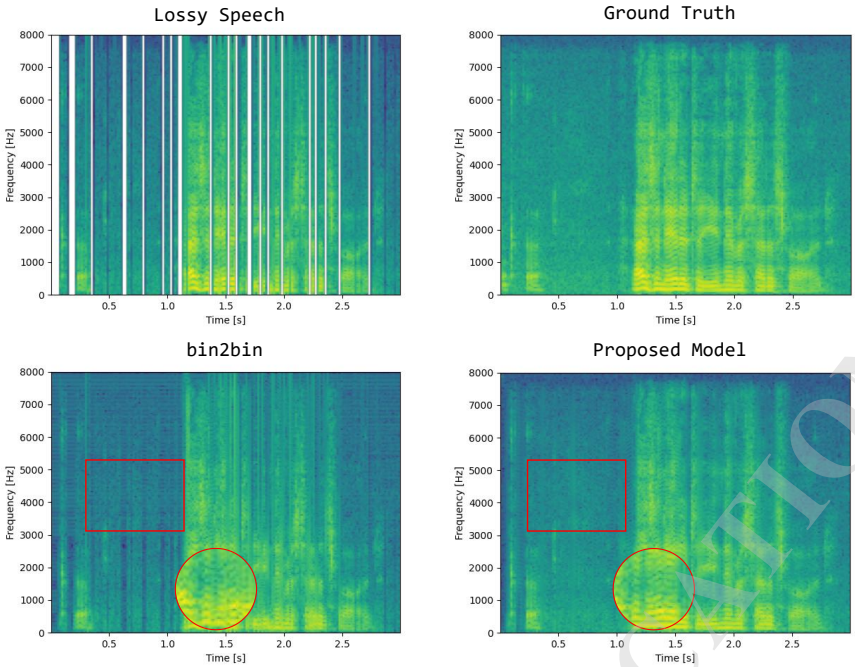


Figure 7: Visualization of the output spectrogram.

Table 3: Overall quality of the output speech under specific loss rate and burst loss.

Maximum burst loss \ Loss rate	10%	20%	30%	40%
20 ms	0.89	0.85	0.82	0.78
40 ms	0.86	0.81	0.80	0.77
60 ms	0.82	0.80	0.76	0.73
80 ms	0.79	0.75	0.73	0.70
100 ms	0.76	0.72	0.69	0.67
120 ms	0.74	0.70	0.66	0.62

improves the spectral structure of the lossy speech, there are some sparse noises in the red box that do not conform to the speech characteristics. This shows that the model has a certain effect in hiding packet loss, but fails to improve the overall quality of speech. At the same time, there is an obvious discontinuity in the red oval area below the red box, indicating that the model is still insufficient in recovering low-frequency local details. On the contrary, the speech output by the proposed model has no non-speech features in the area within the red box, where it is replaced by a more continuous and smooth speech spectrum structure. In the high-frequency area above 3000 Hz, the energy distribution of the spectrogram is more uniform, and the details are richer, indicating that the proposed model can effectively recover the high-frequency details of the speech signal. In addition, the low-frequency part of the red circular area has a smooth transition, indicating that the proposed model generates natural and highly intelligible speech features. In general, the proposed model can effectively remove periodic noise and random noise caused by the nonlinear mapping of neural networks while restoring key details of speech. This is due to the network's ability to accurately model the time-frequency structure of speech signals, which enables high-fidelity restoration of the original signal.

Table 3 presents the overall quality scores of the output speech under different packet loss rates and maximum burst loss durations. This score is the average of DNSMOS, PLCMOS, PESQ, STOI and word accuracy after they are normalized to $[0, 1]$. The test data is VCTK. This normalized average is used only as an auxiliary indicator to illustrate overall trends across different loss conditions and is not intended to replace the interpretation of individual metrics, which are reported and discussed separately in other cases. The data in the table reflects the dual impact of packet loss rate and burst loss duration on speech quality, with an overall trend showing that as the packet loss rate increases and the burst loss duration extends, the output speech quality gradually declines.

