Concealing Audio Packet Loss Using Frequency-consistent Generative Adversarial Networks

Linlin Ou
[1] *Computer Network Information Center*
[1] *Chinese Acadamy of Sciences*
2 School of computer science and technology
2 University of Chinese Acadamy of Science
Beijing, China
oulinlin@cnic.cn

Yuanping Chen
Computer Network Information Center
Chinese Acadamy of Sciences
Beijing, China
ypchen@cashq.ac.cn

***Abstract*—Packet loss is one of the top reasons for speech quality degradation in Voice over IP calls. Packet loss concealment(PLC) is a technique of facing packet loss. This article describes a system based on a deep neural network(DNN) and pre-trained frequency-consistent generative adversarial network(Fre-GAN), which aims to heal audio impairments caused by packet losses, thus impairing the quality of the audio playout. We use the sliding window method to simulate real-time audio processing and extract features within the window, a multi-layer fully connected network is applied to predict the mel-spectrogram of missing audio packets based on the context within the window, and a frequency-consistent generative adversarial network is used to convert the mel-spectrogram to a waveform and backfill it to the part of the dropped frame.We name this system PLCfre-GAN. In the 2022 INTERSPEECH Audio Deep PLC Challenge, we applied PLCfre-GAN to mitigate artificially simulated audio impairments. The processed audio we submitted surpasses the challenge zero-padding baseline and ranks the top 5 with a PLC-MOS score on the blind test dataset of 3.478. Compared with previous machine learning methods, the proposed system has shown considerable improvement in the scores of several evaluation metrics, and the recognition accuracy of synthesized speech is also guaranteed even in the case of frequent loss.**

Keywords—Generative Adversarial Network, Packet Loss Concealment, Audio Synthesis

I. INTRODUCTION

Packet loss concealment (PLC) is any technique that attemptsto handle the effects of packet loss or overly delayed packets.Traditional PLC techniques fill the gaps with zeros to substitute lost fragments, repeating the segment before a lost frame or us ing some form of interpolation between them. Previous statisti cal modeling methods rely on hidden Markov [1] and Gaussian mixture [2] models to predict prospective acoustic parameters.

Recently, deep neural networks were introduced to solve the task of PLC. A deep neural network(DNN) [3] was ad dressed as a non-linear regression function for PLC, in which the model was trained by inputting log-power spectral features, directly acting on time-domain samples. [4] proposed a recur rent neural network(RNN) based speech signal predictor. More recent research results based on generative RNNs, which heals damaged a frame using the preceding frames [5].

