Complex-bin2bin: A Latency-Flexible Generative Neural Model for Audio Packet Loss Concealment

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

***Abstract*—Despite the significant advancements in networking technologies, transmission of data packets in real-time, particularly in speech communications, continues to face challenges due to the possibility of data loss. This loss not only compromises sound quality but also diminishes overall intelligibility. In such cases, Packet Loss Concealment (PLC) techniques could help by reconstructing the missing content and restoring the audio quality. This work proposes a novel method, that improves previous time-frequency generative inpainting approaches. Compared to other state-of-the-art methods, our proposed approach has the flexibility to restore lost packets either in real-time at low latency or in offline mode, without the need to retrain the network. Evaluations conducted against a recent state-of-the-art method, ranked at the top of the 2022 Microsoft PLC competition, and against four DNN-based PLC solutions from the literature, show superior scores in terms of task-specific metrics. The method has also been tested in more challenging scenarios than aforementioned ones, with packet loss rates of up to 50%, showing the ability to help automatic speech recognition (ASR) systems reduce word error rate (WER) by up to almost 50% relative improvement. Additionally, a comparative subjective evaluation has been conducted, confirming the effectiveness of the proposed method in relation to the state of the art. The code is made available in the project repository[1].**

***Index Terms*—Packet Loss Concealment, audio inpainting, audio restoration, neural networks, generative adversarial networks**

I. Introduction

SPEECH transmission over communication channels, still poses significant challenges in terms of latency and speech intelligibility. In contrast to traditional circuit-switched networks, digital transmission introduces packet delays, losses, and jitter, potentially resulting in content loss [1]. To partially alleviate losses due to jitter, a packet buffer may be employed at the receiver side but at the cost of additional latency [2]. Achieving an optimal trade-off between latency and packet losses is a challenge. Packet Loss Concealment (PLC) algorithms can help in case of lost packets (for both real-time and offline scenarios) and packets with excessive jitter (for real-time scenarios).

Over the past three decades, various PLC algorithms have been proposed, ranging from naive methods like “zero fill” and “frame repeat” to modern speech codecs incorporating predictive coding methods such as G.722 [3], G.718 [4], adaptive multi-rate wideband speech codec (AMR-WB) [5], and Opus [6]. These algorithms reconstruct lost packets based on inter-frame correlations, significantly enhancing speech quality in Voice over Internet Protocol (VoIP). However, while introducing very small latency due to their low computational complexity, these methods struggle with long bursts of lost packets, becoming ineffective when the gap exceeds a few tens of milliseconds.

Recent breakthroughs in deep learning have demonstrated the superior speech modeling capabilities of deep neural networks (DNN), leading to their application in PLC algorithms. Despite that, further enhancements in both reconstruction efficiency and perceptual speech quality are needed.

A. Related works

Among the first statistical approaches to the PLC problem is the one presented by Rodbro et al. [7]. They propose to use a hidden Markov model (HMM) on the pitch, gain, and spectral envelope of packets, that can then be used to directly predict future frames to fill in the gaps. Bahat et al. [8] present a dictionary based scheme where the dictionary atoms are learned audio blocks. To find the best replacement for the missing block, they use both a Markov model as well as the feature space distance between the start of the block and the last known good part of the audio sequence. The dictionary is created on the fly from correctly transmitted audio. This approach offers a straightforward way of utilizing audio data to interpolate missing samples, albeit at a high computational cost and with limited adaptability beyond a single speaker.

Kegler et al. [9] and Nair et al. [10] introduce neural network methodologies adapted from computer vision, treating audio inpainting as a vision task. They employ a U-Net [11] architecture to train on masked complex short-time Fourier transform (STFT), including magnitude and phase angle, potentially with an extra channel denoting masked spectral regions. Their model is trained to predict the unmasked STFT, leveraging a perceptual loss based on a Visual Geometry Group (VGG) network. The latter uses a joint time-frequency strategy that also encompasses broader speech enhancement tasks.

In [12] Stimberg et al. present an approach based on conditioning a WaveRNN [13] over recent past context in the time domain and a convolutional conditioning network operating in the frequency domain. This network outputs samples autoregressively, one by one, instead of outputting full blocks of audio data at a time.

Lin et al. [14] introduce a convolutional-recurrent model (CRNN) designed for next frame prediction in the time domain. This model is trained to minimize mean absolute error,

[1] https://github.com/aircarlo/cplx_bin2bin

with or without employing look-ahead. Alongside conventional metrics, they evaluate also a speech recognizer word error rate, demonstrating the potential of DNN-based PLC to improve ASR's ability to recognize text from speech. Similarly, Mohamed et al. [15] propose a recurrent neural network tailored for PLC within a framework for emotion recognition. Recently, Valin et al. proposed an extension to the Opus codec's PLC functionality, called DRED [16], which leverages deep neural networks to introduce redundancy at sender side.

B. Generative Adversarial PLC approaches

Generative Adversarial Networks (GAN) [17] have been shown to be quite effective for audio waveforms generation. The first attempt to adapt the architecture of GANs to speech synthesis dates back to WaveGAN [18]. Derived approaches include cWaveGAN [19], which allows both the generator and discriminator to incorporate additional conditioning information to refine the generation process, Parallel WaveGAN [20] which integrates multiresolution short-time Fourier transform (STFT) loss alongside the adversarial loss to enhance performance and fidelity, SpecGAN [18], MelGAN [21], VocGAN [22] and StyleGAN [23].

Shi et al. [24] proposed the first PLC approach that leverages GANs. They employ a convolutional encoder-decoder network that operates on time-domain audio blocks. Their findings indicate comparable quality, as assessed by various objective metrics such as Perceptual Evaluation of Speech Quality (PESQ) [25], short-time objective intelligibility (STOI) [26], and signal-to-noise ratio (SNR), when compared to a frequency-domain deep neural network. Remarkably, their approach remains competitive even when the baseline method benefits from perfect phase information and only necessitates predicting magnitude. Pascual et al. [27] introduce a GAN-based approach wherein the generator's input is the Mel-spectrogram of the available signal, and the output is the time-domain samples of the corrupted parts. Their study demonstrates enhancements in Mel-Cepstral Distortion [28] and SESQA [29] over several baselines, including real codec systems. Additionally, Wang et al. [30] propose a GAN-based system featuring a fully time-domain U-Net style convolutional generator and a discriminator operating in a mixed time/frequency domain. This setup enables their adversarial loss to capture both intricate short-term details in the waveform and long-term relationships in the spectrum.

Finally, some researchers tackle the PLC issue from a multimodal standpoint: Zhou et al. [31] and Morrone et al. [32] introduce methods that incorporate a video feed of the speaker to aid in the recovery of missing audio segments. The former employ a convolutional neural network with adversarial loss, while the latter utilize a recurrent neural network approach.

C. Scope and organization of the work

This article proposes a time-frequency generative framework for PLC, aiming at addressing challenges posed by joint magnitude-phase recovery. Inspired by our PLC studies [33], [34] employing image inpainting techniques [35] on magnitude spectrograms, we have developed *complex-bin2bin* (also referred to as *cplx-bin2bin*), a novel method that works on complex spectrograms, that overcomes some limitations of the previous works. Its most relevant novelties, compared with current PLC approaches, are: *(a)* the use of complex-valued bins to avoid phase reconstruction artifacts and reduce computation time, *(b)* the adoption of a differentiable loss term, based on a perceptual metric for speech quality evaluation, and *(c)* the use of a smart handling of the audio data that allows a flexible trade-off between reconstruction accuracy and latency at run time, with just one trained backbone.

The paper is organized as follows: chapter II provides an overview of Generative Adversarial Networks (GAN), together with the addition of conditioning signal (cGAN) and least squares (LSGAN) loss. The inpainting process and the neural architecture are illustrated in chapter III, while chapter IV shows the experiments setup, the dataset composition, the loss criteria and illustrates the baseline comparative approaches. Chapter V reports and discusses the obtained results and finally, conclusions and future developments are drawn.

