INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge

Lorenz Diener, Sten Sootla, Solomiya Branets, Ando Saabas, Robert Aichner, Ross Cutler

Microsoft Corporation

lorenzdiener@microsoft.com

Abstract

Audio Packet Loss Concealment (PLC) is the hiding of gaps in audio streams caused by data transmission failures in packet switched networks. This is a common problem, and of increasing importance as end-to-end VoIP telephony and teleconference systems become the default and ever more widely used form of communication in business as well as personal usage.

This paper presents the INTERSPEECH 2022 Audio Deep Packet Loss Concealment challenge. We first give an overview of the PLC problem, and introduce some classical approaches to PLC as well as recent work. We then present the open source data set released as part of this challenge as well as the evaluation methods and metrics used to determine the winner, and give the challenge timeline. We also briefly introduce PLCMOS, a novel data-driven metric that can be used to quickly evaluate the performance PLC systems. Finally, we present the results of the INTERSPEECH 2022 Audio Deep PLC Challenge, and provide a summary of important takeaways.

Index Terms: packet loss concealment, speech enhancement, real-time processing

1. Introduction

With the transition of digital voice communication away from circuit switched analog telephony to end-to-end digital packet-switched telephone systems – voice-over-IP (VoIP) telephony – the issue of how to best process and transmit speech over packet networks has become increasingly important. One important aspect of this is how to handle situations where packet transmission breaks down: Ideally, we would like to build systems that are robust to a small amount of lost packages, and degrade gracefully for longer loss bursts. This process of restoring parts of the signal that are missing where possible, and filling in parts that cannot be recovered in a way that is minimally disturbing to the users of a system, is called “Packet Loss Concealment” (PLC).

Classically, PLC is treated as part of the design of the compression / coding and decompression / decoding algorithms (codecs) used to transmit speech, operating in the coding algorithms feature space. While these techniques allow for basic smoothing over of losses, they can sometimes struggle even with short losses: A study by Sun et al. [1] found that for codecs used in mobile telephony, any packet loss during a voiced speech segment causes almost the entirety of that segment to be degraded. Techniques that can be used independently of the base PLC approach, such as forward error correction – transmitting parts of frames multiple times in anticipation of a loss – can help for short losses, but put additional strain on a network connection that is already straining to transmit all packets.

A less studied but very promising approach that, due to advances in hardware and algorithms, has recently become practically viable, is performing PLC via machine learning and neural networks – Deep PLC. By leveraging audio and call data to build models that can predict future frames in a data-driven manner rather than by making the strong assumptions classical techniques require, it may be possible to transparently hide shorter losses near completely, and smooth over longer loss bursts in a less distracting way than current PLC implementations are able to.

An issue that complicates current Deep PLC research is that as it stands, there are no standard benchmark datasets and evaluation methods, so comparing approaches is extremely limited. With the INTERSPEECH 2022 Deep PLC challenge, we hope to rectify this situation by introducing a realistic evaluation dataset that draws packet loss patterns from actual calls, a subjective quality methodology for evaluation that should make approaches more comparable, and a new objective metric that we hope will help researchers to iterate their approaches more quickly.

In the following paper, we will review classical PLC implementations, contrast them with some recent work in Deep PLC, and finally, present the dataset, evaluation methods and results from the INTERSPEECH 2022 Deep PLC Challenge.

2. Background and related work

2.1. Classical approaches

Classically, PLC is performed in the feature space of the codec used to packetize and en- and decode speech data, preceding the decoding step. This is, of course, especially necessary if the codec in question relies on information from consecutive packages in decoding. The techniques used for this have been iteratively improved with newer generations of codecs, but have in essence remained the same from the original work done for the Global System for Mobile Telecommunication [2] to the modernized Adaptive Multi-Rate Wideband codec used in UMTS [3] and the EVS codec used in Voice-over-LTE [4].

The basic principle of these approaches is to continue decoding as if the change in the coded speech parameters from the last known-good frames continues according to some expert-crafted fixed prediction function (linearly, in the simplest case), with gradual attenuation and replacement of the signal with comfort noise. More modern codecs improve upon this scheme by classifying missing frames into different types (e.g., silence, voiced speech, non-periodic, ...) and using different prediction schemes for different frame types.

An additional technique employed in many codecs (e.g. EVS) on top of these methods is forward error correction: When bad network conditions are detected, the sender may decide to transmit redundant information about past frames so that short losses can be better compensated for if the next frame after the loss is already available. The downside of this technique is that it introduces network overhead and additional latency.

For a deeper survey on past proposed schemes, refer to Thirunavukkarasu et al. [5].

2.2. Machine learning approaches

2.2.1. Audio inpainting

The field of *audio inpainting* concerns itself with filling gaps in an audio signal, with recent approaches largely favoring statistical methods. This is related to the PLC problem, though in a relaxed form: It is often assumed that the entire signal, other than the parts that need to be in-painted, is available, and computation time is not generally considered.

A simple statistical approach is presented by Rodbro et al. [6]. They propose to use a Hidden Markov Model on the pitch, gain and spectral envelope of packets that can then be used to either directly predict future frames, directly fill gaps between frames, or be used as a frame type predictor to choose a prediction scheme.