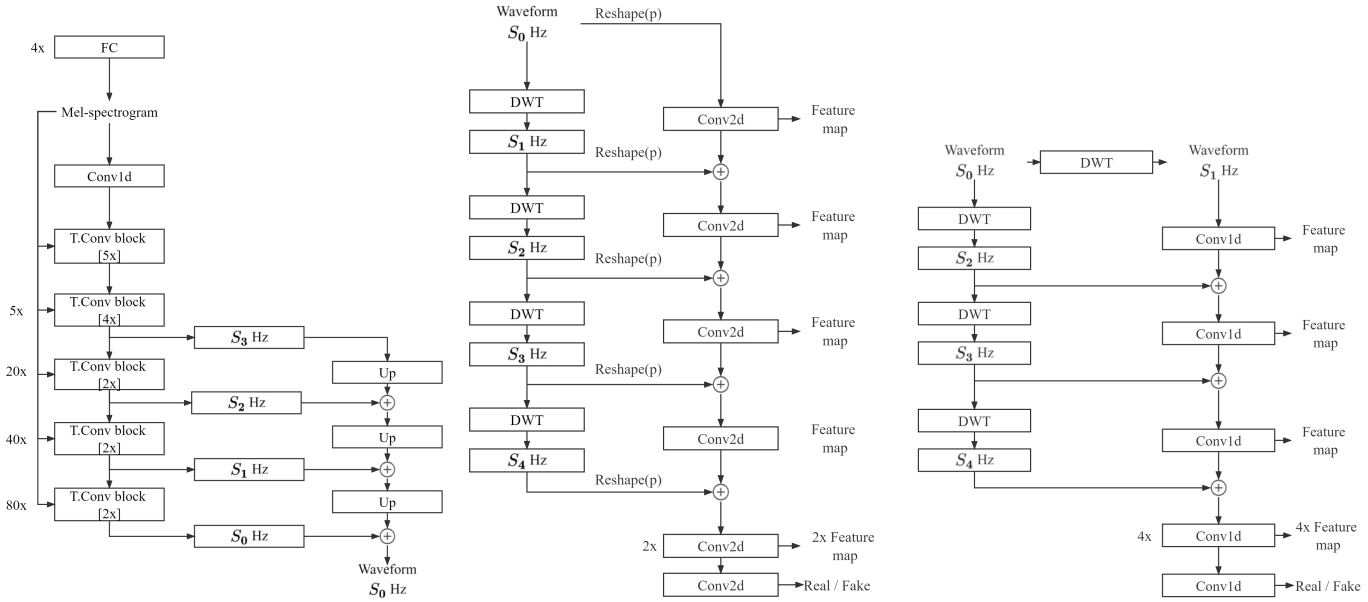
PLC algorithms can be divided into offline and online meth ods. Offline methods handle more chunks of audio including lost packets, and can take advantage of the context in the left and right directions, but they trade latency for premium audio quality. In contrast, online methods must predict missing frames in real-time and can only use left-hand context. The following research like [3] [4] [5] may be classified into this category. Several auto-encoder-based methods were discussed in recent [6] [7]. and frameworks based on generative adversarial net work(GAN) [8] [9] [10] were constructed for PLC. PLC algorithms also are typed into spectral-domain and spectral-domain methods. The performance of time-domain approaches maybe limited because they are challenging to extract commonalities in audio signals. Spectral-domain encoder-decoder models [11] [12] overcome this disadvantage, capturing the dominant frequencies and energy in audio, so they are more suitable for Large-scale speech training. This paper proposes PLCfre-GAN, an online deep neural network framework that synthesizes frequency-consistent audio on par with ground-truth audio in our entry into the Audio Deep PLC Challenge. We show that the GAN boosts the performance of the speech synthesis, surpasses the challenge baseline by a wide margin.

II. RELATED WORK

In the process of generating mel-spectrograms from the wave form, the existing most GAN models [13] [14] [15] [16] will not subdivide the mel-spectrograms containing most of the frequencies or directly use hard rule sampling to generate the wave form, resulting in a gap between synthesis and ground-truth audio in frequency space. This gap leads to spectral artifacts, like mechanical noise because a complicated mixture of various frequencies mixed audio.

MelGAN [13] adopts Multi-Scale Discriminator (MSD) operating on multiple scales of waveforms modulated by Average Pooling (AP). The MSD has been proven to be advantageous for capturing consecutive patterns and long-term dependencies of audio. Recently, HiFi-GAN [16] identifies

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE



(a) Resolution-connected Generator (RCG) (b) Resolution-wise Multi-Period Discriminator (RPD) (c) Resolution-wiseMulti-Scale Discriminator (RSD)

Fig. 1. Architecture of PLCfre-GAN. (a) RCG with DNN. (b) The p^{th} sub-discriminator RPD. (c) A sub-discriminator RSD.

the periodic patterns of audio through Multi-Period Discriminator (MPD) and synthesizes high-fidelity audio. It further improves the audio quality by applying a Multi-Receptive field Fusion (MRF) module in the generator which observes patterns of diverse lengths in parallel. The model outperforms the autoregressive and flow-based vocoder in terms of both quality and inference speed.

Fre-GAN [17] adopts resolution-connected generator and resolution-wise discriminators to learn the spectral distribution of multiple frequency bands at different levels, the architecture uses discrete wavelet transform (DWT) [18] as downsampling method. While traditional downsampling methods such as average pooling and mean pooling may lose high-frequency components, DWT guarantees that all information can be preserved because of its bi-orthogonal .nature. So in the paper, our PLCfre-GAN mainly utilizes Fre-GAN [17] as the primary tool for missing speech synthesis.