As the packet loss rate increases, the restoration quality gradually degrades across all burst loss settings, indicating that higher loss frequency poses a greater challenge even when burst duration is limited. For a fixed packet loss rate, increasing the maximum burst loss duration leads to a more pronounced degradation in quality, reflecting the difficulty of reconstructing longer consecutive missing segments. Notably, the degradation caused by burst loss duration is more severe at higher loss rates, suggesting a compounding effect when frequent and prolonged packet loss occurs simultaneously. These trends indicate that the proposed model is robust to moderate packet loss and short bursts, while its performance degrades gracefully as loss conditions become more extreme, which is consistent with practical communication scenarios.

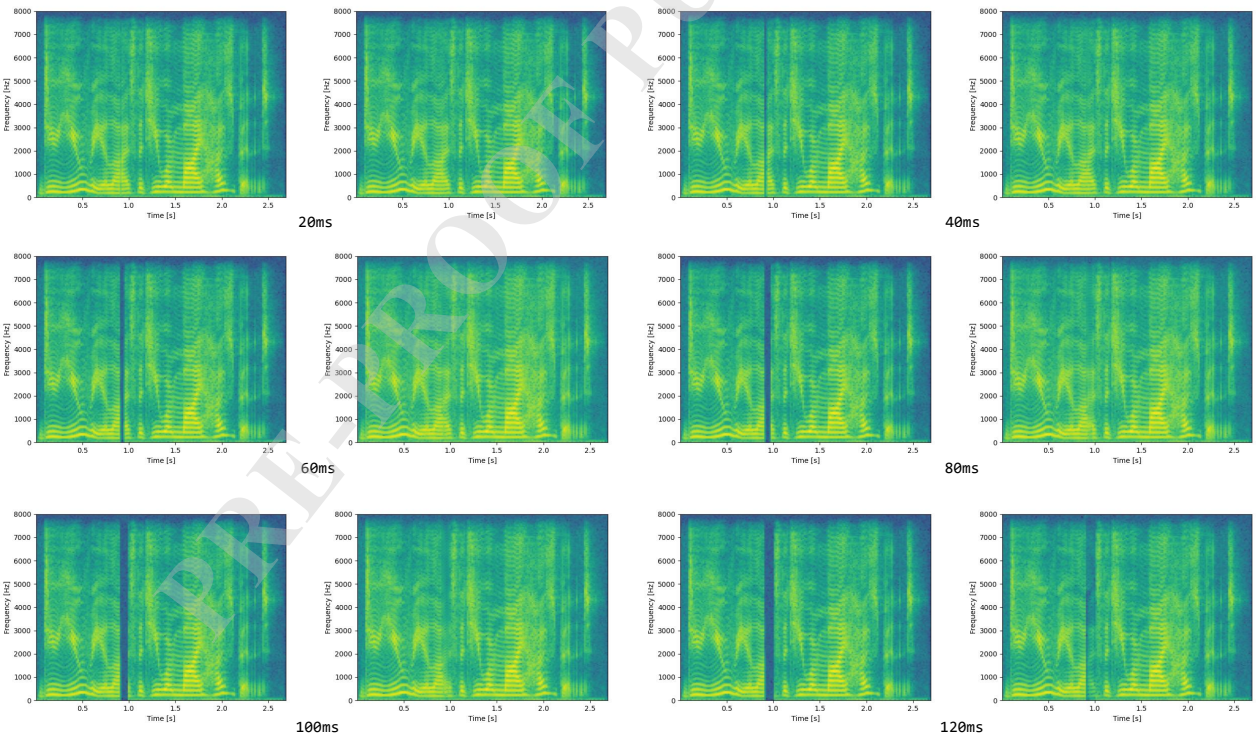


Figure 8: Visualization of the output spectrogram for different burst loss.