II. Least squares Conditional GAN

Since their introduction in 2014, generative adversarial networks (GANs) [17] have emerged as a powerful method in generative modeling using deep learning techniques. GANs have demonstrated their capacity to create novel, realistic and high-quality samples that closely approximate the distribution of training data.

A typical GAN comprises of two networks, namely a generator (G) and a discriminator (D). The discriminator functions as a binary classifier, while the generator operates as a deconvolutional network, converting a random seed (e.g. Gaussian noise, $z \sim \mathcal{N}(0, 1)$ into a data instance. Both G and D are trained simultaneously in a min–max competition with respect to binary cross-entropy loss (eq. 1). The ultimate goal for the generator is to produce samples that closely resemble the distribution of "real" data (x), while the discriminator aims to distinguish between "fake" and "real" samples by penalizing the generator for generating unrealistic outputs.

$$\begin{aligned}\min_{G} \max_{D} \mathcal{L}_{GAN}\left(D, G\right) = \mathbb{E}_{x}\left[\log\left(D(x)\right)\right] + \\ \mathbb{E}_{z}\left[\log\left(1 - D(G(z))\right)\right]\end{aligned} \tag{1}$$

In the original GAN framework, there lacks a mechanism to drive the specific modes of data generation. This problem originates from the need to generate a desired output class in multitarget problems. Additionally, the utilization of the binary cross-entropy loss function in the original GAN formulation may accentuate the issue of vanishing gradients, particularly when updating the generative network with gradients of samples positioned near the decision boundary [36], [37]. This scenario can hinder convergence, making the model unstable. To address these issues and strengthen the controllability and stability of GAN in generating the missing spectrogram sections, two enhancements were made.

Typically, image-to-image translation problems are often formulated as per-pixel regression, hence they treat the output on an “unstructured” space in the sense that each output pixel is considered conditionally independent from all others given the input image. The use of a conditional GAN (cGAN) [38] allows us to learn a structured loss whose objective is to penalise any possible structure that differs between the output and the target. To achieve this, the generator is fed with the spectrogram of the segment containing the damaged portion(s). This has a dual purpose, on the one hand it allows the generator to identify and time-locate reliable and damaged regions without the support of an indicator mask, and on the other hand it provides a sufficiently robust conditioning signal c to guide the generation process towards the production of segments which are pertinent to the surrounding context. In contrast to classical GANs, the authors in [39], verified that the random noise vector z does not exert considerable influence when the conditioning information is sufficiently robust, as in case of inpainting tasks. Consequently, we adopt the same choice of eliminating z, while keeping as much stochastic behavior by incorporating dropout layers into the generator.

To address the convergence issue, we used the least squares loss function instead of the sigmoid cross-entropy loss function for the discriminator. Least squares GAN (LSGAN) [36] is more stable during the learning process as it mitigates the vanishing gradient problem. The objective functions for joint conditional and least squares GAN (which will be referred as LSCGAN) can be defined as follows:

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

$$\min_{G} \mathcal{L}_{LSCGAN}\left(G\right) = \frac{1}{2}\mathbb{E}_{z,c}\left[\left(D\left(G(z)|c\right)-1\right)^{2}\right] \tag{3}$$

III. Proposed method

We first describe the latency-flexible characteristic of our proposed method, which allows to repair damaged segments in any portion of the audio input. To the best of our knowledge, in all previous works the network input and output size, and the way it is trained, constrain the operating context window and the stride step as well as the position of the lost packets in the window.

Our model performs a training procedure under more general conditions, which allows it to be flexible and operate in different inference conditions such as: *(a)* the number of lost packets in a window and their relative position; *(b)* the window length, which can be any size between 20 ms and 1024 ms. The highlighted features allow the proposed algorithm to work in any setting: from latency-critical to latency-tolerating application and even to offline ones. Indeed, by providing the algorithm with short time windows where the lost packet is the last one, packet reconstruction occurs as soon as a packet is lost, or can be regularly predicted to anticipate potential loss. Otherwise the algorithm can work with larger windows and lost packets in any position within them. In this case, the added context and the presence of reliable packets at both sides of lost packets provides useful information to improve the reconstruction quality. All these features allow to trade computational latency and reconstruction quality in real time without the need to change the backbone model at runtime. To the best of our knowledge, the method has not been previously proposed and can be applied to other PLC methods too.

The diagram in Fig. 1 illustrates the generic operation mode of the proposed framework. The algorithm is fed with an audio segment composed of two components: the buffer and the current context. The latter holds the incoming audio packets to be processed, and its length constitues a *stride*, which determines the latency of the system. The buffer context contains past audio packets. These can be either unprocessed (when correctly received) or inpainted in a previous iteration of the algorithm (when they were corrupted or missed). The whole segment size must be 51 frames long, thus the buffer length is adjusted after the size of the current context has been decided. For each audio segment, the prediction of all frames in the current context is done by using information from both the buffer context and the reliable samples in the current context itself. Once the DNN produces an inpainted version of the current context, only the actually reconstructed packets are employed in the final sequence, according to a binary mask that labels the state of each packet. To reduce reconstruction errors caused by waveform transitions at the boundaries of each reconstructed packet, we cross-fade between frames using a Hann function.

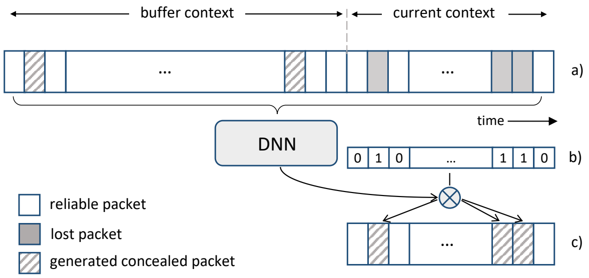


Fig. 1. Overview of the proposed adaptive latency PLC mechanism, operating on a given audio segment. The current context (a) can be varied between 1 and 51 packets during inference, by modifying the buffer length accordingly. (b) represents the binary mask denoting lost (1) and reliable (0) packets, based on which the repaired current context (c) is rearranged.

A. Network architecture

An overview of our complex-bin2bin architecture is presented in Fig. 2. In this paper we use the TCN-DenseUNet architecture in the generator. We chose this architecture given its effective use in various speech processing tasks such as speaker separation [40] and speech dereverberation [41]. TCN-DenseUNet, as the name implies, is a UNet [11] with skip-connections and DenseNet blocks [42] at multiple frequency scales in the encoder and decoder. A temporal convolutional network (TCN) [43] at the middle section leverages long-range information by using dilated convolutions along time. Exponential linear unit (ELU) activations and instance normalizations (IN) are used after convolution and deconvolution blocks. The network takes a real-valued tensor as input with shape $C \times F \times T = 2 \times 257 \times 257$, where C is the number of channels, F the number of STFT frequenciy bins and T the number of STFT frames. The real and imaginary components of the lossy spectrogram, S are concatenated along the channel axis and fed to the network, while the output $\hat{S}$ yields the same size as input. This work evolves from our previous researches on spectrogram inpainting using conditional GANs [33], [34]. Unlike the latter, the use of complex-valued spectrograms throughout the generation process allows to convert the repaired spectrogram back in time domain with a single inverse STFT operation, without the need to resort to approximate algorithms for phase estimation, which are known to be the performance bottleneck of methods operating only on magnitude spectrograms [44]. Additionally, the proposed complex bin2bin architecture is trained according to the novel procedure described in Fig. 1, using more advanced and perceptually-informed losses and different generator and discriminator models.

The discriminator model is a convolutional neural network (CNN) adopting a backbone inspired by PatchGAN [39], and a fully connected readout layer that outputs a scalar value. Since we use a Least Squares Generative Adversarial Network (LSGAN) criterion, compared to [39] we omit the final softmax activation function. The initial layer processes real-valued magnitude spectrograms, and it is fed the input (*clean* or *generated*) and the reference (*lossy*) spectrograms, concatenated along the channel dimension. This layer applies a 2D convolution with reflective padding, followed by a LeakyReLU activation to introduce non-linearity. Following, the model uses a set of three *CNNBlock* modules, each of which consists of a 2D convolutional layer, batch normalization, and a LeakyReLU activation. These blocks increase the number of feature channels, from 64 to 512, while reducing spatial dimensions via strided convolutions. Batch normalization is used to stabilize training by normalizing the feature maps at each layer. All LeakyReLU activations employ a slope of 0.2. After the final convolutional layer, a max-pooling operation is applied, and the resulting feature map is flattened before being passed through the fully connected layer, which produces the scalar output representing the discriminator's decision.