III. Approaches

Our approach employs deep learning neural network. The audio packet loss concealment task is disassembled into two primary parts: predicting the corresponding acoustic features of the lost packet according to the lost audio packet context pattern and generating sound waveforms from the predicted features.

In the first part of the task, we use a multi-layer fully connected neural network to input the mel-spectrograms converted from all audio waveforms within the sliding window in chronological order mel-spectrogram of the last moment to predict the counterpart of the next moment.

In the second part of the task, adjusted Fre-GAN [17] employs a resolution-connected generator and resolution-wise discriminators to learn various levels of spectral distributions over multiple frequency bands. Fre-GAN generator takes a mel-spectrogram as input and upsamples it through transposed convolution blocks until the temporal resolution of the output sequence matches that of the raw waveform.

A. Model Structure

1) Deep neural network

In order to establish the mapping relationship between the missing frame and the reference frame to predict the content of the missing frame, we add a 4-layer DNN module before the Resolution-connected Generator (RCG). The input vector size of the input layer is the number mel bands of several frames in the sliding window, and the output vector size of the out put layer is the number of mel bands of one frame. We have overloaded the method of loading data in PyTorch DataLoader so that it can immediately extract the corresponding mel-spectrogram every time the wav is loaded. Although dealing with a different problem, we still consider it part of the generator due to of the connections on the data stream.

2) Resolution-connected Generator

We refer to the generator design of Fre-GAN [17] and implement it ourselves in PyTorch. The predicted value of the mel-spectrogram of the missing frame as input, and the trans posed convolution blocks with different kernel function sizes are used to upsample it at different frequencies to ensure that the frequency of the output waveform is consistent with the actual waveform frequency. RCG upsamples and sums the top-K waveform outputs corresponding to different resolutions, in this paper we set K to 4.

The transposed convolution block contains the transposed convolution layer and MRF module with the same structure as HiFiGAN [16], we did not change the parameter values

according to the original paper. But we made two renovations: First of all, we adopted Parametric Rectified Linear Unit (PReLU) [19] as the activation function, which can not only improve the fitting ability of the model, but also reduce the risk of overfitting. Secondly, because of the change of sampling parameters (hop size becomes 160), we adjusted the sampling multiple of each stage of upsampling, and the sampling multiples of 5 stages are [5, 4, 2, 2, 2].

3) Resolution-connected Generator

There are two discriminators in PLCfre-GAN, Resolution-wise MPD (RPD) and Resolution-wise MSD (RSD).

In previous MSD works, AP have been used to downsample raw audio [13] [20] [16] [21], but the principle of AP does not comply with Nyquist sampling law, and often discard high frequency contents This disadvantage of the AP leads to a decrease in the quality of the synthesized audio. According to [17], the introduction of DWT alleviates this problem. Both discriminators use DWT technology to ensure that high frequency waveforms are also included in the range of the discriminator.

RPD is a mixture of sub-discriminators, each sub discriminator only accepts equally spaced samples of the input audio, and the spacing is set to p, which is the period. The sub discriminators aim to capture implicit structures that differ from each other by observing at different parts of the input audio. We set the period to [2,3,5,7,11] to try to avoid overlapping. The in put one-dimensional sound length as Z, convert the audio data from a 1D matrix to a 2D one with a length of Z/p and a width of p, and perform a 2D convolution operation on it to simulate downsampling with different periods.

The RSD consists of three sub-discriminators, which are used to discriminate the original audio, 2 × downsampled audio, and 4 × downsampled audio at different scales. RSD can complement RPD by only accepting disjoint samples, evaluating audio sequences continuously.

