A Time-Frequency Generative Adversarial based method for Audio Packet Loss Concealment

Carlo Aironi, Samuele Cornell, Luca Serafini, Stefano Squartini
Department of Information Engineering, Università Politecnica delle Marche, Italy
Email: (c.aironi, s.cornell)@pm.univpm.it, l.serafini@staff.univpm.it, s.squartini@univpm.it

Abstract—Packet loss is a major cause of voice quality degradation in VoIP transmissions with serious impact on intelligibility and user experience. This paper describes a system based on a generative adversarial approach, which aims to repair the lost fragments during the transmission of audio streams. Inspired by the powerful image-to-image translation capability of Generative Adversarial Networks (GANs), we propose *bin2bin*, an improved pix2pix framework to achieve the translation task from magnitude spectrograms of audio frames with lost packets, to non-corrupted speech spectrograms. In order to better maintain the structural information after spectrogram translation, this paper introduces the combination of two STFT-based loss functions, mixed with the traditional GAN objective. Furthermore, we employ a modified PatchGAN structure as discriminator and we lower the concealment time by a proper initialization of the phase reconstruction algorithm. Experimental results show that the proposed method has obvious advantages when compared with the current state-of-the-art methods, as it can better handle both high packet loss rates and large gaps. We make our code publicly available at: `github.com/aircarlo/bin2bin-GAN-PLC`.

Index Terms—Packet Loss Concealment, Spectrogram Inpainting, Conditional Generative Adversarial Networks, *bin2bin*.

I. INTRODUCTION

Speech signals are often subject to localized distortions or even total loss of information, when data is transmitted through unreliable channels. This happens, for example, in applications such as mobile digital communications, video-conferencing systems and Voice over Internet Protocol (VoIP) calls. In such scenarios, audio frames are often encapsulated into packets, which are then routed individually through the network, sometimes taking different paths, resulting in out-of-order delivery. At the destination, the original sequence may be reassembled in the correct order, based on the packet sequence numbers. Hence, a variety of issues can occur, like packet losses, over-delay or jitter.

The process of restoration of missing packets is known as Packet Loss Concealment (PLC) [1]. This term refers to any technique that attempts to overcome the packet-loss problem, by concealing the lost fragments by an estimated reconstruction, which should be meaningful and consistent with the informative content of the speech message. The system should also prevent audible artifacts and decrease listening fatigue, so that the listener remains unaware of any problems that have occurred.

A. Related works

Some techniques refer to a similar task with the terms Audio Inpainting [2], [3], Waveform Interpolation [4] or Extrapolation [5]. These techniques address the reconstruction problem from a sparsity point of view, by approximating the waveform with a combination of frequency atoms, extracted from a given dictionary. However they are not suitable for real-time applications, as the computational cost can lead to excessive latency times.

Most of the current approaches to PLC are based on codecs that implement algorithmic solutions: sender-based techniques like Interleaving and Forward-Error Correction (FEC) [6], or receiver-based concealment techniques, like Silence/Noise Substitution, Waveform Substitution, or Linear Predictive Coding (LPC) [7].

With the rise of Deep Neural Networks (DNN), a significant improvement of quality has been obtained on speech processing tasks, hence also DNN architectures for neural PLC have been successfully investigated: MLP [8], LSTM/RNN [1], [9], Autoencoders [10], [11], GANs [12], [13].

In this study, we apply a Generative method, based on the pix2pix [14] framework which exploits a Fully Convolutional Network (FCN) architecture, to address the spectrogram inpainting task. We show that this solution, while preserving global temporal and spectral information along with local information, can outperform competing approaches, based either on classical digital signal processing solutions or learning methods.

II. GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) [15] have emerged in the past years as a powerful generative modeling technique. A typical GAN consists of two networks, a generator (G) and a discriminator (D). Given an input of random values sampled from a normal distribution, z (latent variable), the generator performs an upsampling in order to obtain a sample of suitable dimensions. On the other hand, the discriminator acts as a binary classifier, trying to distinguish "real" samples x (belonging to the dataset distribution) from "fake" samples, generated by G.