B. Loss criteria

As widely experienced in multiple research works on speech enhancement operating in the time-frequency domain, combining multiple resolution loss criteria can have beneficial effects on several aspects, even more so in the case of generative adversarial networks, where convergence and stability are critical issues [20].

To facilitate the generation of high-resolution slices of inpainted spectrograms, we used a multiresolution STFT criterion, which is based on the evaluation of the repaired and target spectrograms, at three different resolutions in time and frequency. Figure 3 shows the operative adversarial training scheme, used in generator and discriminator update.

We defined each individual STFT loss as the weighted sum of three contributions: the spectral convergence loss ($\mathcal{L}_{sc}$), the log-STFT magnitude loss ($\mathcal{L}_{mag}$) and the phase loss ($\mathcal{L}_{pha}$):

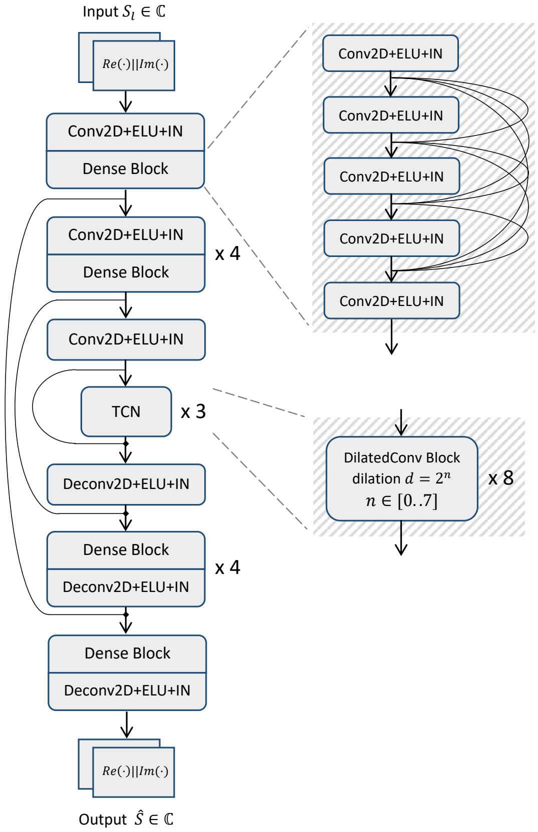


Fig. 2. Generator network composed of the U-Net with temporal convolutional (TCN) bottleneck. As input to the generator is the complex-valued lossy spectrogram S_l, which is processed as a 2-channel feature map, consisting of the real and imaginary part of S_l. The network outputs the repaired spectrogram $\hat{S}$ having the same size as the input.

$$\mathcal{L}_{STFT}(G) = \lambda_1 \cdot \mathcal{L}_{sc}\left(S, \hat{S}\right) + \lambda_2 \cdot \mathcal{L}_{mag}\left(S, \hat{S}\right) + \\ + \lambda_3 \cdot \mathcal{L}_{pha}\left(S, \hat{S}\right) \tag{4}$$

where $S \in \mathbb{C}$ and $\hat{S} \in \mathbb{C}$ denote respectively the STFT of the clean signal (s) and the repaired spectrogram. The optimal weights were chosen as $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 0.1$, during the hyperparameter tuning process. The individual loss terms are defined as follows:

$$\mathcal{L}_{sc}\left(S, \hat{S}\right) = \frac{\sqrt{\sum_{t,f}\left(|S_{t,f}| - |\hat{S}_{t,f}|\right)^2}}{\sqrt{\sum_{t,f}|S_{t,f}|^2}} \tag{5}$$

$$\mathcal{L}_{mag}\left(S, \hat{S}\right) = \frac{\sum_{t,f}|\log|S_{t,f}| - \log|\hat{S}_{t,f}||}{T \cdot N} \tag{6}$$

$$\mathcal{L}_{pha}\left(S, \hat{S}\right) = \frac{\sum_{t,f}\left(\angle S_{t,f} - \angle \hat{S}_{t,f}\right)^2}{T \cdot N} \tag{7}$$

where $|\cdot|$ and $\angle$ represent the STFT magnitude and phase components respectively, while T and N denote the number of time bins and frequency bins of a frame.

As outlined in [45], $\mathcal{L}_{sc}$ highly emphasizes large spectral components, which helps especially in early phases of training, while $\mathcal{L}_{mag}$ accurately fits small amplitude variations, which tends to be more important towards the later phases of training. Finally $\mathcal{L}_{pha}$ helps in phase estimation, although most of the insight about the structure of the speech is obtained from the magnitude [46].

It must be noted that due to the phase wrapping property, the direct Mean Squared Error (MSE) between the real and predicted phases formulated in eq. 7, may not accurately reflect the true prediction errors. An alternative solution, e.g. the one described in [47], formulates an anti-wrapping function. However, from our experiments its impact on the overall loss was negligible, because of the weighting coefficient we assigned to the phase term ($\lambda_3 = 0.1$). Therefore in the remainder of the paper we will use the MSE phase error, for simplicity and its lower computational complexity.

Computing multiresolution STFTs implies consecutive direct and inverse Fourier transformations; this approach enhances STFT consistency, which is beneficial for DNN-based spectrogram reconstruction, as pointed out in [48].

The multiple resolutions of $\mathcal{L}_{STFT}$ are given by the parameters sets reported in table I. All three terms are then averaged and summed to an additional contribution, $\mathcal{L}_{PMSQE}$ [49]. It is a perceptually-motivated speech quality loss, defined by the combination of the MSE difference in the log-power spectra domain, between S and $\hat{S}$, and two terms, $D^{(s)}$ and $D^{(a)}$, defined respectively as *symmetrical* (s) and *asymmetrical* (a) disturbance:

$$MSE = \frac{1}{F}\sum_{f}\frac{1}{\sigma^2}\left(\log\frac{|S_{t,f}|^2}{|\hat{S}_{t,f}|^2}\right)^2 \tag{8}$$

$$\mathcal{L}_{PMSQE} = \frac{1}{T}\sum_{t}\left(MSE + \alpha \cdot D^{(s)} + \beta \cdot D^{(a)}\right) \tag{9}$$

In the equations above, T is the number of frames in the training batch, F is the number of frequency bins, while α and β are weighting factors. The symmetrical and asymmetrical disturbances are computed within the loudness spectrum domain, to closely align with human auditory perception. Power spectra are first transformed into the Bark frequency scale through Q bands. Subsequently, Zwicker's law is applied to convert each Bark spectrum band into the sone loudness scale. Bands with loudness values below an absolute hearing threshold are set to zero, as these frequencies are inaudible to humans. The transformations are carried out on both the target and enhanced power spectra, S and $\hat{S}$, resulting in corresponding target and enhanced loudness spectra, $L = [l_0, ..., l_Q]^\top$ and $\hat{L} = [\hat{l}_0, ..., \hat{l}_Q]^\top$. The symmetrical disturbance term $D^{(s)}$ is first calculated, as proposed in PESQ, as the absolute difference between the loudness spectra. The asymmetrical disturbance term $D^{(a)}$ is then derived by multiplying this difference with a vector of asymmetry ratios. Although these transformations involve non-differentiable operations at specific points, subgradients are used to enable backpropagation.

TABLE I
PARAMETERS USED TO COMPUTE MULTI-RESOLUTION LOSSES. THE WINDOW SIZES AND HOP LENGTHS IN MS ARE DERIVED FROM [20] BY FITTING THE ACTUAL SAMPLING RATE.

