Accepted author version.
Final version Digital Object Identifier 10.1109/OJSP.2025.3526552

The ICASSP 2024 Audio Deep Packet Loss Concealment Grand Challenge

Lorenz Diener[1], Solomiya Branets[1], Ando Saabas[1], Ross Cutler[1]
[1]Microsoft Corporation
Corresponding author: Lorenz Diener (email: lorenzdiener@microsoft.com).

ABSTRACT Audio packet loss concealment hides gaps in VoIP audio streams caused by network packet loss. It operates in real-time with low computational requirements and latency, as demanded by modern communication systems. With the ICASSP 2024 Audio Deep Packet Loss Concealment Grand Challenge, we build on the success of the previous Audio PLC Challenge held at INTERSPEECH 2022. For the 2024 challenge at ICASSP, we update the challenge by introducing an overall harder blind evaluation set and extending the task from wideband to fullband audio, in keeping with current trends in internet telephony. In addition to the Word Accuracy metric, we also use a questionnaire based on an extension of ITU-T P.804 to more closely evaluate the performance of systems specifically on the PLC task. We evaluate a total of 9 systems submitted by different academic and industry teams, 8 of which satisfy the strict real-time performance requirements of the challenge, using both P.804 and Word Accuracy evaluations. Two systems share first place, with one of the systems having the advantage in terms of naturalness, while the other wins in terms of intelligibility. These systems are the current state of the art for Deep PLC.

INDEX TERMS Audio, Digital Signal Processing, Packet Loss Concealment, Speech Coding

I. INTRODUCTION

As voice communication further transitions more and more towards calls that are fully packet-switched end-to-end (rather than being fully circuit-switched, or circuit-switched with a dedicated packet-switched backbone), the need for more robust packet loss concealment – the hiding of gaps in a stream caused by lost or late-arriving packets – has never been more evident. Since the tight latency requirements in real-time communication applications make large buffers and retransmission undesirable if not impossible, degraded network performance leads to audible gaps or annoying distortion in calls at the receiver side. Audio *Packet Loss Concealment* (PLC) is the task of fixing or hiding these gaps and making the audio stream appear as seamless as possible to allow for high quality communication even when packets get lost.

A. MOTIVATION

As algorithms and hardware have advanced, it is now possible to perform PLC using machine learning rather than basic digital signal processing, with the potential for vast quality improvements. In the PLC Challenge held at INTERSPEECH 2022, we for the first time brought together researchers working on the topic to compare approaches on a common dataset, with many interesting approaches and results [1].

B. CHANGES FROM THE 2022 PLC CHALLENGE

In this edition of the challenge, we build on this success, and make some changes based on lessons learned:

1) A more challenging dataset

The dataset in the 2022 PLC Challenge was, while not easy, still largely focused on scenarios where there is only a relatively low amount of packets lost, with not too many packets being lost in a row. Many participants built systems with good performance in these scenarios, but may not be able to perform well for longer sequences of losses. To challenge participants to also tackle harder cases with long burst losses, the dataset in this challenge focuses more on such cases. Additionally, while the 2022 challenge used wideband audio, the audio used in this edition is fullband, making the task once again somewhat more difficult, especially given the latency and compute constraints remain unchanged.

2) Better evaluation procedure

In the 2022 challenge, we performed objective evaluation using an ITU-T P.808 CCR procedure, obtaining a single rating for each file. In this challenge, we switch to the newer ITU-T P.804 [2] standard, in which listeners are asked to evaluate an audio file on multiple scales. The base

P.804 questionnaire asks listeners to rate audio files on four scales: Coloration, Loudness, Noisiness and Discontinuity. The extended variant that we use adds three additional scales on top of this: Reverberation, Signal Quality and Overall Quality.

II. Related work

PRIOR to the 2022 Deep Packet Loss Concealment challenge, work on machine learning-based PLC was relatively sparse. Many groups preferred to focus on the more general problem of audio inpainting (where, in contrast to PLC, there is no requirement for systems to be able to operate with no or minimal future information), or on classical DSP approaches.