Figure 8 further illustrates the phenomenon of speech quality variation with burst loss duration, as shown in Table 3, through spectrogram visualization. When the burst loss duration is 20 ms and 40 ms, the continuity and details of the spectrogram remain relatively clear, with high-frequency components well preserved, which aligns with the higher quality scores

under the corresponding conditions in Table 3. At this point, the main characteristics of the speech signal are retained, and despite a certain degree of packet loss, the network's recovery mechanism still maintains good speech quality. As the burst loss duration increases to 100 ms and 120 ms, the spectrogram begins to show distinct vertical noise streaks and blurred spectral structures, particularly in the high-frequency range (above 4000 Hz). This reflects that prolonged continuous packet loss causes more severe damage to the speech signal, leading to a decline in the quality scores of the output speech in Table 3. Moreover, comparing spectrograms under different packet loss rates reveals that, under the same burst loss duration, a higher packet loss rate results in more noise and missing segments in the spectrogram, along with greater loss of spectral details. For example, at a 20 ms burst loss duration, the spectrogram continuity at a 10% packet loss rate is significantly better than at a 40% packet loss rate, which is consistent with the decreasing trend in the same row of data in Table 3 as the loss rate increases.

Figure 8 and Table 3 jointly reveal the dual-impact mechanism of burst loss duration and packet loss rate on speech quality. Short burst losses primarily affect the details of the speech signal, while prolonged burst losses disrupt the continuity and integrity of the speech signal. An increase in the packet loss rate exacerbates this damaging effect, leading to more spectral information loss. Together, these dual influences determine the performance of packet loss concealment systems under different conditions.

Table 4: Test results on our collected real-world data.

Method	PLCMOS	DNSMOS	MOS
Input	3.02	3.24	2.98
Opus	3.16	3.26	3.20
WebRTC	3.12	3.23	3.11
EVS	3.21	3.29	3.24
PLAAE	3.28	3.40	3.47
TFGAN	3.23	3.32	3.25
bin2bin	3.68	3.51	3.62
RNN	3.36	3.38	3.42
MFM	3.56	3.52	3.55
Proposed model	**3.88***	**3.67***	**3.75***

* Statistically significant improvement over all baseline methods (paired t-test, $p < 0.05$).

Table 4 reports the performance of different packet loss concealment methods on the collected real-world data, reflecting the quality of the restored speech signals. Compared with the degraded input, all PLC methods provide noticeable improvements, indicating that packet loss concealment is effective under practical communication conditions. Traditional codecs such as Opus, WebRTC, and EVS achieve modest gains, with PLCMOS values in the range of 3.12-3.21 and MOS around 3.11-3.24, which reflects the limited capability of their built-in PLC mechanisms in handling combined packet loss and noise. In contrast, learning-based approaches con-

sistently outperform codec-based methods; for example, bin2bin improves PLCMOS to 3.68 and MOS to 3.62, while MFM further increases DNSMOS to 3.52. The proposed model achieves the best overall performance. These results indicate that jointly modeling packet loss concealment and noise suppression is particularly beneficial for real-world scenarios, enabling the proposed model to deliver more natural and intelligible restored speech than both traditional codecs and existing learning-based methods.

Table 5: Results of the ablation study.

Item	PLCMOS	DNSMOS	MOS
w/o adversarial training	3.77	3.60	3.63
w/o $\mathcal{L}_{MSTFT}$	3.62	3.55	3.59
w/o $\mathcal{L}_{pre}$	3.86	3.66	3.65
w/o mamba block	3.76	3.61	3.66
w/o ELA	3.82	3.65	3.69
original setting	**3.88**	**3.67**	**3.75**