B. Training Objectives

To train In order to train a composite neural network composed of DNN and GAN, we use the least-squares objective for training stability [22]. The training objectives for the DNN (1), discriminators (2) and generators (3) are defined as below.

$$L_{\mathrm{DNN}} = \mathrm{E}[\left\|\psi(y) - \psi(\hat{y})\right\|_2] \tag{1}$$

$$\begin{aligned} L_{\mathrm{D}} = &\sum_{\mathrm{n}=0}^{4} \mathrm{E}[\left\|D_n^P(x) - 1\right\|_2 + \left\|D_n^P(\hat{x})\right\|_2] \\ &+ \sum_{\mathrm{m}=0}^{2} \mathrm{E}[\left\|D_m^S(\phi^m(x) - 1)\right\|_2 + \left\|D_m^S(\phi^m(\hat{x}))\right\|_2] \end{aligned} \tag{2}$$

where y and $\hat{y}$ denote ground-truth audio and predicted audio, x and $\hat{x}$ represent predicted audio and generated audio, D^p and D^S are RPD and RSD, ϕ^{m} means m-level DWT. Based on experiments, we set $\lambda_{\mathrm{fm}} = 2$ and $\lambda_{\mathrm{mel}} = 45$ which balance the adversarial losses. The feature matching loss λ_{fm} (4) and the mel-spectrogram loss λ_{mel} (5) defined as below:

$$\begin{aligned} L_{\mathrm{G}} = &\sum_{\mathrm{n}=0}^{4} \mathrm{E}[\left\|D_n^P(\hat{x}) - 1\right\|_2 + \lambda_{fm} L_{fm}(G; D_n^P)] \\ &+ \sum_{m=0}^{2} \mathrm{E}[\left\|D_m^S(\phi(\hat{x}) - 1)\right\|_2 + \lambda_{fm} L_{fm}(G; D_m^S)] \\ &+ \lambda_{mel} L_{mel}(G) + L_{DNN} \end{aligned} \tag{3}$$

$$L_{\mathrm{fm}}(\mathrm{G}; D_k) = \mathrm{E}[\sum_{i=0}^{T-1} \frac{1}{N_i} \left\|D_K^{(i)}(x) - D_K^{(i)}(\hat{x})\right\|_1] \tag{4}$$

$$L_{\mathrm{mel}}(G) = \mathrm{E}[\left\|\psi(x) - \psi(\hat{x})\right\|_1] \tag{5}$$

T represents the number of layers in the discriminator, and ψ is the short time Fourier transform function to transform predicted audio into the corresponding mel-spectrogram. $D_k^{(i)}$ denotes the i^{th} layer feature map of the k^{th} sub-discriminator, and N_i is the number of units in each layer.

According to [23] and successful application to neural vocoder [13] [16], The feature matching loss minimizes L1 distance between the discriminator feature maps of ground truth and generated audio, so RCG adopted it as an ancillary loss to improve training efficiency. Additionally, it add loss of mel-spectrogram to further improve sample quality and training stability, which minimizes L1 distance between the mel-spectrogram of real audio and that of synthesized audio, helping to generate realistic results and stabilize the adversarial training process from the early stages [14] [24].

C. Masked Training

During inference, the left context referenced by the frame to be predicted may also be lost. This phenomenon often occurs when continuous packet loss occurs, resulting in a growing discrepancy between the generated audio and the actual audio. To solve this problem, we randomly replace some frames in the left context with predicted frames during training, and the probability of random replacement is set to 20% based on experience.

IV. Experimental Setup

A. Datasets

The dataset provided for this challenge consists of a set of audio and metadata files divided into three parts: training split, vali dation split and blind test split [25]. A single audio sample approximately 10 seconds in length. We no need to identify dropped frames because the metadata files directly record the missing framesactual audio. To solve this problem, we randomly replace some frames in the left context with predicted frames during training, and the probability of random replacement is set to 20% based on experience.