Loss #	FFT size	Window size	Hop length
$\mathcal{L}_{STFT,1}$	1024	400 (25 ms)	80 (5 ms)
$\mathcal{L}_{STFT,2}$	2048	800 (50 ms)	160 (10 ms)
$\mathcal{L}_{STFT,3}$	512	160 (10 ms)	32 (2 ms)

IV. EXPERIMENTAL SETUP

A. Dataset and training details

The evaluation of the proposed method was carried out using two different criteria for simulating lost packets, according to the most common scenarios used by works dealing with the PLC problem.

First, a series of experiments were conducted using a synthetic dataset, generated from clean speech recordings taken from the VCTK corpus [50]. To simulate the occurrence of lost packets, fragments of 20 ms duration were filled with zeros, each selected randomly, regardless of the state of the preceding packets. As experienced in our previous work, the most effective strategy for training is to use over-corrupted recordings, i.e. with a higher loss rate than that used in the test phase. In addition, to ensure a stable and fast convergence of the generative network, each training sequences were corrupted with varying amount of losses, ranging from 10 % to 60 %.

The second operational scenario was to consider corrupted speech recordings with loss traces observed in actual VoIP calls. For this purpose, the dataset provided for the Microsoft PLC challenge 2022 [51] was used, (hereinafter referred to as MS-PLC) which consists of clean audio clips taken from radio podcasts, and separate lost packet descriptor files, that can be coupled with clean registrations to form a potentially large dataset for our purpose. The traces of lost packets from real video calls have a slightly higher variability than those obtained by randomly simulating losses. The statistical distribution of gap width for both data sets is shown in figure 4. The use of this dataset allows us to compare the proposed method with the top ranked systems of the Microsoft PLC challenge 2022 [51]. The datasets were divided into three partitions: train, validation, and test, ensuring that the test set is the same across all comparison methods to guarantee comparable results. Specifically, the VCTK dataset consists of approximately 44 hours of clean speech from 109 native english speakers, with 2 of them (one male and one female) set aside for testing, amounting to a total of 0.81 hours. The MS-PLC dataset contains about 67 hours of recordings, with the test partition representing approximately 4.5%. The operating frequency of the system was set to 16 kHz and the audio files were resampled to this value from their native frequency. Spectrograms were calculated with a window size of 512 samples and a hop size of 64. Additionally, we applied a set of data augmentation techniques, directly on the raw waveforms, with the aim to improve model performance and generalization abilities.

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2024.3515794

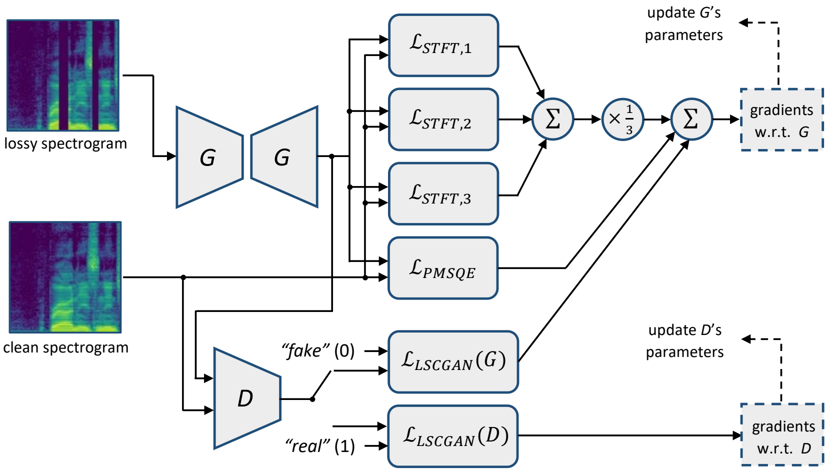


Fig. 3. Illustration of the adversarial training strategy, assisted by the losses: $\mathcal{L}_{LSCGAN}(G)$ of eq. 2, $\mathcal{L}_{LSCGAN}(D)$ of eq. 3, the multi-resolution STFT loss of eq. 4 and the perceptual loss $\mathcal{L}_{PMSQE}$ of eq. 9.

Augmentations include:

- Gaussian noise injection,
- Time stretch, with a rate in $[0.8, 1.25]$,
- Pitch shift, within -4 and $+4$ semitones.

Training is conducted with Adam optimizer, a batch size of 16 and a learning rate of $5 \cdot 10^{-3}$, progressively decreased to $1 \cdot 10^{-3}$, with a cosine profile. The latter value was used for both the generator and the discriminator. As an early stopping criterion, we tracked the progress of one of the evaluation metrics, PLCMOS [52]. This choice led to the end of training around epoch 130, as there were no further improvements beyond.

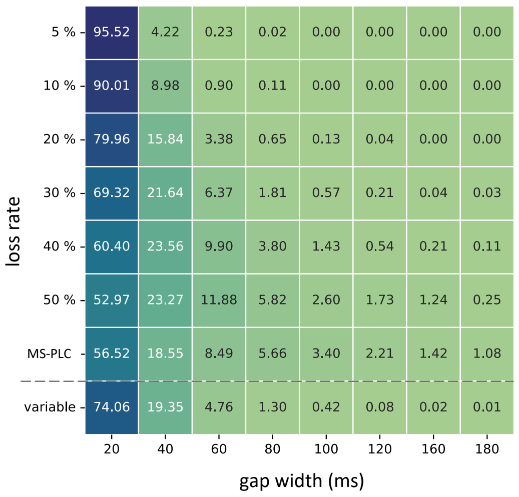


Fig. 4. Heatmap showing the distribution of gap widths characterising the datasets. The first six rows refer to the manually injected gaps, at different rates, MS-PLC refers to thet traces in Microsoft PLC challenge dataset, while *variable* indicates the distribution obtained with varying rates in $[10\,\%, 60\,\%]$. The dashed line separates test configurations (above) from train one (bottom).

B. *Evaluation metrics*

We assessed the performance of the proposed model in terms of several criteria, some of which were also used by the models taken as comparisons. In both sets of experiments, with the synthetic dataset (VCTK) and the real-traces dataset (MS-PLC), we calculated the values of PESQ [25], STOI [26], DNSMOS [53], PLCMOS [52] and Word Error Rate.

Perceptual Evaluation of Speech Quality measure (PESQ) [25] emerged as a valid objective metric on a competition to develop metrics for speech enhancement tasks. The PESQ algorithm operates by simulating human perception of speech quality and assigning scores ranging from -0.5 to 4.5.

Short-Time Objective Intelligibility (STOI) [26] operates on short-time segments of speech signals, typically utilizing a time-frequency representation such as the Short-Time Fourier Transform (STFT). It calculates a correlation-based measure, expressed as a percentage value, between the processed speech and the reference speech in each time-frequency bin, aiming to capture the perceptual intelligibility of the processed speech.

The latter metrics require an aligned reference, which limits their use to scenarios where such a reference is available. Particularly in scenarios involving PLC with a jitter buffer and timescale modification, the reference signal is typically unaligned, potentially leading to additional errors.

Non-intrusive deep neural network (DNN)-based metrics were also effectively utilized in addressing the PLC problem. One of the most prevalent metrics is the Deep Noise Suppression Mean Opinion Score (DNSMOS) [53]. Despite being originally trained for different tasks, many researchers consider it sufficiently aligned with reconstruction quality, especially in scenarios with missing segments. DNSMOS was initially conceived as a non-intrusive metric to predict scores from the ITU-T Rec. P.808 subjective evaluation, which aims to capture the overall quality of an audio clip, and was later upgraded to the P.835 standard. This standard delineates three distinct scores: speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL). Authors stated that the DNSMOS metric exhibits a high correlation with human ratings, showing a Pearson's Correlation Coefficient (PCC) of 0.94 for SIG and 0.98 for BAK and OVRL.

PLCMOS [52] is a newly implemented DNN-based metric formulated by Microsoft researchers as part of an effort to advance research on PLC. The scoring system employs a neural network trained to predict the ratings that human evaluators would assign to an audio file. Unlike the previously described DNSMOS, the PLCMOS model is trained using audio degraded by lossy transmissions, incorporating real packet loss traces observed in VoIP calls, and subsequently restored using various PLC algorithms. As a fully non-intrusive method, PLCMOS does not necessitate a reference signal. It has gained significant popularity as a means of comparing different PLC algorithms in recent times.