Table 5 shows the results of an ablation study designed to evaluate the impact of different components and training strategies on model performance. The "Item" column lists various experimental configurations, including the original setup and several variants where specific components or training methods are removed. The data show that when adversarial training is absent, all three metrics decrease compared to the original setup, indicating that adversarial training positively contributes to improving various aspects of speech quality. For the case without $\mathcal{L}_{MSTFT}$, the PLCMOS and DNSMOS decline, with PLCMOS dropping significantly from 3.88 in the original setup to 3.62, suggesting that $\mathcal{L}_{MSTFT}$ plays a crucial role in inpainting. In the absence of $\mathcal{L}_{pre}$, PLCMOS and DNSMOS show minimal changes, while MOS decreases from 3.75 in the original setup to 3.65. This implies that $\mathcal{L}_{pre}$ has little effect on the objective metrics of speech quality but significantly impacts the overall MOS, which may encompass broader considerations such as subjective evaluations or downstream task accuracy. In other words, $\mathcal{L}_{pre}$ may be important for certain subjective or downstream task performance but does not markedly improve objective speech quality metrics. When the Mamba block is removed, both PLCMOS and DNSMOS decrease compared to the original setup, with DNSMOS dropping from 3.67 to 3.61, demonstrating that it contributes positively to the modeling of long sequences. When it comes to ELA, PLCMOS decreases from 3.88 in the original setup to 3.82, and DNSMOS decreases from 3.67 to 3.65. This may suggest that ELA has a slightly negative effect on PLCMOS but a positive effect on DNSMOS. Overall, the original setup performs best when considering all factors. This indicates that the components and training strategies in the original setup work synergistically to achieve optimal performance across multiple dimensions.

4 Conclusion

In this work, we proposed an efficient audio inpainting model based on a GAN framework with a selective state space model and adversarial training, achieving robust packet loss concealment while maintaining light-weight. By integrating Mamba blocks for efficient long-sequence modeling and ELA attention for local feature refinement, our model outperforms existing methods in both objective metrics and subjective evaluations, particularly in noisy and bandwidth-limited scenarios. The inclusion of WavLM-based perceptual loss further enhances speech quality and intelligibility for the downstream ASR task. Experiments on real-world and benchmark datasets demonstrate our model's superior performance and generalization ability, making it a practical solution for real-time speech communication systems. Our future work will explore further optimization for edge devices and extension to broader speech enhancement applications.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

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

Authors' contribution

Junkang YANG conceptualized the study and wrote the original draft. Hongqing LIU supervised the project. Lu GAN and Liming SHI performed the analysis of experiments. Hiromitsu NISHIZAKI and Chee Siang LEOW contributed to data interpretation. All authors reviewed and approved the final manuscript.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aironi, C., S. Cornell, L. Serafini, and S. Squartini (2023). "A Time-Frequency Generative Adversarial Based Method for Audio Packet Loss Concealment". In: *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 121–125. DOI: 10.23919/EUSIPCO58844.2023.10290027.

Blum, N., S. Lachapelle, and H. Alvestrand (2021). "WebRTC: real-time communication for the open web platform". In: *Commun. ACM* 64.8, pp. 50–54. DOI: 10.1145/3453182.

Chen, S., C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei (2022). "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1505–1518. DOI: 10.1109/JSTSP.2022.3188113.

Chi, Z., S. Huang, L. Dong, S. Ma, B. Zheng, S. Singhal, P. Bajaj, X. Song, X.-L. Mao, H. Huang, and F. Wei (2022). "XLM-E: Cross-lingual Language Model Pre-training via ELECTRA". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6170–6182. DOI: 10.18653/v1/2022.acl-long.427.

Dai, T., J. Wang, H. Guo, J. Li, J. Wang, and Z. Zhu (Aug. 2024). "FreqFormer: Frequency-aware Transformer for Lightweight Image Super-resolution". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 731–739. DOI: 10.24963/ijcai.2024/81.

Davy, S., N. Belton, J. Tobin, O. B. Zuber, L. Dong, and Y. Xuewen (2023). "A causal convolutional approach for packet loss concealment in low powered devices". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096505.

Diener, L., M. Purin, S. Sootla, A. Saabas, R. Aichner, and R. Cutler (2023). "PLCMOS – A Data-driven Non-intrusive Metric for The Evaluation of Packet Loss Concealment Algorithms". In: *Interspeech 2023*, pp. 2533–2537. DOI: 10.21437/Interspeech.2023-1532.

Diener, L., S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler (2022). "INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge". In: *Proc. Interspeech 2022*, pp. 580–584. DOI: 10.21437/Interspeech.2022-10829.

Dubey, H., A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner (2023). "ICASSP 2023 Deep Noise Suppression Challenge". In: *ICASSP*.