A. CODEC PLC

Any audio codec intended for streaming audio transmission has to deal with the problem of what happens when a packet gets lost between sender and receiver, and therefore, has to perform PLC. Typically, this is done by assuming that some attributes, or rate of change of attributes, of the encoded features of the codec stay constant, and then simply continuing to decode. Modern examples of such techniques can be found in, e.g., the UMTS AMR codec [3] or the VoLTE EVS codec [4] that now power modern mobile telecommunication. In addition to basic linear prediction, these codecs also perform a more sophisticated analysis of the signal and try to continue it in a way that matches the type of signal previously received (e.g., treating voiced sections differently from non-voiced sections).

B. DEEP PLC

While audio inpainting has been studied for a long time, research into Deep PLC specifically has only picked up recently. This is because it has only recently become feasible to actually use neural systems for packet loss concealment – hardware has advanced to the point where deploying such models on edge devices is no longer prohibitive. We give a short overview of some recent work in the field.

Lin et al. [5] presents a basic approach based on predicting 320 future samples of wideband audio in the time domain. To achieve this, they use a convolutional-recurrent network trained using mean absolute error loss. The paper highlights some of the difficulties that are typical for the PLC task – phase prediction and continuity as well as longer-term prediction – but are nevertheless able to show improvement over a lossy signal in terms of various objective metrics (PESQ [6], STOI [7], and notably, Word Error Rate in terms of an in-house speech recognizer).

Finding a good loss for audio generation tasks can be challenging. Consequently, many approaches to Deep PLC rely on adversarial approaches. Examples of such approaches are Shi et al. [8] (using a time-domain convolutional encoder-decoder network structure, showing that solving this task is feasible at all), Pascual et al. [9] (presenting a system that maps Mel spectrogram input of received audio data to its time domain continuation and achieving results that compare favorably to codec PLC in terms of MCD [10] as well as SESQUA [11]) and Wang et al. [12] (using a U-Net style architecture and mixed frequency and time domain adversarial losses).

While the previous papers present interesting techniques and serve as a good introduction to the task and typical approaches and problems in Deep PLC, they are not easily comparable – they use different metrics (none of which are necessarily suitable or validated for the PLC task) and evaluate on different, in-house datasets. The 2022 PLC Challenge [1] for the first time allows us to compare approaches on a level playing field and with gold-standard human listening test evaluations. The winners of the challenge, Li et al. [13], present a system with a time domain convolutional encoder-decoder structure, trained with a large variety of different losses, including a spectrogram reconstruction loss, unsupervised audio representation-based loss, speech recognizer based loss, and adversarial losses. Valin et al. [14] place second with a network based on predicting the features of a neural vocoder. They also evaluate their method as part of the Opus codec, replacing the codec PLC entirely.

For further information about the results of the 2022 Deep PLC Challenge, please refer to the 2022 challenge overview paper [1]. For a deeper survey of neural PLC approaches from before then, refer to the 2020 survey paper by Mohamed et al. [15].

Beyond PLC for speech real-time communication, audio signals with other content may present different requirements and challenges. Verma et al. [16] and Mezza et al. [17] present work in Deep PLC for music signals in a networked performance setting.

C. REDUNDANCY AND FORWARD ERROR CORRECTION

Forward Error Correction (FEC) is the transmission of redundant information as part of the audio stream so that when one packet is lost, information from the surrounding packets can be used to fully reconstruct it. This trades off some latency for better audio quality and is typically used together with schemes that try to estimate network quality and bandwidth to avoid transmitting information that is not used. While out of scope for this edition of the challenge, an interesting direction that has emerged in neural PLC is deep redundancy – employing a neural network to encode redundant information to assist in Deep PLC [18]. For a survey of classical FEC schemes, please refer to Thirunavukkarasu et al. [19].

III. CHALLENGE DESCRIPTION

THE task of the 2024 PLC challenge is as follows. Participants are given two sequences:

- A short clip of audio data, mostly speech, provided at a sample rate of 48000 Hz. Some segments of the waveform are zeroed out.

- A sequence of binary values. For each 20 ms (960 samples) frame, these values indicate whether the frame was zeroed out (1) or not (0).

