Hybrid Packet Loss Concealment for Real-Time Networked Music Applications

Alessandro Ilic Mezza, Matteo Amerena, Alberto Bernardini, and Augusto Sarti

Image and Sound Processing Lab
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Milan, Italy

in *IEEE Open Journal of Signal Processing*, vol. 5, pp. 266-273, 2024.

Abstract

Real-time audio communications over IP have become essential to our daily lives. Packet-switched networks, however, are inherently prone to jitter and data losses, thus creating a strong need for effective packet loss concealment (PLC) techniques. Though solutions based on deep learning have made significant progress in that direction as far as speech is concerned, extending the use of such methods to applications of Networked Music Performance (NMP) presents significant challenges, including high fidelity requirements, higher sampling rates, and stringent temporal constraints associated to the simultaneous interaction between remote musicians. In this article, we present PARCnet, a hybrid PLC method that utilizes a feed-forward neural network to estimate the time-domain residual signal of a parallel linear autoregressive model. Objective metrics and a listening test show that PARCnet provides state-of-the-art results while enabling real-time operation on CPU.

Parallel Linear Predictor

Assuming a linear AR model for the short-time signal under consideration, we can estimate the missing packets as the linear combination of past samples plus a residual noise term.

$$y[n] = \sum_{i=1}^{p} \varphi_i y[n-i] + \varepsilon[n]$$

In practice, however, the residual of a finite-memory linear model is far from being white, to the detriment of audio quality. The key idea behind PARCnet is to let a feed-forward neural network predict the residual term from the past valid context in a nonlinear fashion.

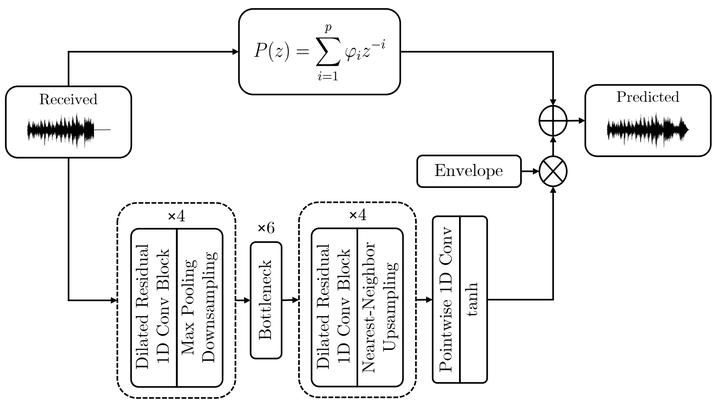
PARCnet

PARCnet is a hybrid PLC method for real-time NMP applications. It comprises two parallel modules: a linear predictor and a deep neural branch, both operating in the time domain. PARCnet recasts the problem of causal PLC into that of predicting the residual signal of an AR model instead of trying to generate a coherent waveform from scratch.

$$\begin{cases} \hat{y}[n] = y[n] + f_\theta(\mathbf{x})_{n-k} \\ y[n] = \sum_{i=1}^{p} \varphi_i y[n-i] \end{cases}$$

PARCnet neural branch embeds a fully-convolutional causal information bottleneck model implementing a **feed-forward frame-by-frame inference mechanism** that drastically expedites computations compared to existing deep autoregressive PLC methods.

PARCnet is trained to reconstruct 8-packet sequences, where the last packet is assumed to be “lost” and thus zero-filled.

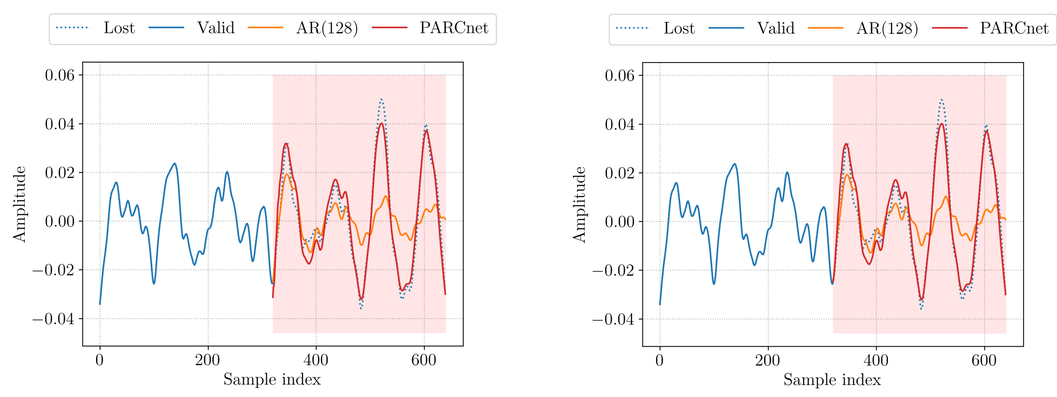


We integrate AR(128) into PARCnet linear branch. The AR model is fitted via the autocorrelation method by applying white noise compensation and the Levinson-Durbin algorithm. The autocorrelation function is computed over a 100 ms context window with stride of 10 ms, where the signal is assumed to be wide-sense stationary.