Ebner, P. P. and A. Eltelt (2020). "Audio inpainting with generative adversarial network". In: *arXiv preprint arXiv:2003.07704*.

Fu, S.-W., C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao (2021). "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement". In: *Proc. Interspeech 2021*, pp. 201–205. DOI: 10.21437/Interspeech.2021-599.

Gu, A. and T. Dao (2023). "Mamba: Linear-time sequence modeling with selective state spaces". In: *arXiv preprint arXiv:2312.00752*.

Guan, Y., G. Yu, A. Li, C. Zheng, and J. Wang (2022). "TMGAN-PLC: Audio Packet Loss Concealment using Temporal Memory Generative Adversarial Network". In: *Proc. Interspeech 2022*, pp. 565–569. DOI: 10.21437/Interspeech.2022-644.

Hao, S., X. Li, W. Peng, Z. Fan, Z. Ji, and I. Ganchev (2024). "YOLO-CXR: A Novel Detection Network for Locating Multiple Small Lesions in Chest X-Ray Images". In: *IEEE Access* 12, pp. 156003–156019. DOI: 10.1109/ACCESS.2024.3482102.

He, K., X. Zhang, S. Ren, and J. Sun (June 2016). "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Janicki, A. and B. Ksiundefinedżak (2008). "Packet Loss Concealment Algorithm for VoIP Transmission in Unreliable Networks". In: *Proceedings of the 2008 Conference on New Trends in Multimedia and Network Information Systems*, pp. 23–33. DOI: 10.5555/1565754.1565759.

Kumar, K., R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville (2019). "Melgan: Generative adversarial networks for conditional waveform synthesis". In: *Advances in neural information processing systems* 32.

Lagrange, M., S. Marchand, and J.-B. Rault (2005). "Long interpolation of audio signals using linear prediction in sinusoidal modeling". In: *Journal of the Audio Engineering Society* 53.10, pp. 891–905.

Lecomte, J., T. Vaillancourt, S. Bruhn, H. Sung, K. Peng, K. Kikuiri, B. Wang, S. Subasingha, and J. Faure (2015). "Packet-loss concealment technology advances in EVS". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5708–5712. DOI: 10.1109/ICASSP.2015.7179065.

Lee, B.-K. and J.-H. Chang (2015). "Packet loss concealment based on deep neural networks for digital speech transmission". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.2, pp. 378–387. DOI: 10.1109/TASLP.2015.2509780.

Li, N., X. Zheng, C. Zhang, L. Guo, and B. Yu (2022). "End-to-End Multi-Loss Training for Low Delay Packet Loss Concealment". In: *Proc. Interspeech 2022*, pp. 585–589. DOI: 10.21437/Interspeech.2022-11439.

Lin, J., Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen (2021). "A time-domain convolutional recurrent network for packet loss concealment". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7148–7152. DOI: 10.1109/ICASSP39728.2021.9413595.

Liu, B., Q. Song, M. Yang, W. Yuan, and T. Wang (2022). "PLCNet: Real-time Packet Loss Concealment with Semi-supervised Generative Adversarial Network". In: *Proc. Interspeech 2022*, pp. 575–579. DOI: 10.21437/Interspeech.2022-10428.

Lotfidereshgi, R. and P. Gournay (2018a). "Speech Prediction Using an Adaptive Recurrent Neural Network with Application to Packet Loss Concealment". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5394–5398. DOI: 10.1109/ICASSP.2018.8462185.

— (2018b). "Speech prediction using an adaptive recurrent neural network with application to packet loss concealment". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5394–5398. DOI: 10.1109/ICASSP.2018.8462185.

Maher, R. C. (1994). "A method for extrapolation of missing digital audio data". In: *Journal of the Audio Engineering Society* 42.5, pp. 350–357.

Miotello, F., M. Pezzoli, L. Comanducci, F. Antonacci, and A. Sarti (2024). "Deep Prior-Based Audio Inpainting Using Multi-Resolution Harmonic Convolutional Neural Networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32, pp. 113–123. DOI: 10.1109/TASLP.2023.3324556.