Both G and D are trained simultaneously in a min–max competition with respect to binary cross-entropy loss. The final objective for G is to output samples that follow as close as possible the "real" data distribution, while D learns to spot the fake samples from real ones, by penalizing G for producing implausible results.

Given the success achieved in the field of image processing, GANs have also been effective in speech processing tasks. In

this regard, WaveGAN [16] represents the pioneering attempt to adapt a deep convolutional GAN (DCGAN) structure for speech, by compressing the two-dimensional image input into one-dimensional. It laid the foundations for GAN-based practical audio synthesis and for converting different image generation GANs to operate on waveforms.

Several extensions have been derived from WaveGAN; to name a few, cWaveGAN [17], which allows conditioning both G and D with additional information to drive the generation process, and Parallel WaveGAN [18], which uses a multi-resolution STFT loss along with the adversarial loss.

As outlined in [16], in the generative setting, working with compressed time-frequency representations may be problematic as the generated spectrograms are non-invertible, hence they cannot be listened to without lossy estimations, nevertheless, the practice of bootstrapping image recognition algorithms for audio tasks has become commonplace; examples include SpecGAN [16], MelGAN [19], VocGAN [20] and StyleGAN [21].

A. Pix2pix

Pix2pix is a conditional GAN (cGAN) originally developed in 2017 by Phillip Isola, et al. [14] for synthesizing photos from label maps, reconstructing objects from edge maps and colorizing images. Unlike a vanilla GAN which uses only random noise seeds to trigger generation, a cGAN introduces a sort of supervision by feeding the generator with the target information c, categorical labels or contextual samples. The discriminator is also conditioned by c, to help distinguish more accurately the matching and alignment of two images:

$$\min_G \max_D \mathcal{L}_{cGAN}(D, G) = \mathbb{E}_{x,c}\left[\log\left(D(x|c)\right)\right] + \\ \mathbb{E}_{z,c}\left[\log\left(1 - D(G(z|c))\right)\right] \quad (1)$$

Unlike other cGAN-based works (e.g. [22] [23]), Isola et al. demonstrate that the input noise vector z does not have a significant impact if the conditioning information is strong enough, so they removed it, getting the same stochastic behavior by adding dropout layers to the generator.

III. Neural Concealment Architecture

An overview of our bin2bin architecture is presented in Fig. 1. The main contribution of this paper is the adaptation of the pix2pix architecture, for the audio packet loss concealment task, through an in-depth evaluation of both generative and discriminative processes, optimized to inpaint spectrograms gaps. We adopt the term bin2bin as a direct translation of pix2pix, inspired by the fundamental unit (bin) of the discretized time and frequency axes of the spectrogram.

A. Generator

In the proposed bin2bin scheme, the generator architecture makes use of the U-Net [24] structural design with the insertion of skip-connections between affine layers. The U-Net is composed of a convolutional encoder that down-samples the input image in the first half of the architecture, and a decoder that upsamples the latent representation applying 2D transposed-convolutions.

The clean signal s and its lossy counterpart $\tilde{s}$, are first transformed into time-frequency spectrograms. In the provided implementation, all STFTs are computed with a 512 points Hann window, corresponding to 32 milliseconds at the sample rate of 16000 Hz, and a hop size of 64. The STFT parameters have been chosen to ensure a balanced resolution between the regions to be reconstructed and the reliable parts acting as conditioning contexts.

Our generator G accepts $1 \times 256 \times 256$ inputs, where each dimension represents, respectively, the number of *Channels*, *Frequency* and *Time* bins, hence, a portion of such size is extracted at a random time, from the aforementioned spectrograms S and $\tilde{S}$, regardless of the amount of lost fragments present inside.

Only the log-magnitude spectrogram is fed into the generator; for the training stage, the phase information is discarded, while for the test stage it is used to initialize the Griffin-Lim [25] phase reconstruction algorithm.