The objective is to fill in the gaps – and potentially process the remaining audio for better continuity – in a way that maximizes both the intelligibility and naturalness of the resulting speech. The system must have an algorithmic latency of at most one 20 ms frame, which can be split between buffer size and lookahead as needed. Additionally, it should achieve a real-time factor of less than one on an Intel Core i5 quad-core machine clocked at 2.4 GHz or equivalent processors.

A. DATASET CONSTRUCTION

The dataset for the ICASSP 2024 challenge is built upon the same framework as in the 2022 challenge, leveraging real-world packet loss patterns combined with data that is either in the public domain or crowd-sourced to be used in the challenge. This allows us to have a dataset that is very realistic in terms of containing packet loss patterns that systems might have to actually deal with in the real world, while not containing any personal information from actual calls.

1) Audio data

We use audio data that is in the public domain (conversational speech, sourced from the LibriVox Community Podcast)[1] or was collected by us explicitly for use in challenges (crowd-sourced read speech using a wide variety of both mobile phones and laptops for recording), allowing us to have a realistic and varied dataset while avoiding the potential for privacy issues.

Audio segments were selected by filtering using DNS-MOS [20] and manual inspection to avoid very noisy base audio clips and were cut to 10 to 15 seconds of length using the WebRTC Voice Activity Detection to avoid cutting off parts of words. For the public domain data, where we do not have control over the recording setup, we select only files where a substantial amount of energy is present in the upper-frequency bands (i.e., files that actually contain more than just wideband audio). All clips were normalized to -6 dBFS peak amplitude.

2) Packet loss traces

Since we want to evaluate systems under conditions that are as realistic as possible, we collect traces of packet losses from real Microsoft Teams calls. We convert these traces, which contain transmission timing and loss information for all packets in a call, into a more basic binary loss indicator by treating all packets that either do not arrive or arrive too late to be used in decoding in the actual call as lost, with a packet size of 20 ms. We then cut the traces into 15 second segments (750 values) and apply a stratified sampling strategy to get broad coverage of cases that we expect to be difficult for PLC systems to deal with.

Packet loss traces were selected as follows: First, we select packet loss trace segments according to the longest burst loss present (exclusive higher edge). We then select traces for 5 equally sized packet loss brackets (0% to 10%, 10% to 20%, 20% to 30%, 30% to 40%, above 40%) for each of these burst loss ranges. We select a total of 600 traces:

- **0–120 ms burst:** 20 traces per loss bracket, 5 × 20 = 100 traces total
- **120–500 ms burst:** 40 traces per loss bracket, 5 × 40 = 200 traces total
- **500–1000 ms burst:** 40 traces per loss bracket, 5 × 40 = 200 traces total
- **1000–3000 ms burst:** 20 traces per loss bracket, 5 × 20 = 100 traces total

While we expect any PLC system to perform well on very short bursts, and a good system to at least partially conceal losses for medium-length bursts, we also include data with very long burst losses in excess of a second. We do not expect systems that have to operate under real-time operation constraints to be able to fill these gaps with speech resembling the ground truth audio, but rather, to degrade gracefully in a way that minimally impacts communication.

We post-process traces by setting the first 25 frames to not be lost since it would be uninteresting to evaluate a PLC system's ability to generate audio with no prior information. Note that this post-processing is performed after trace selection, thus potentially shifting the distribution of files in loss percent and burst loss brackets.

In addition to these newly selected traces, we include a total of 200 traces also used in the 2022 PLC Challenge to allow for limited comparability.

3) Final assembly

We combine each trace with one audio segment, zeroing out samples for which the loss indicator in the trace indicates that the packet has been lost, producing a total of 800 files. We additionally add four files that have been specially constructed to allow us to check for violations of the latency requirement. These files are identical up to a certain point, after which they diverge. This allows us to check that the system output is also identical up to that point plus the allowed latency. On Oct. 11, 2023, we first released a validation set constructed in this way (including a script to check the algorithmic latency requirement), followed by a blind set with no references on Dec. 1, 2023. Participants submitted the output of their systems for this blind set for evaluation by the deadline of Dec. 7, 2023.