Mohamed, M. M. and B. W. Schuller (2020). "Concealnet: An end-to-end neural network for packet loss concealment in deep speech emotion recognition". In: *arXiv preprint arXiv:2005.07777*.

Moliner, E., J. Lehtinen, and V. Välimäki (2023). "Solving audio inverse problems with a diffusion model". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095637.

Pascual, S., J. Serrà, and J. Pons (2021). "Adversarial Auto-Encoding for Packet Loss Concealment". In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 71–75. DOI: 10.1109/WASPAA52581.2021.9632730.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140, pp. 1–67. DOI: 10.5555/3455716.3455856.

Reddy, C. K. A., V. Gopal, and R. Cutler (2022). "Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890. DOI: 10.1109/ICASSP43922.2022.9746108.

Rix, A., J. Beerends, M. Hollier, and A. Hekstra (2001). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Vol. 2, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

Rodbro, C. A., M. N. Murthi, S. V. Andersen, and S. H. Jensen (2006). "Hidden Markov model-based packet loss concealment for voice over IP". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.5, pp. 1609–1623. DOI: 10.1109/TSA.2005.858561.

Sun, Y., H. Zhang, F. Huang, Q. Gao, P. Li, D. Li, and G. Luo (2025). "Deliod a lightweight detection model for intestinal organoids based on deep learning". In: *Scientific Reports* 15.1, p. 5040. DOI: https://doi.org/10.1038/s41598-025-89409-y.

Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen (2011). "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7, pp. 2125–2136. DOI: 10.1109/TASL.2011.2114881.

Valin, J.-M., G. Maxwell, T. B. Terriberry, and K. Vos (2016). *High-Quality, Low-Delay Music Coding in the Opus Codec*. arXiv: 1602.04845 [cs.MM]. URL: https://arxiv.org/abs/1602.04845.

Valin, J.-M., A. Mustafa, C. Montgomery, T. B. Terriberry, M. Klingbeil, P. Smaragdis, and A. Krishnaswamy (2022). "Real-Time Packet Loss Concealment With Mixed Generative and Predictive Model". In: *Proc. Interspeech 2022*, pp. 570–574. DOI: 10.21437/Interspeech.2022-903.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Wang, J., Y. Guan, C. Zheng, R. Peng, and X. Li (Oct. 2021). "A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission". In: *The Journal of the Acoustical Society of America* 150.4, pp. 2577–2588. DOI: 10.1121/10.0006528.

Westhausen, N. L. and B. T. Meyer (2022). "tPLCnet: Real-time Deep Packet Loss Concealment in the Time Domain Using a Short Temporal Context". In: *Proc. Interspeech 2022*, pp. 2903–2907. DOI: 10.21437/Interspeech.2022-10157.

Wichern, G., J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux (2019). "WHAM!: Extending Speech Separation to Noisy Environments". In: *Proc. Interspeech*.

Xu, W. and Y. Wan (2024). "ELA: Efficient local attention for deep convolutional neural networks". In: *arXiv preprint arXiv:2403.01123*.

Yamagishi, J., C. Veaux, and K. MacDonald (2019). "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)". In: URL: https://api.semanticscholar.org/CorpusID:213060286.

Yang, D.-H., D. Kim, and J.-H. Chang (2023). "Masked Frequency Modeling for Improving Packet Loss Concealment in Speech Transmission Systems". In: *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. DOI: 10.1109/WASPAA58266.2023.10248056.

Yang, J., H. Liu, L. Gan, and X. Jing (Nov. 2024). "Spectral network based on lattice convolution and adversarial training for noise-robust speech super-resolution". In: *The Journal of the Acoustical Society of America* 156.5, pp. 3143–3157. DOI: 10.1121/10.0034364.

Yen, H., F. G. Germain, G. Wichern, and J. Le Roux (2023). "Cold diffusion for speech enhancement". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096064.

Zhao, H. (2023). "A GAN Speech Inpainting Model for Audio Editing Software". In: *Proc. INTERSPEECH 2023*, pp. 5127–5131. DOI: 10.21437/Interspeech.2023-904.