We train PARCnet with a **multiresolution spectro-temporal loss** including time-domain MSE, spectral convergence, and the L^1-norm of the regularized log-magnitude error.

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \frac{\lambda}{Q} \sum_{q=1}^{Q} \left(\mathcal{L}_{\text{sc}}^{(q)} + \mathcal{L}_{\log}^{(q)} \right)$$

Since AR(128) is proficient in estimating the first few samples, whereas the feed-forward neural network might introduce discontinuities. **the neural contribution is modulated with a time-domain envelope** (upward ramp).



Evaluation

We train all methods using 28 h of piano music from MAESTRO [2] resampled at 32 kHz. We test on 1 h of held-out data and, following [1], simulate evenly-spaced 10 ms losses with a loss rate of 10%.

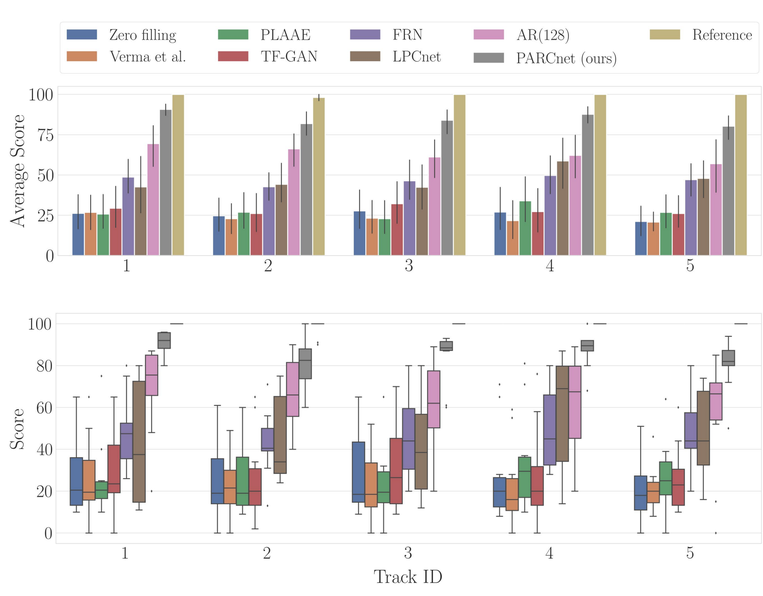
To ease outbound transitions, every AR-based model, including PARCnet, forecast 25% more samples than a packet length and apply a linear crossfade.

	NMSE (dB) ↓	Mel-SC ↓	PEAQ ↑	PLCMOS ↑	# Params ↓	CPU Time ↓
Zero-Insertion	0.0	0.242	−3.86 ± 0.03	1.88	—	—
AR(32)	−1.02	0.238	−2.35 ± 0.33	1.93	—	**1.06 ms**◇
AR(64)	−2.03	0.214	−2.15 ± 0.33	1.95	—	1.54 ms◇
AR(128)	−3.39	0.187	−1.84 ± 0.32	1.98	—	1.84 ms◇
PLAAE	1.7	0.198	−3.61 ± 0.17	1.91	1 M	34 ms*
FRN	0.85	0.214	−3.33 ± 0.26	1.92	8.6 M	14.9 ms
LPCNet	0.3	0.202	−2.43 ± 0.51	1.93	5.9 M	7.1 ms†
Verma et al.	−1.3	0.175	−3.89 ± 0.03	1.92	19 M	13.5 ms
TF-GAN	−1.6	0.149	−3.77 ± 0.20	1.94	2.2 M	18 ms
PARCnet (Ours)	**−5.2**	**0.136**	**−1.42 ± 0.19**	**2.07**	**416 k**	8.1 ms‡

◇ Accelerated using Numba JIT compiler; including model fitting. * Including post-inference maximum-correlation alignment.

† Obtained using the highly-optimized C implementation of the inference model provided by the authors. ‡ Only considering neural network inference.

We also conducted a **MUSHRA test**, with 16 musically-trained participants aged 25 to 37 rating five 10-second piano excerpts from MAESTRO on a scale of 0 to 100.



Audio Examples

Paper

PARCnet GitHub

2024 IEEE International Conference on Acoustics, Speech and Signal Processing
14-19 April 2024, Seoul, South Korea

References

[1] 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 *Proc. Conf. of Open Innovations Association*, 2020, pp. 268–275.

[2] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in Proc. Int. Conf. Learn. Representations, 2019.