[1] Librivox Contributors, "The LibriVox community podcast", https://librivox.org/category/librivox-community-podcast/

The blind dataset (without ground truth data) and validation dataset (including ground truth data) can be downloaded from our challenge website.[2]

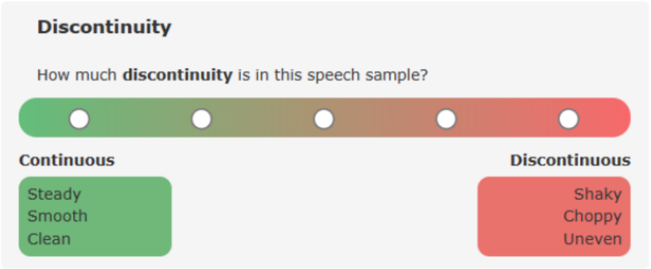
B. EVALUATION PROCEDURE

We evaluate the subjective perceived quality with the P.804 crowd-sourced evaluation procedure using the Amazon Mechanical Turk crowd-sourcing service. P.804 is a multi-dimensional audio quality evaluation questionnaire, evaluating a total of 4 aspects on a 1 to 5 scale, without reference, which, following Naderi et al. [2] we extend to 7 for our testing:

- Coloration
- Noisiness
- Loudness
- Discontinuity
- Reverb (extended P.804 only)
- Signal quality (extended P.804 only)
- Overall quality (extended P.804 only)

For quality control, we include both two *gold questions* (clips where the expected answer for a scale is known ahead of time, with either very low or very high quality) and one *trapping question* (questions where the rating clip is replaced by instructions to select a specific answer regardless of quality). Following the recommended procedures, we only use answers from listeners that consistently answer these quality control questions correctly [2]. After quality filtering, we obtain on average ~5 ratings for each clip.

To evaluate intelligibility, we use automatic speech recognition (based on the Azure Cognitive Services speech recognizer). We compare the speech recognizer output for each system against human-checked ground truth transcriptions, calculating the average Word Accuracy (*WAcc*, calculated as 1 - Word Error Rate).



Discontinuity

How much **discontinuity** is in this speech sample?

○ ○ ○ ○ ○

FIGURE 1. **P.804 "Discontinuity" factor question. Nb: The best score, 5, is on the left, while the worst, 1, is on the right.**

The final score according to which we rank systems is computed as the average of three values: P.804 Discontinuity (normalized to be between 0 and 1), Overall quality (normalized to be between 0 and 1), and Word Accuracy. The Discontinuity factor asks raters to evaluate how discontinuous the audio is, with the low end being audio that is shaky, choppy and uneven (1), and the high end (5) being audio that is steady, smooth and clean, while the Overall factor, raters are asked to give one score to the entire sample, from Bad (1) to Excellent (5). See Figure 1 for an example of how questions were presented to raters.

Averaging between Discontinuity score, Overall score and WAcc puts the focus on the aspects most important for the PLC challenge: That packet losses are concealed from humans with minimal impact on overall quality and without any sacrifices to intelligibility.

IV. RESULTS AND DISCUSSION

We evaluate the results for a total of 8 participants as well as clean (ground truth) and lossy (system input audio, as sent to participants) audio in the manner described in section B. The results can be found in Table 1. We also include one system that did not meet latency requirements (marked DNF). We perform statistical testing (one-tailed related-sample t-test between systems adjacent to each other in the scoreboard, no family-wise error rate correction) on the final score to see whether the differences we obtain are significant. Based on this, the winners of the ICASSP 2024 Audio Deep Packet Loss Concealment Grand Challenge are, sharing first place, teams 1024K and NWPU & ByteAudio.

A. 2024 WINNERS

An interesting aspect of the results of this year's challenge is that we have two winning submitted systems (with final scores that do not differ significantly), which achieve their score in different ways. For a more thorough explanation of these systems, please refer to the cited papers.