Bahat et al. [7] present a dictionary based scheme: They learn a dictionary of audio blocks and continue an audio sequence by finding the best matching audio block to continue a sequence with. To do this, they use both a Markov model as well as the feature space distance between the start of the block and the last known good part of the audio sequence. The dictionary is created on the fly from correctly transmitted audio. This is a simple and effective way to leverage audio data to fill audio gaps, though computationally expensive and not easily adaptable beyond a single speaker.

Neural network approaches that consider the whole spectrogram are presented by Kegler et al.and Nair et al.. The former treats the task of audio inpainting fully like a vision task: They train a U-Net [8] model to map masked spectrograms (magnitude and phase angle), optionally with an additional channel indicating whether a part of the spectrum was masked or not, to the unmasked spectrum, using a perceptual VGG network based loss. They report an improvement over a baseline linear predictive coding approach for most cases. The latter use a joint time-frequency approach that also performs general speech enhancement: They use a time-domain U-Net to fill gaps and then use a frequency domain U-Net on the magnitude output of the first network to remove distortions (while keeping the phase from the time-domain U-Net signal).

Others approach the problem from a multi-modal perspective: Both Zhou et al. [9] and Morrone et al. [10] present approaches where an available video feed of the speaker is used to assist in the restoration of lost audio segments, using a convolutional neural network with adversarial loss and a recurrent neural network approach respectively. While approaches such as these are outside of the purview of current deep PLC research, they may become relevant in the context of video telephony, where in future techniques available video frames may help reconstruct the audio stream and vice-versa.

2.2.2. Real-time deep packet loss concealment

In recent years, both hardware and algorithms have advanced dramatically, allowing neural network based approaches to enter practical use in many areas of speech processing. This is also happening for PLC. Here, we give an overview of recent work targeting real-time, causal usage.

Stimbert et al.present an approach based on WaveRNN [11]: They condition a WaveRNN on the recent time domain history as well as, through a convolutional conditioning network operating in the frequency domain, more long term history. Unlike other methods presented in this section, this network outputs samples autoregressively instead of outputting full blocks of audio data at a time. They present methods for how their network can be used for real-time inference, operating on the jitter buffer of the NetEQ codec used in the Google Duo telecommunication software [12], and state that their implementation of this method significantly improved user satisfaction metrics.

One issue when trying to train neural networks for audio generation is what loss to use. Generative adversarial networks provide a framework within which a loss does not need to be defined, but can be co-learned by training a discriminator network that tries to classify a sample as being generated or from the training set, and then training the generation network to try to fool the discriminator. GAN based approaches have been shown to be able to efficiently generate audio waveforms [13]. A basic initial PLC approach based on GANs is presented by Shi et al. [14]. They train a convolutional encoder-decoder network operating on time domain audio blocks. They report comparable quality in terms of several objective metrics (PESQ [15], STOI [16], SNR) compared to a frequency deep neural network even when this baseline has perfect phase information available and only needs to predict magnitude. Pascual et al. [17] present a GAN based approach where the generator input is the Mel-spectrogram of the available signal, and the output is the time-domain continuation of this signal. They show that this method improves upon several baselines (including some of the systems used in real codecs, such as NetEQ and ITU G722.1) in terms of Mel-Cepstral Distortion [18] as well as SESQUA [19]. Finally, Wang et al. [20] present a GAN based system with a fully time-domain U-Net style convolutional generator and mixed time/frequency domain discriminator, which allows their adversarial loss to both capture fine short-term details in the waveform as well as long-term relationships in the spectrum. They evaluate in terms of PESQ, STOI and SNR and are able to show an improvement over several strong baselines – even over non-causal models when their own model is being used causally.

Other methods show that Deep PLC systems can be trained even with simpler losses. Lin et al. [21] present a convolutional-recurrent model performing next frame prediction in the time domain, trained to minimize the mean absolute error, with or without look-ahead. In addition to the usual metrics, they also perform an evaluation in terms of speech recognizer word error rate, and are able to show that Deep PLC has the potential to improve this metric as well. Similarly, Mohamed et al. [22] present a recurrent neural network for packet loss concealment in the context of performing far-end emotion recognition. They are able to show that using such a network before their emotion classifier improves classifier accuracy, further illustrating the potential of Deep PLC to improve the performance on downstream speech processing tasks.

3. Challenge description

The objective of the INTERSPEECH 2022 Audio Deep Packet Loss Concealment challenge is to fill regions where audio has been lost due to packet loss in such a way as to ideally hide these losses from listeners and to maximize intelligibility.

Note that effects from de- and encoding or buffering and time stretching are not considered as part of this challenge.

3.1. Dataset and evaluation

The data set provided for the INTERSPEECH 2022 Audio Deep Packet Loss Concealment challenge consists of a set of audio and meta-data files, divided into three splits: A train split, validation split and blind test split. The first two will be released on January 19th, 2022, while the blind test split will released on March 1st, 2022. The training and validation splits contain, for each included sample of audio:

- A clean audio file
- A ”lossy” audio file with packet loss regions zeroed out
- A text file with loss metadata

The metadata indicates, with one line foe each 20 millisecond segment of audio, whether the packet has been lost (file contains a ”1”) or not (file contains a ”0”). The ”lossy” files have the corresponding segments zeroed out. Additionally, the dataset includes, for each split, a file with file-level metadata about the included clips in csv format. The training set is intended as a quick-start set for training models to perform PLC, though participants are free to use any other data sets as they desire to improve their models.