Word Error Rate (WER) is the primary accuracy metric used to evaluate Automatic Speech Recognition (ASR) systems, so it plays an important role in judging the correct gap reconstruction. Obviously, the impact of small and sparse gaps is significantly smaller than bursts of close gaps can have, so we expect different and non-comparable WER values between the two datasets considered. To calculate WER, we use the pre-trained ASR Whisper [54] model. Specifically, the `medium.en` model was chosen since it is adequately accurate and lightweight at the same time, to allow reasonably fast evaluations.

C. Comparative methods

The baseline systems used for evaluation of VCTK data include two strictly causal solutions, DNN and CRN, two methods designed for offline use, SEGAN and Wave-U-Net, and TFGAN-PLC. This latter allows two latency options, 20 ms and 160 ms, but still requires two separate models for each latency condition.

DNN [55] is a deep approach to predicting lost speech frames by resorting to the FFT features of previous correctly received frames. Specifically, two DNNs with three hidden layers and 2048 neurons each are employed to separately predict the magnitude and phase of the candidate frame. CRN [14] is a convolutional encoder-decoder architecture with LSTMs which has achieved excellent results in speech enhancement with magnitude-only mapping. The SEGAN-based speech enhancement approach [24] works end-to-end with the raw audio signals and reconstructs the lost frames directly in the time domain. Unlike the original SEGAN paper, a reduced configuration is used for the PLC task, with less output channels and shorter time frames. Wave-U-Net [56] is an application of the 1D convolutional U-Net architecture, originally designed to perform end-to-end speech enhancement, for the PLC task. TFGAN-PLC [30] is an end-to-end PLC approach adopting a time-frequency hybrid generative adversarial network with the integration of time-domain and frequency-domain discriminators.

Finally, the bin2bin [33] system from our previous study was included in the comparison. It operates at a fixed stride value of 1024 ms, performing the reconstruction of the magnitude spectrogram, while approximating the signal phase with the Griffin-Lim [57] algorithm, in post-processing.

Additionally, real-world traces from MS-PLC dataset were tested with LPCNet [58], an autoregressive neural vocoder that improves on WaveRNN [13] using linear prediction. It allows causal operation, reconstructing 20 ms lost packets as they occur, or by looking at 5 ms lookahead, which will be considered having 25 ms stride. A variant of LPCNet, named LPCNet-dc is also tested, in which the authors state a special handling for DC offsets.

V. Results and discussion

Comparative results are provided in the following for all the aforementioned methods, according to PESQ, STOI, DNSMOS and PLCMOS scores. Table II reports PESQ scores at different loss rates for the reference clean speech, the lossy speech (with zero-fill where packets are lost) and the neural network approaches that accept a stride of 20 ms or 160 ms. In addition we provide results for the bin2bin and the cplx-bin2bin networks, with a 1024 ms (being the only ones that can accept such a long stride) to show the benefit of the latter in operating the joint magnitute and phase reconstruction. As can be seen, the proposed method outperforms all the comparative methods in terms of PESQ. With very high loss rates (40% and 50%) there is no data for the comparative methods, but we tested our method showing that an improvement in terms of PESQ can be still achieved with respect to the zero-fill lossy

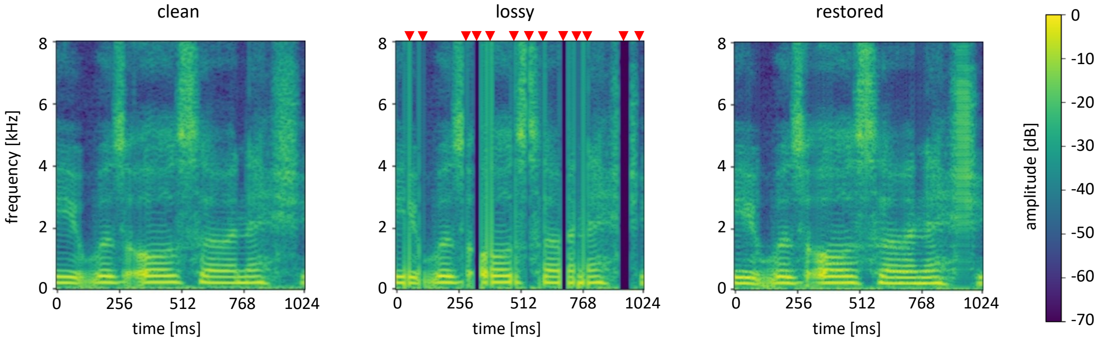


Fig. 5. Magnitude spectrograms (in dB) of an example reconstruction. Left: target signal. Center: lossy signal with red markers indicating gap displacements. Right: restored signal by using the complex-bin2bin network.

speech. Please note that according to PESQ, the clean audio gets a slightly higher result (4.64) than the ideal score (4.5), due to implementation details [59].

Evaluations are also conducted on the same data using the STOI index, as shown in table III. In this case, the TFGAN-PLC scores better with low loss rates (5 % with stride 160 ms, and 5-10 % with stride 20 ms), but the proposed method scores better with higher loss rates, making it more suitable to heavy loss scenarios. In addition to PESQ and STOI we also evaluated the PLCMOS and DNSMOS scores for the proposed method, which are shown in table VII and V. In all cases, the proposed algorithm is able to increase both metrics, over the lossy speech. Specifically, the PLCMOS score is increased from a minimum of 39 % (loss rate 50 %, stride 20 ms) to a maximum of 86 % (loss rate 20 %, stride 1024 ms).

For the sake of completeness, we also conducted tests with the cplx-bin2bin network with varying length of the stride, to assess the PESQ, STOI, PLCMOS and DNSMOS performance. These are shown in figure 7. All scores follow a similar pattern, i.e. that with longer strides the concealment performance increases. Specifically, the performance rises quickly in terms of PESQ and PLCMOS when the stride increases from 20 ms to values between 300-400 ms. Then a plateau is reached, meaning that the added context between 400 and 1000 ms is of little help to increase the reconstruction quality. It is interesting to note that with high loss rates, an increase in the stride can highly improve the STOI, the PLCMOS and the DNSMOS.

Another method to evaluate the proposed method and its ability to restore the original speech signal is to assess the WER on the reconstructed signal. Table VIII shows that with a stride of 20 ms the network can only slightly increase the performance[2]. However, with a larger stride the network benefits from the added context and is capable of reducing the WER up to a half, with loss rates as high as 50%.

TABLE II
PESQ SCORES FOR COMPLEX-BIN2BIN AND THE COMPARATIVE DNN SOLUTIONS, EVALUATED AT DIFFERENT LOSS RATES AND STRIDE.

	stride	Loss rate					
Model	(ms)	5 %	10 %	20 %	30 %	40 %	50 %
Clean	-	4.64	4.64	4.64	4.64	4.64	4.64
Lossy (zero-fill)	-	2.61	1.84	1.33	1.18	1.10	1.03
DNN	20	2.73	1.89	1.54	1.39	-	-
CRNN	20	2.79	1.93	1.66	1.48	-	-
TFGAN-PLC	20	2.94	2.16	1.87	1.63	-	-
cplx-bin2bin	20	**3.30**	**2.64**	**1.99**	**1.75**	1.51	1.47
SEGAN	160	2.76	1.95	1.63	1.49	-	-
Wave UNet	160	2.87	2.11	1.76	1.54	-	-
TFGAN-PLC	160	3.24	2.59	2.14	1.86	-	-
cplx-bin2bin	160	**3.73**	**3.23**	**2.65**	**2.26**	1.94	1.68
bin2bin	1024	3.06	2.97	2.29	2.10	1.82	1.59
cplx-bin2bin	1024	**3.76**	**3.41**	**2.89**	**2.49**	**2.15**	**1.87**

TABLE III
STOI SCORES FOR COMPLEX-BIN2BIN AND THE COMPARATIVE DNN SOLUTIONS, EVALUATED AT DIFFERENT LOSS RATES AND STRIDE.