Team 1024K [21] improve on the winning system from the 2022 challenge [13], improving the encoder-decoder structure by adding a recurrent bottleneck and adjusting parameters to allow the system to deal with full-band audio. They use a very small buffer size (1ms), giving them ample room for lookahead (with their system using 4ms – so the overall system achieves an even lower latency than required). They also introduce a two-stage training procedure to speed up convergence, training first with a relatively low amount of packet loss and always using ground truth data as input to the model, even when such data would not be available during actual inference (i.e., when more than one frame is lost). In the second stage, they train with higher loss rates and also train on the systems output autoregressively. The system fills gaps very smoothly, allowing it to achieve the best scores in both the Overall as well as the Discontinuity factors. However, it tends to fill gaps with hallucinated audio, leading to a regression in word accuracy compared to the lossy data.

Team NWPU & ByteAudio [22] use a similar overall structure (convolutional-recurrent encoder-decoder network), however, there are some key differences. Their network operates in the frequency-domain, and focuses heavily on the wideband part of the spectrum (with most compute allocated

[2] https://aka.ms/plc_challenge

TABLE 1. ICASSP 2024 Audio Deep PLC Challenge results. Scores are averaged over all files in the blind set. The differences between systems are significant at $p < 0.05$ except where indicated (ns bracket). Columns that contributed to the final evaluation score in the challenge are highlighted, and the best system for every metric is bolded.

		Extended P.804 Scores										
Place	System	Coloration	Noisiness	Loudness	Discontinuity	Reverb	Signal	Overall	WAcc	PLCMOS	Final Score	
	Raw (Clean)	4.43	4.17	4.50	4.55	4.40	4.34	4.01	0.98	4.38	0.87	
1 (tied)	1024K [21]	4.05	**4.26**	**4.38**	**3.90**	**4.21**	**3.78**	**3.49**	0.81	**4.13**	0.72	ns
1 (tied)	NWPU & ByteAudio [22]	**4.11**	4.03	4.35	3.82	4.14	3.73	3.44	**0.84**	3.94	0.72	
3	SpeechGroupIoA [23]	4.05	4.23	4.35	3.67	4.15	3.64	3.37	0.81	3.99	0.69	
4	HWYW [24]	3.99	4.14	4.31	3.49	4.09	3.49	3.21	0.81	3.53	0.66	
5	LEIBUS [25]	3.74	3.82	4.17	2.94	3.87	2.98	2.75	0.84	3.09	0.59	
6	Regenerate	3.53	3.42	3.91	2.90	3.64	2.83	2.56	0.83	3.05	0.57	
	Raw (Lossy)	3.60	3.67	3.98	2.47	3.72	2.58	2.37	0.83	2.64	0.51	
7	CQUPT_ISARL	2.93	3.19	3.77	2.65	3.13	2.34	2.11	0.81	3.0	0.50	
8	NJUAcstcs	2.92	3.18	3.77	2.68	3.15	2.39	2.17	0.64	3.09	0.45	
(DNF)	Enchanto	3.73	3.59	4.17	3.36	3.85	3.21	2.91	0.82	3.35	0.63	

TABLE 2. Key aspects of the top 5 participants models

Team	Model Arch	Loss	#params	Features	Training data
1024K	Two convolutional-recurrent encoder-decoders (PLC stage & enhancement stage)	SI-SNR, Multi-Scale STFT MSE, PFPL [26], ASR-based	~3.67M	Time-domain (PLC stage) & STFT (enhancement stage)	PLC Challenge 2022 & AISHELL [27] & VCTK [28] & Librispeech [29], simulated loss (Gilbert-Elliot) [30] & real loss traces (PLC Challenge 2022)
NWPU & ByteAudio	Convolutional-recurrent encoder-decoder with band splitting	Power-law compressed STFT MSE, Time-Domain MSE, F0 prediction, Adversarial, Metric-GAN, ASR-based	3.81M	Power-law compressed STFT	DNS Challenge [31], simulated loss (Gilbert-Elliot) [30]
SpeechGroup IoA	Convolutional-recurrent encoder-decoder with band splitting & Linear-GRU enhancement stage	Multi-Scale STFT MSE, Time-Domain MSE, Log-compressed magnitude MSE, Adversarial with feature matching	4.9M	STFT	DNS Challenge [31], simulated loss (three state Markov model)
HWYW	Fully convolutional encoder-decoder with PQMF	Mel reconstruction, Adversarial	~2.36M	Time-domain PQMF	DNS Challenge [31], real loss traces (PLC Challenge 2022)
LEIBUS	Convolutional-recurrent encoder-decoder, explicit packet loss detection head & GRU-Conv enhancement stage	SI-SNR, loss probability MSE, STFT MSE	11.5M	Log-Mel & STFT	DNS Challenge [31], simulated loss (uniform random)