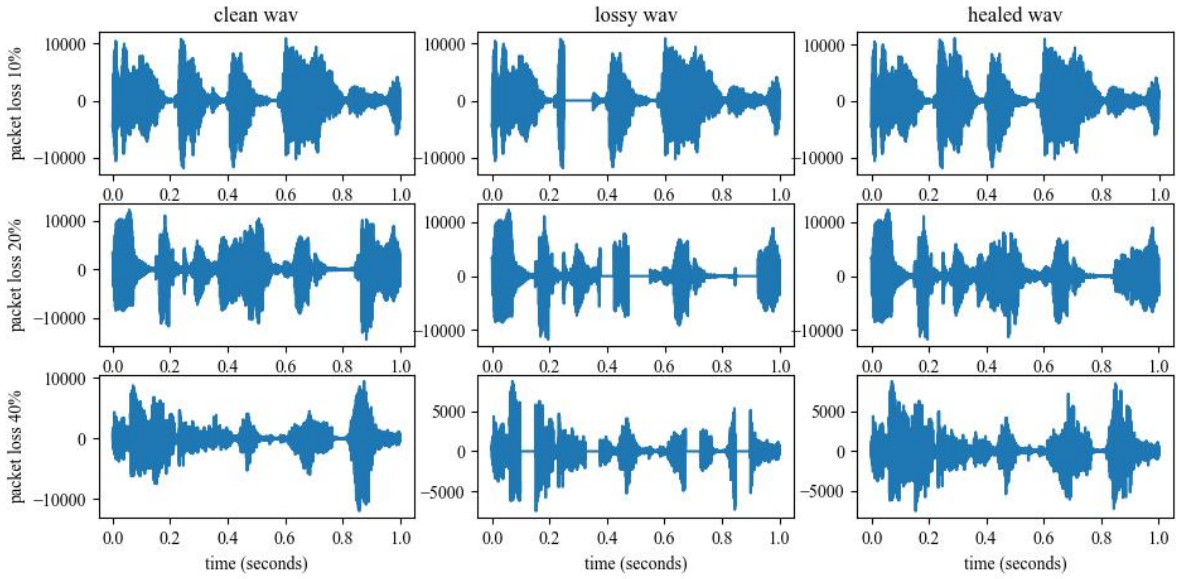


Fig. 2. Comparison of PLCfre-GAN speech compensation results. Each column of subgraphs represents clean, lossy and PLCfre-GAN healed waveforms from left to right and each row of subgraphs represents wavforms with different packet loss levels from top to bottom.

B. Evaluation Metrics

Speech enhancement is universally evaluate with the evaluated with the perceptual evaluation of speech quality (PESQ) score [26] and the short time objective intelligibility (STOI) score [27]. Both scores can compare the enhanced speech signal with a clean reference signal, and act as objective metrics of the speech quality and intelligibility by human perception.

Virtual speech quality objective Listener (ViSQOL) [28] is introduced as a subjective indicator, replacing the traditional MOS indicator. The metric has been particularly designed to be robust for quality issues associated with VoIP transmission. ViSQOL is used for the validation and blind test dataset of the competition dataset.

To calculate the word error rate (WER) [29] for repaired speech we leveraged the open source automatic speech recognition tool VOSK [30] (using the vosk-model-en-us-0.22-lgraph model) to extract their texts and adopted the LJSpeech dataset with text scripts to aid testing.

V. RESULTS

A. Performance

We ran and timed the packet loss compensation task of the blind test dataset (966 wavs) several times under the officially required hardware environment. The average time to process the entire blind test set is 1743 minutes; The average processing time is 1.8 minutes. Consider the win size, hop size and duration of wav, the approximate processing time of each frame is 12 ms, which meets the requirement of the competition that the algorithm latency should less than 20 ms.

Figure 2 visually displays the waveforms near the regener ated area of clean audio, lossy audio and healed audio with different packet loss rates. PLCfre-GAN effectively restores the original audio, and the envelope of the restored audio is roughly the same as the original audio.