	stride	Loss rate					
Model	(ms)	5 %	10 %	20 %	30 %	40 %	50 %
Clean	-	100	100	100	100	100	100
Lossy	-	95.45	91.20	83.94	75.75	68.87	63.99
DNN	20	95.73	92.57	85.36	78.84	-	-
CRNN	20	96.25	92.77	86.11	79.24	-	-
TFGAN-PLC	20	**97.69**	**94.68**	88.93	83.72	-	-
cplx-bin2bin	20	97.15	94.63	**89.72**	**84.59**	79.92	75.39
SEGAN	160	96.82	94.20	87.03	81.37	-	-
Wave UNet	160	97.15	94.23	87.68	82.17	-	-
TFGAN-PLC	160	**98.45**	95.82	90.11	86.39	-	-
cplx-bin2bin	160	97.74	**96.00**	**93.05**	**89.99**	86.30	81.71
bin2bin	1024	95.85	93.38	90.60	87.98	83.99	80.03
cplx-bin2bin	1024	**97.81**	**96.78**	**94.45**	**91.65**	**88.41**	**84.25**

When comparing bin2bin and cplx-bin2bin architectures, the former had worse results across all considered objective metrics, with a significant drop in DNSMOS values, and a less pronounced decline in PESQ, STOI, PLCMOS, and WER. This became clearer after conducting the subjective evaluation test. The sequences reconstructed with bin2bin maintained a good level of intelligibility, even at high loss rates (hence the high WER values), however, the lack of a cross-fade mechanism in reassembling the packet stream, and the imprecise effect of the phase approximation, resulted in noticeable distortion, which was responsible for the low DNSMOS (which evaluates noise suppression) and the low subjective MOS rates.

To visually assess the qualitative results of the proposed model, we report on Figure 5 (center) the spectrogram of a 1024 ms long segment (equivalent to nearly 50 packets) corrupted by losses of varying width. The example is taken from a real validation sample. The regions affected by losses are not sharply demarcated because the STFT operation introduces an inherent cross-fade that smooth the transition in time. On figure 5 (left) is the spectrogram of the same clean segment, while the image on the right shows the output of the reconstruction network. Although this visual evaluation does not allow for quantifying the presence of other types of distortions potentially introduced by the network, it can be observed that the differences between the reconstructed and clean spectrogram are imperceptible, and the typical formant frequencies of the speech signal are reconstructed seamlessly.

A. Subjective evaluation test

Since objective metrics may not accurately reflect the perceived quality of the concealed losses, we conducted a subjective listening experiment. The test was conducted according to the MUSHRA [60] (MUltiple Stimuli with Hidden Reference and Anchor) standard, using the webMUSHRA [61] framework. Participants were asked to rate the perceptual quality of each audio sample on a scale from 0 (bad) to 100 (excellent). The test conditions included a low quality anchor (a version of the reference with zero-filled gaps), six reconstructed versions

[2] With loss rates 5-10 % the WER of the reconstructed signal is slightly higher than the one computed on the lossy speech, which may imply that the pre-trained ASR Whisper model used in this work is somewhat already robust to drops of small audio segments.

of the anchor (using SEGAN, CRNN, WaveUNet, TFGAN-PLC, bin2bin and the proposed cplx-bin2bin), and an hidden reference (the original gap-free signal). Listeners were able, if needed, to manually replay and loop playback to focus on the gap locations for detailed evaluation. The audio items varied in packet loss rate, from 5% to 30%, and were randomly drawn from the VCTK test-set. The test consisted of 11 sessions, each containing the eight audio items to be rated. The first two sessions presented the same audio sample and were used to evaluate the reliability of each listener: we wanted to assess their ability to assign equal ratings when presented with the same stimulus. Therefore, we applied a screening criterion to the test results (in accordance with the ITU-R BS.1534 recommendation [60]), which involved excluding the two evaluators who exhibited the highest variability during such preliminary sessions. A total of 20 volunteers, all of whom reported no hearing impairment, participated in the study. The average age of the listeners was 31.5 years old. We normalized the scores for each user and session, effectively mitigating variability caused by differing loss rates as well as user judgment biases. The scores were then linearly scaled to span a range from 0 (minimum) to 100 (maximum). A subset of the samples used in the test are available on the accompanying webpage[3]. The results of the listening test are presented in Figure 6 and Table IV. As can be seen, cplx-bin2bin achieves the best score among the inpainting methods, and its median value is on par with the reference gap-free audio, showing a significant improvements from TFGAN-PLC, which is the second in the ranking. While the original bin2bin method scored quite well with respect to the objective metrics, it does not perform well in the listening test because of phase reconstruction artifacts introduced by the Griffin-Lim procedure. This issue negatively affects audio quality as it introduces a “buzzy noise” experienced by evaluators.

This finding prompts the need for more subjective tests when audio inpainting algorithms are proposed, or the need for a new objective metrics that account for perceptual factors.

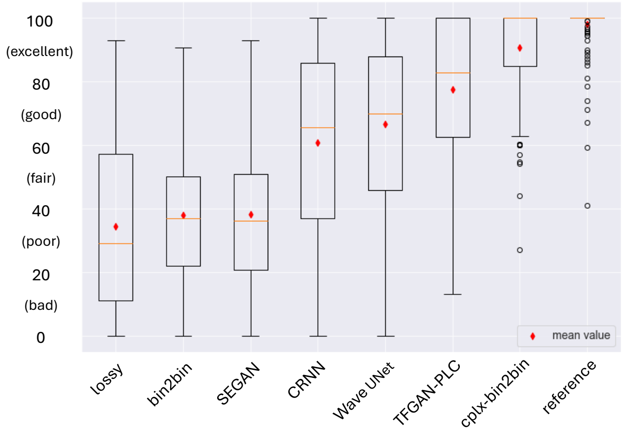


Fig. 6. Boxplot diagram representing the results of the subjective listening experiment for evaluated PLC methods.

TABLE IV
STATISTICS FROM THE SUBJECTIVE LISTENING EXPERIMENT, FOR EVALUATED PLC METHODS.

Model	stride (ms)	subjective score mean	std. dev.
Clean (reference)	-	97.87	7.49
Lossy (zero-fill)	-	34.37	27.48
CRNN	20	60.71	30.13
SEGAN	160	38.19	22.71
Wave UNet	160	66.38	24.84
TFGAN-PLC	160	77.48	22.81
cplx-bin2bin	160	**90.49**	**13.51**
bin2bin	1024	38.04	20.18

VI. Conclusions

In this work, we proposed a novel approach for Audio PLC that provides flexible handling of latency and is comparable or superior to other state of the art DNN solutions. Its flexibility lies in the ability of recovering spectrograms with lost segments in arbitrary positions, therefore, it can be employed to repair the latest audio packet, as well as any other in the input temporal context. Other PLC architectures can adopt this approach, making it valuable for future research works. Our PLC architecture is based on a generative bin2bin network that handles complex spectrograms, thus restoring the phase and magnitude information jointly. The system also employs an audio quality metric, PESQ, implemented as a differentiable DSP algorithm, in order to use it as a loss function during the training.

Experiments were conducted on different datasets. On voice recordings with randomly inserted gaps, at rates ranging from 5 % to 50 %, the proposed model outperformed five recent alternative approaches, based on neural networks, with improvements up to 22.2 % (20 ms latency) and 23.8 % (160 ms latency) for PESQ, and improvements of 1.04 % (20 ms latency) and 4.17 % (160 ms latency) for STOI. Furthermore, experiments conducted on corrupted signals with lost packet distributions taken from real-world communications networks, the complex-bin2bin model showed improvements of 14.08 % (PESQ), 1.17 % (STOI), 26.2 % (PLCMOS), 0.64 % (DNSMOS ovrl), 0.29 % (DNSMOS sig), 0.26 % (DNSMOS bak), succeeding in lowering the word error rate perceived by an automatic recognition system by 1.27 percentage points. To complete the evaluation, we conducted a subjective listening test that provided a ranking of the algorithms and clearly highlights the superior quality of our proposed cplx-bin2bin over all others.