to the lower frequencies and the higher ones being treated separately). In addition to the typical reconstruction and adversarial losses, they add a MetricGAN [32] loss as well as a loss based on the Whisper [33] speech recognizer. This allows the system to actually improve the word accuracy score compared to the baseline, while also generating high quality and natural sounding output.

We summarize key aspects of the top 5 participants in Table 2. For further details, refer to the participants' conference papers or associated OJSP publications.

B. BREAKDOWN BY BURST SUBSETS

We break down the lossy clip as well as the first place model scores by burst subset, comparing the Overall score (Table 3) and WAcc (Table 4). In doing so, we can observe an interesting effect: While WAcc behaves predictably (decreasing as the burst loss length increases), the situation is different for the subjective listening test score. Here we find that the score actually increases for the lossy clips as burst losses get very large, while for the system outputs, it first decreases, and then, as losses get too long for the systems to meaningfully compensate, plateaus. This is in contrast to the 2022 challenge, where we did not observe such an effect. This could be because, to our listeners, a long quiet burst is preferable to many abrupt transitions between speech and silence. In the 2022 challenge, bursts were not long enough for this to happen, but when getting to upwards of half a second, the effect becomes pronounced. This highlights the importance of relying on multiple metrics to capture all aspects of a task.

TABLE 3. Overall MOS for different burst subsets. The scores are averaged over all of files in the blind set. Significant differences between participant models are marked with * (two-tailed related-sample t-test, $p < 0.05$).

Burst subset	Lossy	1024K	NWPU & ByteAudio
0ms to 120ms	2.11	3.74	3.75
120ms to 500ms	2.17	3.45	3.43
500ms to 1000ms	2.36	3.32*	3.19*
1000ms to 3000ms	2.61	3.34	3.22

TABLE 4. WAcc for different burst subsets. Scores averaged over all files in blind set. Significant differences between participant models are marked with * (two-tailed related-sample t-test, $p < 0.05$).

Burst subset	Lossy	1024K	NWPU & ByteAudio
0ms to 120ms	0.93	0.93*	0.95*
120ms to 500ms	0.83	0.83*	0.86*
500ms to 1000ms	0.77	0.77*	0.79*
1000ms to 3000ms	0.69	0.69*	0.70*

C. COMPARISON TO THE 2022 DEEP PLC CHALLENGE

To gauge improvement compared to the systems from the 2022 challenge, we perform a subjective P.808 listening test (as used in the 2022 challenge) on the data of the top 3 systems from the 2022 and 2024 challenges, using the shared-trace subset. To mitigate the effect of the audio data being different, we include clean ground truth audio from both the 2022 and 2024 sets, allowing us to compute DMOS for both editions – “how much worse is the P.808 MOS for this file compared to a perfect reconstruction?” – and downsample all clips to 16000 Hz for rating. We obtain a total of 5 ratings for each of the 200 clips in the overlap subset, for a total of 1000 ratings per system. The results are shown in Table 5.

The 2024 systems have to operate on full-band audio, while the 2022 systems, with the same compute and latency budget, did not need to. Despite this, there has been a substantial improvement in both the scores achieved by the best systems as well as those achieved by systems on average.

TABLE 5. Comparison of DMOS scores between 2022 and 2024 challenges, including 95% confidence intervals.

