Restoring Music Integrity via Predictive Modeling and Spectral Inpainting

Carlo Aironi*, Leonardo Gabrielli*, Samuele Cornell† and Stefano Squartini*
*Università Politecnica delle Marche, Italy †Carnegie Mellon University, USA
Email: c.aironi@staff.univpm.it, (l.gabrielli, s.squartini)@univpm.it, samuele.cornell@ieee.org

***Abstract*—This technical report addresses the issue of packet loss concealment (PLC) in Networked Music Performance (NMP) by presenting an improved version of our *bin2bin* [1] model, used to participate in the 2024 edition of the IS^2 Music PLC Challenge. The approach combines Linear Predictive Coding (LPC) with a *bin2bin* Generative Adversarial Network (GAN). Unlike existing methods such as PARCnet, which estimate the LPC error using an ANN, our approach uses the LPC-generated audio to condition a generative *bin2bin* convolutional model for spectrogram inpainting, trained under the GAN paradigm. The report discusses the architecture and compares it to the previous version, highlighting its improvements in restoring the original audio quality. The solution is capable of running in real-time and in a fully causal setting, in compliance with the 2025 IS^2 Music PLC Challenge requirements.**

***Index Terms*—Packet Loss Concealment, Generative Adversarial Network, Linear Predictive Coding**

I. Introduction

Networked Music Performance (NMP) depends on high-quality audio transmission with low latency to ensure a smooth and immersive user experience. Nonetheless, the inherent variability of network conditions frequently results in packet loss, which can significantly impair audio quality and disrupt the performance. Therefore, effective Packet Loss Concealment (PLC) is essential to preserve the integrity and quality of the transmitted audio in NMP scenarios.

Conventional PLC approaches, including interpolation and packet repetition, often fall short of meeting the stringent quality requirements of NMP. While more sophisticated techniques have been explored since the 1990s [2], they continue to face challenges in achieving high audio fidelity while keeping latency low.

In this technical report, we propose an improved version of a PLC architecture that was previously proposed at the 2024 IS^2 PLC Challenge [1]. The previously proposed architecture reached second place after the PARCnet baseline, motivating the authors to seek further refinements that could potentially lead to surpassing the baseline. As in the originally proposed method, we integrate Linear Predictive Coding (LPC) with a Generative Adversarial Network (GAN). The GAN stems from the *bin2bin* inpainting architecture [3], [4], but is now improved to be lighter in terms of computational cost. Further adjustments on the LPC contribute to making the method feasible in real-time.

The baseline system provided by the challenge organizers is a modified version of PARCnet [5], which also uses a linear predictor (LP) in conjunction with an artificial neural network (ANN). In PARCnet, the ANN is designed to estimate the LPC error to refine time-domain residuals. In contrast, our approach employs a *bin2bin* GAN to generate audio conditioned on the preliminary signal estimate from the LP model. This strategy combines the predictive efficiency of LPC with the generative power of GANs to reconstruct lost packets with high fidelity.

II. Method

The objective of the challenge is to conceal one or more lost audio packets of very short duration (512 samples at a 44 100 Hz sampling rate). It is assumed that the positions of the losses are provided in advance via a binary mask.

The proposed concealment method is depicted in Figure 1. When a loss is detected, the system must generate a plausible reconstruction in a fully causal manner, with no future context (i.e., no lookahead). Assuming that all preceding packets are valid (even if some have been reconstructed in prior steps), the system begins by performing an inpainting hypothesis using Linear Prediction (LP). The LP model is adapted to estimate $p = 256$ coefficients from a context consisting of the 7 most recent consecutive frames (3584 samples, or 81.3 ms). It is tasked with generating a segment twice the length of the missing portion, i.e., 1024 samples. This approach supports future sample prediction, which is beneficial for a potential crossfade step, and it also enhances the *bin2bin* network's performance by positioning the corrupted region slightly away from the context edge.

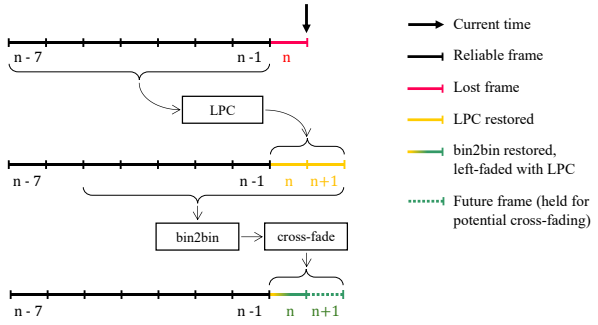

Fig. 1. Illustration of the proposed inpainting method.

A. *bin2bin-v2* Model

Our *bin2bin-v2* architecture leverages the U-Net design, incorporating skip-connections between homogeneous layers.

The U-Net consists of a convolutional encoder that downsamples the input spectrogram and a decoder that upsamples the latent representation. With respect to our previous model, the U-Net CNN employs depthwise separable convolution layers, to obtain savings in terms of memory footprint and learnable parameters.