B. Objective and subjective evaluation

Table 1 shows the wide-band PESQ score, STOI score, WERs and ViSQOL score of the various systems. We stats these metrics on the validation set before the release of the blind test set. According to the packet loss rate, we divide the test data into 3 categories: less than 10%, 10%-20% and

TABLE I. EVALUATION INDEX VALUES OF PLC UNDER DIFFERENT PACKET LOSS RATES

Packet Loss Rate		**Objective**			**Subjective**
		WB-PESQ	***STOI***	***WER (%)***	***ViSQOL***
PLC fre-GAN	0-10%	0-10%	0-10%	0-10%	0-10%
	3.51	3.51	3.51	3.51	3.51
	0.9754	0.9754	0.9754	0.9754	0.9754
	9.85	9.85	9.85	9.85	9.85
Fre-GAN	0-10%	0-10%	0-10%	0-10%	0-10%
	3.42	3.42	3.42	3.42	3.42
	0.9689	0.9689	0.9689	0.9689	0.9689
	9.31	9.31	9.31	9.31	9.31

TABLE II. 2022 INTERSPEECH AUDIO DEEP PLC CHALLENGE RESULTS

Rank	Team	PLCMOS	DNSMOS	CMOS	WAcc	Final Score
	Clean (reference)	4.51	3.89	0	0.97	0.98
1	Kuaishou	4.28	3.8	-0.55	0.88	0.85
2	Amazon	3.74	3.79	-0.64	0.88	0.84
3 (tied)	Alibaba Inc.	3.83	3.68	-0.81	0.87	0.8
3 (tied)	Oldenburg University	3.98	3.69	-0.84	0.86	0.79
5	SRIB	3.28	3.51	-1.1	0.86	0.75
6	UCAS	3.48	3.74	-1.04	0.83	0.74
	Zero-filling baseline	2.9	3.44	-1.23	0.86	0.73
7	Shenzhen University	2.9	3.48	-1.31	0.86	0.71

greater than 20%, their proportions are 47%, 20.9% and 31.7%, respectively.

We combine the original Fre-GAN and DNN modules as baseline, which without using the PReLU activation function and the number of DWTs in the discriminator is one less than the current scheme. The results show that the model we designed has a significant advance in the effect of speech generation.

C. *Official Results*

The final score consists of two equally weighted components: the crowd sourced mean opinion score and word accuracy. The mean opinion score was obtained by asking raters to rate samples using a crowd sourcing framework [31], with 5 ratings for each clip in the blind test dataset, and word accuracy was calculated using Microsoft Azure Cognitive Services speech recognition.

Judging from the official results shown in Table 2, our system is much better than the baseline in terms of PLCMOS [32] and DNSMOS [33], but other indicators are close to or even inferior to it. This reflects a problem that the audio we generate basically meets the PLC task requirements in terms of frequency, pitch, etc., but there is still a lot of room for improvement in the simulation of noise, echo, and human voice color and intonation in actual scenarios.

VI. CONCLUSION

This paper introduces PLCfre-GAN, a neural network system that handles audio packet loss in real time for the task of 2022 INTERSPEECH Audio Deep PLC Challenge. We learn about the frequency space inconsistency between generated and real audio, and solve this problem with the cited and upgraded net work architectures and lossless downsampling methods.

For future work, there are many interesting directions, in cluding how to denoise synthetic audio and further adapt to the human auditory system Feature extraction schemes such as early use of CNN and RNN [5] [34], [35] and other methods to predict the content of missing frames such as working on an earlier level appear promising. Also, ensemble-learning could be beneficial given the small size of the training data.

REFERENCES

[1] C. A. Rodbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen,"Hidden markov model-based packet loss concealment for voiceover ip," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1609–1623, 2006.

[2] J. Lindblom, J. Samuelsson, and P. Hedelin, "Model based spectrum prediction," in 2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat.No. 00EX421). IEEE, 2000, pp. 117–119.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting, " in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 283–292.

[9] Y. Shi, N. Zheng, Y. Kang, and W. Rong, "Speech loss compensation by generative adversarial networks, " in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019, pp. 347–351.

[10] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "Gacela: A generative adversarial context encoder for long audio inpainting of

music,″ IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 1, pp. 120–131, 2020.