B. Discriminator

The discriminator is built on a custom architecture, specifically designed for the pix2pix framework, called PatchGAN [14]. It is basically a fully convolutional network that maps the input image into an $N \times N$ feature map of outputs Y, in which each patch y_{ij} indicates whether the corresponding portion of input is real or fake. The patches originate from overlapped receptive fields, which can be retrieved through simple backtracking operations.

In the original paper [14], an ablation study was conducted to determine the best configuration of D (number of conv. layers, kernels size), to maximize the evaluated metrics. In this work we focused on a similar aspect: we tested the effect of varying the size of the discriminator convolutional kernels, to achieve a rectangular receptive field, instead of the square dimension (70×70 pixels) used in pix2pix. We motivated this decision by observing that the portions of the spectrogram to be concealed extend over the entire frequency dimension, and a relatively small part of the time dimension. We traded-off between the complexity of D and the desired shape, obtaining an optimal receptive field of 162×24, with rectangular 8×2 kernels for all conv layers.

C. Post-processing

The generator output represents the magnitudes of the TF coefficients, both of the reliable and lost regions. The synthesis by the inverse STFT introduces an inherent cross-fading, which significantly reduces artifacts. For the phase reconstruction we used a modified version of Griffin-Lim [25] algorithm, by providing the phase of the lossy frame as an initial estimate. In this way the synthesis of the reconstructed waveform is considerably sped up; we can obtain maximum quality, with less than 10 iterations of the algorithm.
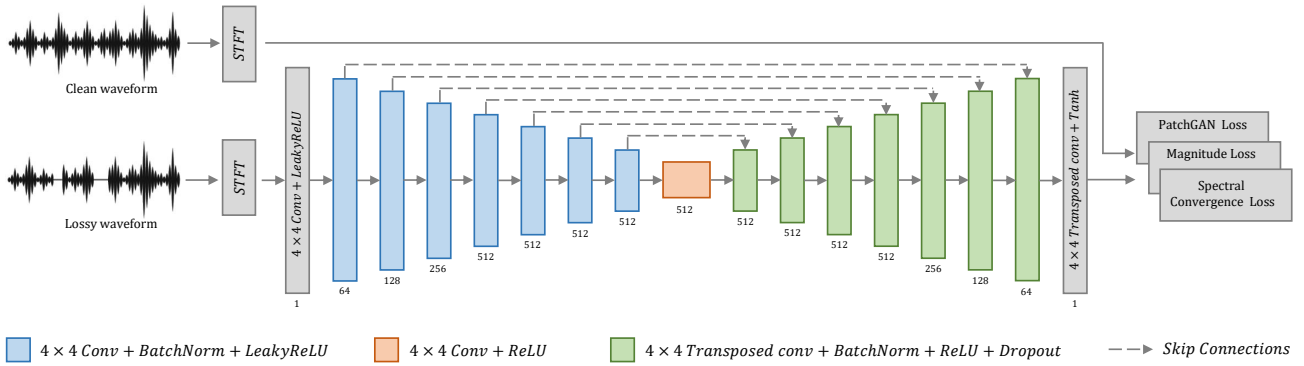
Fig. 1. The proposed framework is composed of the U-Net for spectrogram inpainting. Deep feature loss for training the U-Net is obtained by ensembling the discriminator loss (binary cross-entropy between patches), along with the spectral distances ($\mathcal{L}_{mag}$ and $\mathcal{L}_{sc}$), between the representations of the recovered and the actual STFT log-magnitudes.

D. Loss functions

The generator model is trained by mixing the GAN objective with a traditional pixel-wise loss, between the generated reconstruction of the source spectrogram and the expected target spectrogram.

Differently from the original paper, we have found it more beneficial to use loss functions related to the perceptual quality of the audio signal: log-STFT magnitude loss ($\mathcal{L}_{mag}$) and Spectral Convergence loss ($\mathcal{L}_{sc}$), defined as follows:

$$\mathcal{L}_{mag}\left(S, \tilde{S}\right) = \frac{\sum_{t,f}|\log|S_{t,f}| - \log|\tilde{S}_{t,f}||}{T \cdot N} \quad (2)$$

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

where $|S_{t,f}|$ and $|\tilde{S}_{t,f}|$ represent the STFT magnitude vector of s and $\tilde{s}$ respectively, at time t, while T and N denote the number of time bins and frequency bins of a frame.