Table I compares the original *bin2bin* U-Net employing regular 2D convolutions and the novel *bin2bin-v2* U-Net. The computational cost savings are remarkable, though the network can maintain a high complexity, which is necessary to obtain the best performance.

TABLE I
COMPARISON BETWEEN ORIGINAL (*bin2bin*) AND DEPTHWISE-SEPARABLE (*bin2bin-v2*) GENERATOR ARCHITECTURES.

Metric	Original	Depthwise-Separable
Total Parameters	54 408 385	1 152 449
Trainable Parameters	54 408 385	1 152 449
Total Mult-Adds (G)	17.84	1.10
Input Size (MB)	0.26	0.26
Forward/Backward Size (MB)	56.43	112.72
Parameters Size (MB)	217.63	4.61
Estimated Total Size (MB)	274.33	117.59

The encoder consists of five convolutional stages, with the number of kernels progressively increasing as [16, 32, 64, 64, 64]. Each stage applies a depthwise convolution with 4×4 kernels and stride 2 for downsampling, followed by a pointwise projection using 1×1 kernels, batch normalization, and a Leaky-ReLU activation. At the bottleneck, a standard 3×3 convolutional layer processes the compressed representation. The decoder mirrors the encoder but differs in a few key aspects: it uses transposed convolutions for upsampling instead of standard convolutions, and replaces Leaky-ReLU with ReLU as the activation function. The final transposed convolution restores the original spatial resolution and maps the output to the target domain using a sigmoid activation.

The generator G is fed with an STFT magnitude signal of size $1 \times 256 \times 256$, (channels, frequency bins, time frames). During inference, the magnitude STFT reconstructed by the generator is combined with the original STFT phase spectrogram (predicted by the LP) to reconstruct the audio signal in the time domain.

The GAN Discriminator is a typical classification network, starting with an initial convolutional layer followed by a series of CNN blocks. Each CNN block features a convolutional layer with reflection padding, batch normalization, and a LeakyReLU activation function. The network concludes with a fully connected layer that outputs a scalar value, representing the discriminator's assessment of the input data.

In addition to improving the CNN performance, the LPC was also improved to obtain a speedup. Specifically, `numpy` optimized operators were introduced, redundant computations were removed, in-place operations were employed and slicing was made for efficient. All these changes contributed to a 50% reduction of the LPC computation time compared to the implementation used in our submission to the 2024 challenge.

B. Loss criteria

The discriminator is trained using a least-square conditional loss function to make the training more stable and alleviate the vanishing gradient problem [6]. The objective functions for the joint conditional and least squares GAN (referred to as LSCGAN) are defined as follows:

$$\min_D \mathcal{L}_{LSCGAN}(D) = \frac{1}{2}\mathbb{E}_{x,c}\left[\left(D(x|c)-1\right)^2\right] + \\ + \frac{1}{2}\mathbb{E}_{z,c}\left[\left(D(G(z)|c)\right)^2\right] \tag{1}$$

$$\min_G \mathcal{L}_{LSCGAN}(G) = \frac{1}{2}\mathbb{E}_{z,c}\left[\left(D\left(G(z)|c\right)-1\right)^2\right] \tag{2}$$

The generator model is trained by combining the GAN objective with two conventional pixel-wise losses between the generated source spectrogram reconstruction and the expected target spectrogram. Based on our previous studies, we have found it beneficial to use loss functions that relate to the perceptual quality of the audio signal. These include the log-STFT magnitude loss ($\mathcal{L}_{mag}$) and spectral convergence loss ($\mathcal{L}_{sc}$), defined as follows:

$$\mathcal{L}_{sc}\left(S,\hat{S}\right) = \frac{\sqrt{\sum_{t,f}\left(|S_{t,f}| - |\hat{S}_{t,f}|\right)^2}}{\sqrt{\sum_{t,f}|S_{t,f}|^2}} \tag{3}$$

$$\mathcal{L}_{mag}\left(S,\hat{S}\right) = \frac{\sum_{t,f}|\log|S_{t,f}| - \log|\hat{S}_{t,f}||}{T \cdot N} \tag{4}$$

where $|S_{t,f}|$ and $|\hat{S}_{t,f}|$ represent the STFT magnitude vector of the target and the generated signal respectively, at time t, while T and N denote the number of time and frequency bins. The spectral loss is given a weight of 10, while the adversarial is given a weight of 1.

C. DNN training protocol

The training procedure involves processing chunks of 4096 samples, which equate to approximately 93 ms at the sampling rate of 44 100 Hz, and correspond to 8 quantized gaps. As the first step, the audio context, randomly extracted from the dataset, is corrupted by inserting zeros at the end for a duration of 1024 samples, then a coarse reconstruction is performed with the LP. The resulting audio segment is then transformed into the time-frequency domain and fed into the *bin2bin-v2* generator network, which refines the reconstruction acting as a conditional-GAN (cGAN) [7]. The reliable portion of spectrogram, preceding the gap, serves as the conditioning signal, along with the rough inpainted bins provided by the LP.