[11] S. K. Pedram, S. Vaseghi, and B. Langari, ″ Audio packet loss concealment using spectral motion, ″ in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 6707–6710.

[12] D. Le, X. Zhang, W. Zheng, C. F¨ugen, G. Zweig, and M. L. Seltzer, ″ From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition, ″ in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 457–464.

[13] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Br´ebisson, Y. Bengio, and A. C. Courville, ″Melgan: Generative adversarial networks for conditional waveform synthesis,″ Advances in neural information processing systems, vol. 32, 2019.

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

[15] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee,″Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis,″ arXiv preprint arXiv:2012.07267, 2020.

[16] J. Kong, J. Kim, and J. Bae, ″Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,″ Advances in Neural Information Processing Systems, vol. 33, pp. 17 022–17 033, 2020.

[17] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, ″Fre-gan: Adversarial frequency-consistent audio synthesis,″ arXiv preprint arXiv:2106.02297, 2021.

[18] M. J. Shensa et al., ″The discrete wavelet transform: wedding the a trous and mallat algorithms,″ IEEE Transactions on signal processing, vol. 40, no. 10, pp. 2464–2482, 1992.

[19] K. He, X. Zhang, S. Ren, and J. Sun, ″ Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, ″ in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[20] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, ″Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network, ″ arXiv preprint arXiv:2007.15256, 2020.

[21] W. Jang, D. Lim, and J. Yoon, ″ Universal melgan: A robust neu ral vocoder for high-fidelity waveform generation in multiple do mains, ″ arXiv preprint arXiv:2011.09631, 2020.

[22] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley,″Least squares generative adversarial networks, ″ in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.

[23] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, ″ Autoencoding beyond pixels using a learned similarity metric, ″ in International conference on machine learning. PMLR, 2016, pp. 1558–1566.

[24] 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 recogni tion, 2017, pp. 1125–1134.

[25] ″ INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge,″ GitHub, Jun. 17, 2022. https://github.com/microsoft/PLC-Challenge.

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

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, ″An al gorithm for intelligibility prediction of time–frequency weighted noisy speech,″ IEEE Transactions on Audio, Speech, and Lan guage Processing, vol. 19, no. 7, pp. 2125–2136, 2011.

[28] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O′Gorman, and A. Hines, ″Visqol v3: An open source production ready objec tive speech and audio metric,″ in 2020 twelfth international con ference on quality of multimedia experience (QoMEX). IEEE, 2020, pp. 1–6.

[29] A. Ali and S. Renals, ″Word error rate estimation for speech recognition: e-wer,″ in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 20–24.

[30] R. Kolobov, O. Okhapkina, O. Omelchishina, A. Platunov, R. Bedyakin, V. Moshkin, D. Menshikov, and N. Mikhaylovskiy, ″ Mediaspeech: Multilanguage asr benchmark and dataset, ″ arXiv preprint arXiv:2103.16193, 2021.

[31] R. Cutler, B. Nadari, M. Loide, S. Sootla, and A. Saabas, ″ Crowd sourcing approach for subjective evaluation of echo impairment,″ in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 406–410.

[32] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, ″ Interspeech 2022 audio deep packet loss concealment challenge,″ in INTERSPEECH 2022 - 23rd Annual Conference of the International Speech Communication Association, 2022 (submitted).

[33] C. K. Reddy, V. Gopal, and R. Cutler, ″ Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise sup pressors,″ in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6493–6497.

[34] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, ″A time-domain convolutional recurrent network for packet loss concealment, ″ in ICASSP 2021-2021 IEEE Interna tional Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7148–7152.

[35] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, ″ Audio visual speech inpainting with deep learning, ″ in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6653–6657.

Your Name	Title*	Affiliation	Research Field	Personal website
Linlin Ou	master student	Management Information Technology Department, Computer Network Information Center, Chinese Acadamy of Science	Audio/Video AI Enhancement, Audio/Video Coding	
Yuanping Chen	full professor	Management Information Technology Department, Computer Network Information Center, Chinese Acadamy of Science	Key technology research, data intelligence applications for large information systems	