As outlined in [26], $\mathcal{L}_{sc}$ 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.

The goal of the adversarial loss is to drive the generator model to output T-F representations that are plausible in the target domain, whereas the spectral losses regularize the generator model to output spectrograms that are a plausible translation of the source context. The combination of the adversarial loss and the spectral losses is controlled by the hyperparameters λ_1 and λ_2, both set to 250, since it has been observed that the spectral loss is more important for reconstruction than the adversarial one.

$$\mathcal{L} = \mathcal{L}_{cGAN} + \lambda_1\mathcal{L}_{mag} + \lambda_2\mathcal{L}_{sc} \quad (4)$$

The discriminator model is trained in a standalone manner in the same way as in a traditional GAN model, minimizing the negative log-likelihood of identifying real and fake images,

although conditioned on the clean spectrogram, which is concatenated with $G(\tilde{S})$ to form the input of D.

We followed a common practice in training generative networks [27], 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 models were trained for 50 epochs, following an early stopping policy based on the spectral losses observed on the validation set. We used the Adam [28] optimizer with a learning rate of 0.0002 for both the generator and the discriminator, and a batch size of 8.

IV. DATASETS

We used the VCTK Corpus (Centre for Speech Technology Voice Cloning Toolkit) [29] set of data to simulate loss traces, for training and evaluation of the speech PLC model.

VCTK contains about 44 hours of clean speech from 109 English speakers, 47 males and 62 females, with different accents. To comply with the policy followed by the comparing methods, we downsampled the audio to 16 kHz, trimmed leading and trailing silence, and split into three subsets: train, validation and test, the latter containing 5 speakers held out from the train and validation sets. We assumed that the lost packets have a duration multiple of 20 ms, and were simulated by zeroing samples of the clean waveform, finally we limited to 120 ms the maximum gap length, equivalent to 6 consecutive packets. Fig. 3 shows the distribution of lost gaps, obtained by injecting packets with rates in the range 10% - 40%.

V. RESULTS AND COMPARISONS

The proposed PLC method has been compared with three algorithmic solutions, represented by the general purpose codecs Opus [30], WebRTC [31] and Enhanced Voice Services (EVS) [32], and against four state-of-the-art deep PLC methods: the wave-to-wave generative adversarial network (PLCNet) [33], the mel-to-wave non-autoregressive adversarial auto-encoder (PLAAE) [34], the wave-to-wave adaptive recurrent neural network (RNN) [9] and the time-frequency hybrid generative
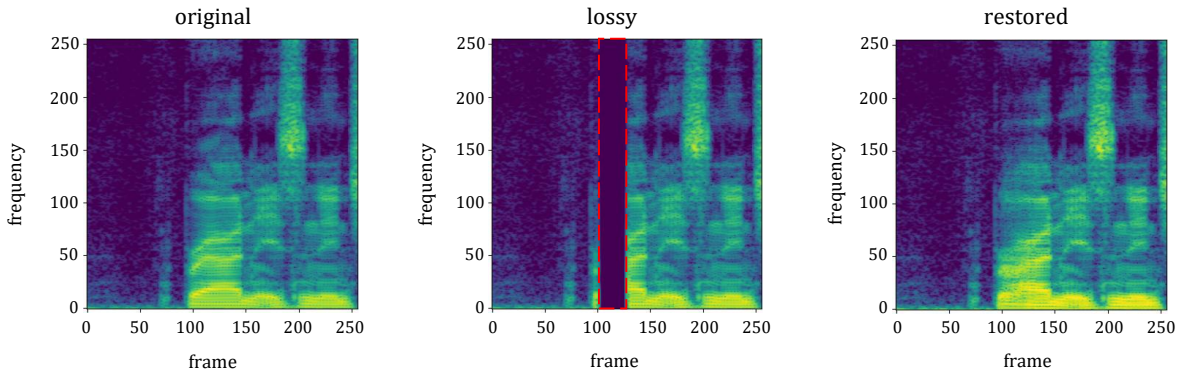
Fig. 2. Magnitude spectrograms (in dB) of an example reconstruction. Left: original signal. Center: lossy signal with 120 ms wide gap (in red-dashed box) Right: reconstruction by the bin2bin network. The axes of the plots indicate the frequency bin and the frame index.
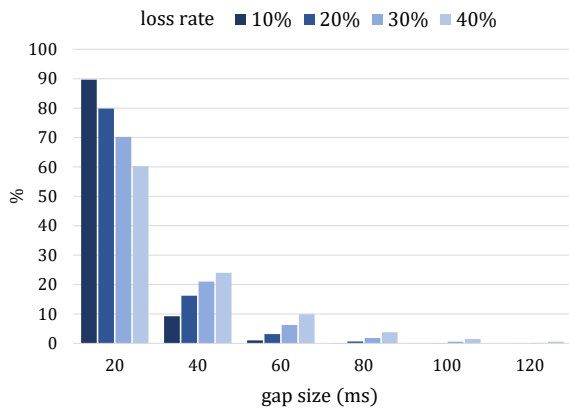