We followed a common practice in training generative networks, which consists in balancing the evolution of training by iterating n_G times the generator weights update, for every one of *D*. We used the value n_G = 10. The model was trained for an arbitrary number of epochs. Due to the lack of an objective metric that directly correlates with the quality of the

reconstruction, we opted to systematically evaluate the training progress by listening to the generated audio samples. Based on these evaluations, we determined our criteria for selecting the best network checkpoint. Finally, we used the Adam optimizer for both the generator and the discriminator, with a learning rate of 0.0002, progressively decreased to half, with a cosine profile, and a batch size of 16.

D. Datasets

To ensure robust network generalization during the training phase, we used an ensemble of three music signals collections. The first, as recommended by the challenge organizers, is Medley Solos DB [8], comprising nearly 18 hours of solo instrument recordings across a taxonomy of eight instruments: clarinet, distorted electric guitar, female singer, flute, piano, tenor saxophone, trumpet, and violin. Additionally, to enable the network to learn an extended frequency range for each musical instrument timbre, we included the GoodSounds collection [9], which contains monophonic recordings of both sustained notes and scales. Finally, we augmented the training set by generating 45 additional hours of synthetic audio clips from MIDI sequences. For this purpose, we selected nine representative instruments from the categories of keyboards, strings and woodwinds, using the GeneralUser GS soundfont.[1]. The MIDI sequences were sourced from the MAESTRO dataset [10] and synthesized using FluidSynth [2].

III. Experiments

According to the challenge guidelines, the test set was inpainted using the proposed technique. Processing times were measured on an Intel i7-6850K (3.6 GHz) processor from 2016 which, in benchmarks, offers comparable performance to the recommended hardware specified in the guidelines, namely, an Intel i5-10400F or equivalent. The average processing time for each frame is 8.9 ms, thereby enabling real-time processing of standard audio frames with a duration of 11.6 ms. This measurement was obtained using 4 CPU cores, as utilizing all available cores would introduce parallelization overhead that increases computation time.

The project code, along with minimal instructions for setting up experiments and reproducing results, is available on our GitHub repository[3].

We assessed the reconstruction quality of the proposed method through both subjective and objective evaluations. Subjectively, random samples from the test set were selected for informal listening tests, which confirmed a perceptual improvement of *bin2bin-v2* over the previous *bin2bin* model. Afterwards, we employed an objective evaluation metric, the PLCMOS [11], a reference-free quality estimator specifically designed for packet loss concealment tasks, with scores ranging from 1 (bad) to 5 (excellent). Our enhanced method achieved an average PLCMOS of 2.594, corresponding to a 9.68% relative increase compared to the score of 2.365 obtained by our previous model. In contrast, the lossy (unrepaired) input files scored significantly lower, with an average PLCMOS of 2.008.

As expected, the model is less effective at restoring musical excerpts characterized by percussive sounds, low harmonic content and dominant low-frequency components, where the spectrogram resolution is lower and convolutional layers struggle in generating the appropriate structures.

[1]https://schristiancollins.com/generaluser.php

[2]https://www.fluidsynth.org

[3]https://github.com/aircarlo/bin2bin_LPC

IV. Conclusions

This technical report described an approach to PLC based on a enhanced version of the previously proposed combination of LPC and *bin2bin* GAN. By combining the robustness of LPC with the generative capabilities of GANs, our proposed method will hopefully increase the current benchmark for PLC in NMP, to ensure higher quality, real-time audio transmission despite network-induced packet loss.

References

[1] C. Aironi, L. Gabrielli, S. Cornell, and S. Squartini, "Enhancing music packet loss concealment with generative spectrogram inpainting," *2024 IS2 PLC Challenge Technical Report*, 2024. [Online]. Available: http://internetofsounds.net/public_downloads/Aironi_tech_report.pdf

[2] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[3] C. Aironi, S. Cornell, L. Serafini, and S. Squartini, "A Time-Frequency Generative Adversarial Based Method for Audio Packet Loss Concealment," in *31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 121–125.

[4] C. Aironi, L. Gabrielli, S. Cornell, and S. Squartini, "Complex-bin2bin: A latency-flexible generative neural model for audio packet loss concealment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] A. I. Mezza, M. Amerena, A. Bernardini, and A. Sarti, "Hybrid packet loss concealment for real-time networked music applications," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 266–273, 2024.

[6] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "On the Effectiveness of Least Squares Generative Adversarial Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2947–2960, dec 2019.

[7] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv*, 2014.

[8] V. Lostanlen, C.-E. Cella, R. Bittner, and S. Essid, "Medley-solos-DB: a cross-collection dataset for musical instrument recognition," Sep. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3464194

[9] G. Bandiera, O. Romani Picas, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "Good-sounds.org: a framework to explore goodness in instrumental sounds," 08 2016.

[10] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[11] L. Diener, M. Purin, S. Sootla, A. Saabas, R. Aichner, and R. Cutler, "Plcmos–a data-driven non-intrusive metric for the evaluation of packet loss concealment algorithms," *arXiv preprint arXiv:2305.15127*, 2023.