Model	P.808 DMOS
1st place (shared) 2024: 1024K	-0.24 ± 0.11
1st place (shared) 2024: NWPU & ByteAudio	-0.30 ± 0.12
3rd place 2024: SpeechGroupIoA	-0.37 ± 0.12
1st place 2022: Kuaishou	-0.39 ± 0.12
2nd place 2022: Amazon	-0.55 ± 0.13
3rd place (shared) 2022: ByteDance	-0.63 ± 0.13
3rd place (shared) 2022: Oldenburg University	-0.64 ± 0.13
Lossy (2024 data)	-1.13 ± 0.14

D. PERFORMANCE OF PLCMOS

Participants were encouraged to use the PLCMOS metric [34] to help them develop their systems. This comes with some caveats: PLCMOS was trained on wideband data (so system output has to be downsampled before being evaluated), and data with shorter burst losses than used in the 2024 Deep PLC Challenge. To evaluate the ability of PLCMOS to evaluate the 2024 systems, we compute the system-wise Pearson and Spearman correlation coefficients between the PLCMOS score and P.804 Discontinuity, Overall, and WAcc scores. The results are shown in Table 6.

TABLE 6. *Pearson Correlation Coefficient* (PCC) and *Spearman Rank Correlation Coefficient* (SRCC) for PLCMOS.

	PCC	**SRCC**
P.804 Discontinuity	0.96	0.92
Overall	0.94	0.91
WAcc	0.45	0.13

It can be seen that, while it still performs quite well, the metrics’ ability to predict relevant subjective evaluation ranks is worse than on the PLCMOS evaluation set (which includes data from the 2022 PLC Challenge). The correlation with WAcc, which is intended to evaluate a different aspect of system performance, is also – unsurprisingly – very low. This illustrates both the need for audio quality metrics that correlate well with human perception and that generalize, and the absolute need to, as long as such metrics do not exist and have not been broadly validated, perform sufficient testing with human listeners.

V. CONCLUSION

WITH the 2024 PLC Challenge, we aimed to move the field of Deep PLC forward by establishing a harder task with strong applicability to actual real-time communication systems. The results have clearly demonstrated the feasibility of solving this task, and shown that it is possible the improve both intelligibility and naturalness of the resulting speech while staying within stringent latency and compute requirements. Participants have shown that this task can be solved with both time- and frequency-domain approaches, and the winning systems demonstrate that using appropriate, but also varied losses is important to building a system that performs well.

We hope to continue to organize challenges in the same vein, moving real-time communication closer to being able to deliver perfect audio quality under ever more adverse network conditions. We would like to thank all participants for their submissions, which we hope will inspire further interesting research directions, and look forward to seeing their future work in this growing field.

REFERENCES

[1] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge," in *Proc. Interspeech 2022*, 2022. doi: 10.21437/Interspeech.2022-10829 pp. 580–584.

[2] B. Naderi, R. Cutler, and N.-C. Ristea, "Multi-dimensional speech quality assessment in crowdsourcing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. doi: 10.1109/ICASSP48485.2024.10447225

[3] 3rd Generation Partnership Project, "Adaptive Multi-Rate (AMR) speech codec; Error concealment of lost frames," in *Adaptive Multi-Rate (AMR) speech codec*, 2004.

[4] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015. doi: 10.1109/ICASSP.2015.7179065 pp. 5708–5712.

[5] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A Time-Domain Convolutional Recurrent Network for Packet Loss Concealment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021. doi: 10.1109/ICASSP39728.2021.9413595 pp. 7148–7152.

[6] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2001. doi: 10.1109/ICASSP.2001.941023 pp. 749–752.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011. doi: 10.1109/TASL.2011.2114881

[8] Y. Shi, N. Zheng, Y. Kang, and W. Rong, "Speech Loss Compensation by Generative Adversarial Networks," in *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2019. doi: 10.1109/APSIPAASC47483.2019.9023132 pp. 347–351.

[9] S. Pascual, J. Serrà, and J. Pons, "Adversarial auto-encoding for packet loss concealment," in *Proceedings of the 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021. doi: 10.1109/WASPAA52581.2021.9632730 pp. 71–75.

[10] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of the IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, May 1993. doi: 10.1109/PACRIM.1993.407206 pp. 125–128.

[11] J. Serrà, J. Pons, and S. Pascual, "Sesqa: semi-supervised learning for speech quality assessment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. doi: 10.1109/ICASSP39728.2021.9414052 pp. 381–385.