Fig. 3. Gap size distribution obtained by injecting zero-valued frames, with different rates.

adversarial network (TFGAN) [12]. In addition, the evaluation metrics obtained by simply zero-filling the lost gaps were also reported as a baseline.

We evaluated the performances of the proposed generative inpainting method, in terms of Wide-Band Perceptual Evaluation of Speech Quality (PESQ) [35] and Short-Time Objective Intelligibility (STOI) [36]. The implementations used in this paper are from [37] for PESQ, and from [38] for STOI.

Table I shows the experimental results for PESQ and STOI, under different packet loss rates, compared with the PLCNet method, It can be seen that the proposed model can achieve a significant improvement in performance, the more the loss rate increases, so it is also able to cope better with large gaps of adjacent lost packets. The improvement is notable on PESQ scores; it ranges from +6.0% (loss rate 10%) to +27.5% (loss rate 40%). The STOI shows less noticeable gains, only for higher loss rates: +2.3% (loss rate 30%) and +7.8% (loss rate 40%).

Table II summarizes the results of the proposed method with all the competing approaches. Values represent the average score of PESQ and STOI under all packet loss rates investigated. Compared with the best performing network among

previous state-of-the-art systems (PLCNet), bin2bin improves PESQ by 15.3% and STOI by 2.4%, while, in comparison with the best codec-based concealment (EVS), the improvement rises up to 43.9% for PESQ and 12.8% for STOI.

Figure 2 shows the qualitative results of a concealed 120 ms wide gap, within a test sample. This represents the worst scenario, in terms of extent of lost fragments, the network is trained to face.

In addition, we timed the forward execution of the bin2bin inpainting process, both in a CPU environment (Intel core i7-6850K) and a GPU environment (Nvidia Titan Xp), obtaining real-time (RT) factor values of 0.17 and 0.11 respectively.

TABLE I
OBJECTIVE SCORES FOR BIN2BIN AND PLCNET, UNDER DIFFERENT PACKET LOSS RATES

	Packet Loss Rate	zero-fill	PLCNet	bin2bin
PESQ	10 %	2.13	3.12	**3.31**
	20 %	1.04	2.60	**2.87**
	30 %	0.89	2.04	**2.50**
	40 %	0.81	1.71	**2.18**
STOI	10 %	0.86	**0.93**	0.92
	20 %	0.81	**0.90**	**0.90**
	30 %	0.73	0.85	**0.87**
	40 %	0.61	0.77	**0.83**