The blind set does not contain clean audio, and contains additional audio files for word recognition rate evaluation.

3.1.1. Dataset construction

The dataset was constructed by stratified sampling of actual packet loss traces observed in calls made by Microsoft Teams users, applied to randomly chosen segments of audio from a podcast dataset.

The traces were sampled in the following way: First, 10 second segments with at least one packet loss were randomly extracted from packet loss traces from Teams calls. These segments were then divided into three subsets according to the maximum burst loss length in the trace:

- up to 120 milliseconds
- between 120 and 320 milliseconds
- between 320 and 1000 milliseconds

Traces with burst losses longer than 1000 milliseconds were discarded. Each subset was then divided into 14 cells according to packet loss quantiles. Finally, an equal amount of traces was sampled from each cell (with more traces being sampled for the subsets with shorter maximum burst losses).

To create audio with packet losses, clips of audio of approximately 10 seconds of length were sampled from a base public domain podcast data set, cutting in low volume regions to attempt to not split up words. The losses from the sampled traces were then applied to the resulting audio files by zeroing out the corresponding regions in the audio clips.

3.1.2. Evaluation

The evaluation criterion is split into two parts, weighted equally: Crowd-Sourced Mean Opinion Score, and Word Accuracy. The Mean Opinion Score is calculated by having raters rate clips using a crowd-sourcing framework [23], obtaining 5 ratings for each clip in the blind set. The word accuracy is calculated using Microsoft Azure cognitive services speech recognition.

3.2. Timeline

Jan. 19, 2022	Training + validation dataset will be released
Mar. 1, 2021	Blind test dataset will be released
Mar. 8, 2022	Deadline for submission of challenge results
Mar. 18, 2022	Announcement of the challenge results to authors
Mar. 21, 2022	INTERSPEECH 2022 submission deadline
Jun. 23, 2022	Camera-ready papers due
Sep. 18, 2022	Start of INTERSPEECH 2022

3.3. PLC-MOS

(To be filled for final paper)

4. Results

(To be filled for final paper)

5. Conclusion

(To be filled for final paper)

6. References

[1] L. F. Sun, G. Wade, B. M. Lines, and E. C. Ifeachor, “Impact of Packet Loss Location on Perceived Speech Quality,” in *In 2nd IP-Telephony Workshop*, 2001, pp. 114–122.

[2] K. Hellwig, P. Vary, D. Massaloux, J. Petit, C. Galand, and M. Rosso, “Speech codec for the European mobile radio system,” in *1989 IEEE Global Telecommunications Conference and Exhibition ’Communications Technology for the 1990s and Beyond’*, Nov. 1989, pp. 1065–1069 vol.2.

[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 *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5708–5712, iSSN: 2379-190X.

[5] 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, publisher: Inderscience Publishers.

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

[7] Y. Bahat, Y. Y. Schechner, and M. Elad, “Self-content-based audio inpainting,” *Signal Processing*, vol. 111, pp. 61–72, Jun. 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *arXiv:1505.04597 [cs]*, May 2015, arXiv: 1505.04597 version: 1.

[9] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, “Vision-Infused Deep Audio Inpainting,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 283–292.

[10] 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)*, Jun. 2021, pp. 6653–6657, iSSN: 2379-190X.

[11] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," *arXiv:1802.08435 [cs, eess]*, Feb. 2018, arXiv: 1802.08435 version: 1.

[12] "Improving Audio Quality in Duo with WaveNetEQ."

[13] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *arXiv:2010.05646 [cs, eess]*, Oct. 2020, arXiv: 2010.05646.

[14] 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)*, Nov. 2019, pp. 347–351, iSSN: 2640-0103.

[15] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2. Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.

[16] 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, conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[17] S. Pascual, J. Serrà, and J. Pons, "Adversarial Auto-Encoding for Packet Loss Concealment," *arXiv:2107.03100 [cs, eess]*, Jul. 2021, arXiv: 2107.03100.

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

[19] J. Serrà, J. Pons, and S. Pascual, "SESQA: semi-supervised learning for speech quality assessment," *arXiv:2010.00368 [cs, eess]*, Feb. 2021, arXiv: 2010.00368.

[20] 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, publisher: Acoustical Society of America.

[21] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7148–7152, iSSN: 2379-190X.

[22] M. M. Mohamed and B. W. Schuller, "ConcealNet: An End-to-end Neural Network for Packet Loss Concealment in Deep Speech Emotion Recognition," *arXiv:2005.07777 [cs, eess]*, May 2020, arXiv: 2005.07777.

[23] R. Cutler, B. Nadari, M. Loide, S. Sootla, and A. Saabas, "Crowdsourcing approach for subjective evaluation of echo impairment," 2020, arXiv: 2010.13063.