Despite the positive results obtained by our novel PLC approach with respect to the state of art, we believe that there is still room for improvements. First, ongoing developments are aimed to incorporate an additional ASR-based loss criterion in order to enhance the overall performance, especially at low latency conditions. Moreover, future investigations will focus on alternative techniques, such as diffusion models [62] or attention-guided generative models [63], which have already shown remarkable properties in various application domains.

[3] https://aircarlo.github.io/cplx_bin2bin/

TABLE V
DNSMOS SCORES FOR COMPLEX-BIN2BIN EVALUATED AT DIFFERENT LOSS RATES AND STRIDE.

		Loss rate																	
	stride	5 %			10 %			20 %			30 %			40 %			50 %		
Model	(ms)	ovrl	sig	bak	ovrl	sig	bak	ovrl	sig	bak	ovrl	sig	bak	ovrl	sig	bak	ovrl	sig	bak
Clean		3.18	3.46	3.99	3.18	3.46	3.99	3.18	3.46	3.99	3.18	3.46	3.99	3.18	3.46	3.99	3.18	3.46	3.99
Lossy		3.12	3.40	3.97	3.00	3.28	3.92	2.64	2.91	3.75	2.19	2.43	3.55	1.79	1.98	3.34	1.51	1.64	3.20
cplx-bin2bin	20	3.14	3.45	4.00	3.06	3.38	3.96	2.90	3.23	3.89	2.71	3.05	3.81	2.49	2.83	3.69	2.24	2.58	3.56
cplx-bin2bin	160	3.17	3.47	4.01	3.12	3.43	3.99	3.04	3.35	3.96	2.95	3.27	3.93	2.86	3.19	3.90	2.76	3.08	3.86
cplx-bin2bin	1024	3.12	3.43	3.99	3.10	3.41	3.99	3.05	3.36	3.98	3.00	3.31	3.97	2.93	3.25	3.95	2.86	3.83	3.93
bin2bin	1024	3.09	3.37	3.89	2.99	3.31	3.90	2.60	2.99	3.69	2.21	2.63	3.55	1.99	2.25	3.44	1.65	1.97	3.41

TABLE VI
OVERALL METRICS FOR TESTING CPLX-BIN2BIN AND LPCNET ON MS-PLC REAL TRACES DATASET.

	stride ms	PESQ	STOI	PLCMOS	DNSMOS (ovrl)	DNSMOS (sig)	DNSMOS (bak)	WER
Clean (reference)		4.56	100.00	4.33	3.24	3.58	3.94	9.85 %
Lossy (zero-fill)		2.19	83.91	2.68	2.56	2.77	3.56	20.07 %
LPCNet causal	20	2.70	90.82	3.58	3.11	3.47	3.85	16.93 %
LPCNet-dc causal	20	2.71	90.95	3.54	3.13	3.47	3.91	17.39 %
LPCNet noncausal	25	2.76	91.36	3.62	3.10	3.45	3.87	17.23 %
LPCNet-dc noncausal	25	2.77	91.49	3.59	3.13	3.47	3.91	17.02 %
cplx-bin2bin	20	**3.16**	**92.56**	**3.95**	**3.15**	**3.48**	**3.92**	**15.66 %**

TABLE VII
PLCMOS SCORES FOR COMPLEX-BIN2BIN EVALUATED AT DIFFERENT LOSS RATES AND STRIDE.

	stride	Loss rate					
Model	(ms)	5 %	10 %	20 %	30 %	40 %	50 %
Clean	-	4.274	4.274	4.274	4.274	4.274	4.274
Lossy	-	4.025	3.586	2.721	2.179	1.822	1.547
cplx-bin2bin	20	4.179	3.982	3.567	3.163	2.853	2.613
cplx-bin2bin	160	4.229	4.131	3.939	3.728	3.477	3.168
cplx-bin2bin	1024	4.222	4.174	4.064	3.927	3.742	3.518
bin2bin	1024	4.169	4.001	3.408	3.102	2.799	2.559

TABLE VIII
WORD ERROR RATES OBTAINED ON THE SYNTHETIC DATASET, WITH DIFFERENT LOSS RATES.

	stride	Loss rate					
Model	(ms)	5 %	10 %	20 %	30 %	40 %	50 %
Clean		1.76	1.76	1.76	1.76	1.76	1.76
Lossy		1.95	2.19	2.94	3.86	6.44	12.98
cplx-bin2bin	20	2.10	2.27	2.82	3.74	5.51	12.01
cplx-bin2bin	160	1.95	2.04	2.53	3.02	3.89	6.84
cplx-bin2bin	1024	1.89	1.99	2.28	3.0	3.46	6.30
bin2bin	1024	1.99	1.94	2.30	2.98	3.77	6.91

REFERENCES

[1] D. Bhadra, P. Soni, C. Joshi, N. Vyas, and R. Jhaveri, "Packet loss probability in wireless networks: A survey," in *International Conference on Communications and Signal Processing (ICCSP)*, 04 2015, pp. 1348–1354.

[2] L. Gabrielli and S. Squartini, *Wireless Networked Music Performance*. Springer, 2016.

[3] J. Thyssen, R. Zopf, J.-H. Chen, and N. Shetty, "A Candidate for the ITU-T G.722 Packet Loss Concealment Standard," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. 549–552.

[4] M. Jelinek, T. Vaillancourt, and J. Gibbs, "G.718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 117–123, 2009.

[5] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, 2002.

[6] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-Quality, Low-Delay Music Coding in the Opus Codec," in *135th AES Convention, October 17–20 2013, New York, USA*, 2013.

[7] C. Rodbro, M. Murthi, S. Andersen, and S. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1609–1623, 2006.

[8] Y. Bahat, Y. Y. Schechner, and M. Elad, "Self-content-based audio inpainting," *Signal Processing*, vol. 111, no. C, p. 61–72, jun 2015.

[9] M. Kegler, P. Beckmann, and M. Cernak, "Deep Speech Inpainting of Time-Frequency Masks," *ArXiv*, vol. abs/1910.09058, 2019.

[10] A. A. Nair and K. Koishida, "Cascaded time + time-frequency U-Net for Speech Enhancement: jointly addressing Clipping, Codec distortions, and Gaps," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7153–7157.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