TABLE II
AVERAGE OBJECTIVE SCORES FOR THE COMPARISON PLC SOLUTIONS, UNDER PACKET LOSS RATE 10%-40%

	PESQ	STOI
bin2bin	**2.72**	**0.88**
PLCNet	2.36	0.86
PLAAE	2.04	0.84
TF-GAN	1.97	0.81
RNN	1.91	0.77
EVS	1.89	0.78
Opus	1.77	0.77
WebRTC	1.70	0.70
zero-fill	1.22	0.75

VI. Conclusions

In this paper, we proposed an end-to-end pipeline for spectrogram inpainting and audio concealment using a cGAN-based architecture, inspired by the popular pix2pix framework. We combined the classical discriminative loss with a linear combination of two loss functions, that are correlated with the perceptual quality of speech. In addition, we adapted the receptive field of the PatchGAN discriminator and we used a custom initialization of the Griffin-Lim algorithm to speed up post-processing. We demonstrated experimentally that the proposed method is capable of simultaneously identifying and recovering missing parts, thus outperforming the state-of-the-art DNN method by +15.3% on PESQ and +2.4% on STOI, respectively. Finally, inference time evaluation suggests that this approach can be integrated into a real-time application, even with a mid-range hardware setting.

As future developments we plan to investigate the generator to directly process complex-valued spectrograms, in order to incorporate the phase reconstruction directly into the generative model.

References

[1] 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.

[2] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2011.

[3] O. Mokrý and P. Rajmic, "Audio inpainting: Revisited and reweighted," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2906–2918, 2020.

[4] M. Lagrange, S. Marchand, and J.-B. Rault, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 891–905, 2005.

[5] R. C. Maher, "A method for extrapolation of missing digital audio data," *Journal of the Audio Engineering Society*, vol. 42, no. 5, pp. 350–357, 1994.

[6] T.-K. Chua and D. C. Pheanis, "Qos evaluation of sender-based loss-recovery techniques for voip," *IEEE Network*, vol. 20, no. 6, pp. 14–22, 2006.

[7] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.

[8] 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, 2015.

[9] R. Lotfidereshgi and P. Gournay, "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)*. IEEE, 2018, pp. 5394–5398.

[10] Y.-L. Chang, K.-Y. Lee, P.-Y. Wu, H.-y. Lee, and W. Hsu, "Deep long audio inpainting," *arXiv preprint arXiv:1911.06476*, 2019.

[11] M. Kegler, P. Beckmann, and M. Cernak, "Deep Speech Inpainting of Time-Frequency Masks," in *Proc. Interspeech 2020*, 2020, pp. 3276–3280.

[12] 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.

[13] P. P. Ebner and A. Eltelt, "Audio inpainting with generative adversarial network," *arXiv preprint arXiv:2003.07704*, 2020.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[16] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

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

[18] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[19] 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.

[20] 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*, 2020, pp. 200–204.

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

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[23] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *International Conference on Learning Representations*, 2018.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[25] 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.

[26] 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.

[27] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

[30] 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.

[31] N. Blum, S. Lachapelle, and H. Alvestrand, "Webrtc: Real-time communication for the open web platform," *Commun. ACM*, p. 50–54, jul 2021.

[32] J. Lecomte, T. Vaillancourt, S. Bruhn, H. Sung, K. Peng, K. Kikuiri, B. Wang, S. Subasingha, and J. Faure, "Packet-loss concealment technology advances in evs," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5708–5712.

[33] B. Liu, Q. Song, M. Yang, W. Yuan, and T. Wang, "Plcnet: Real-time packet loss concealment with semi-supervised generative adversarial network," *Proc. Interspeech 2022*, pp. 575–579, 2022.

[34] S. Pascual, J. Serrà, and J. Pons, "Adversarial auto-encoding for packet loss concealment," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 71–75.

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

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[37] M. Wang, C. Boeddeker, R. G. Dantas, and A. Seelan, "Pesq (perceptual evaluation of speech quality) wrapper for python users," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6549559

[38] M. Pariente, "Pystoi, python implementation of stoi," 2018. [Online]. Available: https://github.com/mpariente/pystoi