[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, Oct. 2021. doi: 10.1121/10.0006528

[13] N. Li, X. Zheng, C. Zhang, L. Guo, and B. Yu, "End-to-end multi-loss training for low delay packet loss concealment." in *INTERSPEECH*, 2022, pp. 585–589.

[14] J.-M. Valin, A. Mustafa, C. Montgomery, T. B. Terriberry, M. Klingbeil, P. Smaragdis, and A. Krishnaswamy, "Real-Time Packet Loss Concealment With Mixed Generative and Predictive Model," in *Proc. Interspeech 2022*, 2022. doi: 10.21437/Interspeech.2022-903. ISSN 2958-1796 pp. 570–574.

[15] M. M. Mohamed, M. A. Nessiem, and B. W. Schuller, "On Deep Speech Packet Loss Concealment: A Mini-Survey," *arXiv:2005.07794 [cs, eess]*, May 2020, arXiv: 2005.07794.

[16] P. Verma, A. I. Mezza, C. Chafe, and C. Rottondi, "A deep learning approach for low-latency packet loss concealment of audio signals in networked music performance applications," in *2020 27th Conference of open innovations association (FRUCT)*, 2020. doi: 10.23919/FRUCT49677.2020.9210988 pp. 268–275.

[17] A. I. Mezza, M. Amerena, A. Bernardini, and A. Sarti, "Hybrid packet loss concealment for real-time networked music applications," *IEEE Open Journal of Signal Processing*, pp. 266–273, 2023. doi: 10.1109/OJSP.2023.3343318

[18] J.-M. Valin, J. Büthe, and A. Mustafa, "Low-bitrate redundancy coding of speech using a rate-distortion-optimized variational autoencoder," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10096528

[19] E. Thirunavukkarasu and E. Karthikeyan, "A survey on VoIP packet loss techniques," *International Journal of Communication Networks and Distributed Systems*, vol. 14, no. 1, pp. 106–116, Jan. 2015. doi: 10.1504/IJCNDS.2015.066029 Publisher: Inderscience Publishers.

[20] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. doi: 10.1109/ICASSP39728.2021.9414878

[21] N. Li, G. Yu, C. Zhang, C. Zhou, Q. Huang, and B. Yu, "Multi-stage training for cross-domain full-band audio packet loss concealment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2024. doi: 10.1109/ICASSPW62465.2024.10626444

[22] Z. Zhang, J. Sun, X. Xia, C. Huang, Y. Xiao, and L. Xie, "Bs-plcnet: Band-split packet loss concealment network with multi-task learning framework and multi-discriminators," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW))*, 2024. doi: 10.1109/ICASSPW62465.2024.10627343

[23] L. Dai, Y. Ke, H. Zhang, F. Hao, X. Luo, X. Li, and C. Zheng, "A time-frequency band-split neural network for real-time full-band packet loss concealment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2024. doi: 10.1109/ICASSPW62465.2024.10626621

[24] B. Irvin, S. Yin, S. Zhang, and M. Stamenovic, "A fullband neural network for audio packet loss concealment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2024. doi: 10.1109/ICASSPW62465.2024.10627667

[25] X. Sun, Q. Li, K. Ma, L. Wang, and Y. Wang, "Two-stage neural network model with packet loss detection for icassp 2024 plc challenge," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2024. doi: 10.1109/ICASSPW62465.2024.10626654

[26] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving Perceptual Quality by Phone-Fortified Perceptual Loss Using Wasserstein Distance

for Speech Enhancement," in *Proc. Interspeech 2021*, 2021. doi: 10.21437/Interspeech.2021-582. ISSN 2958-1796 pp. 196–200.

[27] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017. doi: 10.1109/ICSDA.2017.8384449

[28] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015. doi: 10.1109/ICASSP.2015.7178964 pp. 5206–5210.

[30] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *The Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, 1963.

[31] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. doi: 10.1109/ICASSP43922.2022.9747230

[32] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*, 2019, pp. 2031–2041.

[33] 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*, 2023, pp. 28 492–28 518.

[34] L. Diener, S. Sootla, M. Purin, 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*, Aug. 2023. doi: 10.21437/Interspeech.2023-1532