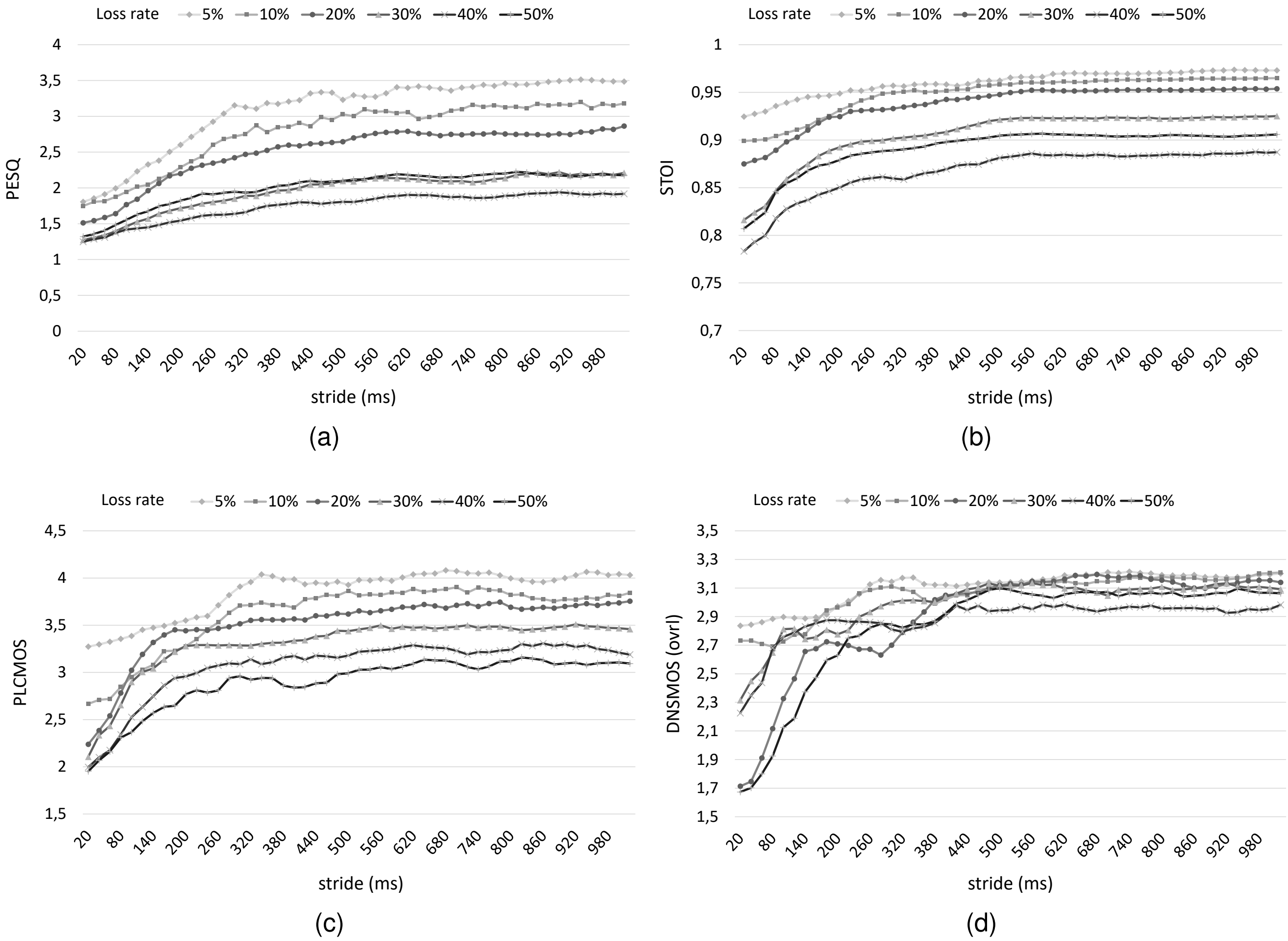
[12] F. Stimberg, A. Narest, A. Bazzica, L. Kolmodin, P. Barrera González, O. Sharonova, H. Lundin, and T. C. Walters, "WaveNetEQ — Packet Loss Concealment with WaveRNN," in *54th Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 672–676.

[13] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., vol. 80, 10-15 Jul 2018, pp. 2410–2419.

[14] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A Time-Domain Convolutional Recurrent Network for Packet Loss Concealment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7148–7152.

[15] M. M. Mohamed and B. W. Schuller, "Concealnet: An end-to-end Neural Network for Packet Loss Concealment in Deep Speech Emotion Recognition," *arXiv preprint arXiv:2005.07777*, 2020.

[16] J.-M. Valin, J. Büthe, and A. Mustafa, "Low-Bitrate Redundancy Coding of Speech Using A Rate-Distortion-Optimized Variational Autoencoder,"



(a) (b) (c) (d)

Fig. 7. Trend of PESQ (a), STOI (b), PLCMOS (c) and DNSMOS ovrl (d) values, for recovering a selected file, by varying the stride and loss rate, through the entire admissible ranges.

2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2022.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[18] C. Donahue, J. McAuley, and M. Puckette, "Adversarial Audio Synthesis," in *7th International Conference on Learning Representations, ICLR*, 2019, pp. 4974–4989.

[19] C. Y. Lee, A. Toffy, G. J. Jung, and W.-J. Han, "Conditional WaveGAN," *arXiv preprint arXiv:1809.10636*, 2018.

[20] R. Yamamoto, E. Song, and J. M. Kim, "Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[21] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[22] J. Yang, J. Lee, Y. Kim, H.-Y. Cho, and I. Kim, "VocGAN: A High-Fidelity Real-Time Vocoder with a Hierarchically-Nested Adversarial Network," in *Proc. INTERSPEECH*, 2020, pp. 200–204.

[23] K. Palkama, L. Juvela, and A. Ilin, "Conditional Spoken Digit Generation with StyleGAN," in *Proc. INTERSPEECH*, 2020, pp. 3166–3170.

[24] Y. Shi, N. Zheng, Y. Kang, and W. Rong, "Speech Loss Compensation by Generative Adversarial Networks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 347–351.

[25] I.-T. Recommendation, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4214–4217.

[27] S. Pascual, J. Serrà, and J. Pons, "Adversarial Auto-Encoding for Packet Loss Concealment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 71–75.

[28] Q. Chen, M. Tan, Y. Qi, J. Zhou, Y. Li, and Q. Wu, "V2C: Visual Voice Cloning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 210–21 219.

[29] J. Serrà, J. Pons, and S. Pascual, "SESQA: Semi-Supervised Learning for Speech Quality Assessment," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 381–385.

[30] J. Wang, Y. Guan, C. Zheng, R. Peng, and X. Li, "A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2577–2588, 2021.

[31] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-Infused Deep Audio Inpainting," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 283–292.

[32] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, "Audio-Visual Speech Inpainting with Deep Learning," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6653–6657.

[33] 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.

[34] C. Aironi, S. Cornell, L. Gabrielli, and S. Squartini, "A Score-aware

Generative Approach for Music Signals Inpainting," in *4th International Symposium on the Internet of Sounds*, 2023, pp. 1–7.

[35] X. Dong and R. Hua, "GAN Based Image Inpainting Methods: A Taxonomy," in *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, 2022, pp. 145–150.

[36] X. Mao, Q. Li, H. Xie, R. K. Lau, Z. Wang, and S. Smolley, "Least Squares Generative Adversarial Networks," in *IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2017, pp. 2813–2821.

[37] 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.

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

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[40] Y. Liu and D. Wang, "Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, p. 2092–2102, dec 2019.

[41] Z.-Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 941–950, 2020.

[42] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 2261–2269.

[43] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[44] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[45] S. Ö. Arık, H. Jun, and G. Diamos, "Fast Spectrogram Inversion using Multi-head Convolutional Neural Networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2018.

[46] J. Zhang, M. D. Plumbley, and W. Wang, "Weighted Magnitude-Phase Loss for Speech Dereverberation," in *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5794–5798.

[47] Y. Ai and Z.-H. Ling, "Low-Latency Neural Speech Phase Prediction based on Parallel Estimation Architecture and Anti-Wrapping Losses for Speech Generation Tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2283–2296, 2024.

[48] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 900–904.

[49] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[50] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[51] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "INTERSPEECH 2022 Audio Deep Packet Loss Concealment challenge," in *Proc. INTERSPEECH*, 09 2023, pp. 580–584.

[52] 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," in *Proc. INTERSPEECH*, 2023, pp. 2533–2537.

[53] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 886–890.

[54] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[55] B.-K. Lee and J.-H. Chang, "Packet Loss Concealment Based on Deep Neural Networks for Digital Speech Transmission," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 378–387, 2016.

[56] M. N. Ali, A. Brutti, and D. Falavigna, "Speech Enhancement Using Dilated Wave-U-Net: an Experimental Analysis," in *27th Conference of Open Innovations Association (FRUCT)*, 2020, pp. 3–9.

[57] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[58] J. Valin, A. Mustafa, C. Montgomery, T. Terriberry, M. Klingbeil, P. Smaragdis, and A. Krishnaswamy, "Real-Time Packet Loss Concealment With Mixed Generative and Predictive Model," in *Proc. INTERSPEECH*, 2022, pp. 570–574.

[59] R. G. D. Miao Wang, Christoph Boeddeker and A. Seelan, "PESQ wrapper for Python users," may 2022, https://doi.org/10.5281/zenodo.6549559.

[60] I.-R. Recommendation, "Method for the subjective assessment of intermediate quality level of audio systems," *Rec. ITU-R BS.1534*, 2015.

[61] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, Feb 2018.

[62] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional Diffusion Probabilistic Model for Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7402–7406.

[63] H. Tang, H. Liu, D. Xu, P. H. S. Torr, and N. Sebe, "AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 1972–1987, 2023.