The IEEE-IS² 2024 Music Packet Loss Concealment Challenge

Alessandro Ilic Mezza
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
alessandroilic.mezza@polimi.it

Alberto Bernardini
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
alberto.bernardini@polimi.it

***Abstract*—We present the IEEE-IS² 2024 Music Packet Loss Concealment Challenge. We begin by detailing the challenge rules, followed by an overview of the provided baseline system, the blind test set, and the evaluation methodology used to determine the final ranking. This inaugural edition aimed to foster collaboration between researchers and practitioners from the fields of signal processing, machine learning, and networked music performance, while also laying the groundwork for future advancements in packet loss concealment for music signals.**

***Index Terms*—Packet loss concealment, Internet of Sounds, networked immersive audio, networked music performance**

I. Introduction

Packet loss, either by missing packets or high packet jitter, is one of the main problems and, in turn, engineering challenges for real-life Networked Music Performance (NMP) applications. While Packet Loss Concealment (PLC) for Voice over IP has recently attracted a great deal of attention, as also evidenced by the recent INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge [1] and the 2024 ICASSP Audio Deep Packet Loss Concealment Grand Challenge [2], PLC for NMP has been considerably less studied.

The IEEE-IS² 2024 Music Packet Loss Concealment Challenge[1] aimed to provide a platform for researchers and practitioners working on the topic to share their work and compare different methods within a unified benchmark, in an effort to encourage further advancements in the field of Music PLC.

II. Challenge Overview

A. Challenge Rules

The IEEE-IS² 2024 Music Packet Loss Concealment Challenge lunched on May 13, 2024; the blind test set was released on July 3, 2024; the submission window closed on July 20, 2024. Each team was allowed to submit up to two systems for evaluation.

The systems had to be designed to process audio files at a sampling rate of 44.1 kHz, and predict packets of 512 samples, corresponding to approximately 11.6 ms.

Whereas smaller packets are sometimes preferred in NMP application, the choice of using packets of 512 samples was meant to be challenging for the proposed PLC systems and encourage the participants to tackle harder test cases with long-term losses.

Additionally, motivated by the tight latency requirements of real-time networked musical interaction, only causal systems were deemed eligible for the Challenge. Namely, at any given time, only previously received packets or prediction thereof may be used to predict the next audio frame. In other words, differently from other audio deep PLC challenges, systems were not allowed any look ahead. Other than that, there were no limitations on the eligible PLC methods, which may comprise one or more deep-learning models, traditional signal processing algorithms, or a hybrid approach.

We did not provide training data, nor did we indicate a list of eligible training datasets. However, the Challenge prescribed that participants only used data from publicly-available and freely-accessible datasets. No limit, instead, was posed on data augmentation, as long as the models were kept blind to metadata and other auxiliary information other than packet loss traces.

We encouraged all participants to develop PLC systems that would respect real-time constraints as strictly as possible. However, slower-than-real-time inference was not accounted as a reason for disqualifying a submission.

TABLE I
Challenge Ranking

	Average score ± SD	Median score	Trials won	Ranking
PARCnet-IS² (Baseline)	58.06 ± 22.13	59.5	9	1st
Aironi et al. (full) [3]	49.14 ± 22.09	50.5	1	2nd
Aironi et al. (lite) [3]	48.06 ± 22.48	50.0	–	3rd
Daniotti et al. [4]	41.70 ± 21.38	41.0	–	4th
Zero-filling	5.14 ± 8.55	0.0	–	–

B. Baseline System

We released a baseline system for the IEEE-IS² 2024 Music Packet Loss Concealment Challenge. The system, dubbed **PARCnet-IS²**, is a modified PARCnet architecture [5] trained on Medley-solos-DB [6].

PARCnet comprises two parallel modules, an autoregressive linear predictor (AR model) and a feed-forward neural network. The linear predictor is fitted in real-time within a sliding context window using the autocorrelation method

[1]Challenge web page: https://internetofsounds2024.ieee-is2.org/program/ieee-is²-2024-music-packet-loss-concealment-challenge

with white noise compensation, while the neural network is trained to estimate the residual of the AR model. Compared to the original method, PARCnet-IS² incorporates several minor modifications. Namely, (i) PARCnet-IS² was trained for 250,000 steps using a L^1-loss instead of a L^2-loss; (ii) the audio signals were sampled at 44.1 kHz instead of 32 kHz; (iii) the system is designed to predict packets of 512 samples instead of 320; (iv) the neural network valid context was increased from 7 to 8 packets; (v) the order of the parallel AR model was increased to 256; (vi) the extra prediction length, which allows to cross-fade between subsequent packets, either valid or predicted, was increased from 80 to 256 samples. For more details, we refer the readers to [5]. PARCnet-IS² is available online.[2]

C. Blind Test Set

The IEEE-IS² 2024 Music Packet Loss Concealment Challenge blind test set[3] consists of 162 single-channel audio files in a 16bit-44.1kHz wav format extracted from AVAD-VR [7], a publicly available dataset of anechoic audio and 3D-video recordings of several small music ensemble performances. Every test audio file consists of a 11.6-second clip of a closed-miked classical or jazz performance obtained by segmenting the full recording with no overlap. The blind set thus comprises various acoustic instruments, including violin, cello, clarinet, sax, double bass, and classical guitar. Clips in which silence made up more than 30% of the total duration were discarded.

The audio clips are artificially degraded by dropping packets (zero-filling) according to predetermined "packet traces," i.e., text files containing a string of binary digits: 0 if a packet was correctly received and 1 if the packet was lost. Every digit in a packet trace corresponds to 512 samples. Traces do not contain explicit temporal information, and the packet rate is implicitly determined by the audio sampling rate. The packet traces used to create the blind test set were repurposed from the blind set of the INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge.[4]

Said traces are measured and represent a real network scenario. The text files are divided into three subsets according to the maximum burst loss length: **Subset 1.** bursts of up to 6 consecutive packets; **Subset 2.** bursts of 6 to 16 consecutive packets; **Subset 3.** bursts of 16 to 50 consecutive packets. We sampled packet traces from Subset 1 (with a probability of 90%) and Subset 2 (with a probability of 10%). We did not sample traces from Subset 3. For each audio clip, we sampled and concatenated up to three traces so as to encompass the entire clip duration.

The clean, untampered versions of the degraded audio clips in the test set were kept private, making it a *blind* set; these files were used for both objective and subjective evaluations.

[2]Available: https://github.com/polimi-ispl/2024-music-plc-challenge/tree/main/parcnet-is2

[3]Available: https://github.com/polimi-ispl/2024-music-plc-challenge

[4]Available: https://github.com/microsoft/PLC-Challenge

D. Evaluation Procedure

Challenge participants were asked to download the blind test set, process each and every clip with the proposed PLC method, and submit the enhanced audio files. Similarly to [2], no model was collected and run during the evaluation process.

The Challenge ranking was determined through a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test. A subset of ten audio files was manually selected from the blind test set so as to encompass different musical instruments and playing styles.

The MUSHRA test thus consisted of ten *trials*, where the test conditions were compared with the clean, untampered track in terms of Basic Audio Quality (BAQ). In each trial, the test condition that received the highest average score was considered the *winner* of that trial. The final ranking was determined by counting the *number of trials won* by each PLC method. The average MUSHRA scores computed across all trials were considered as a tie-breaker.

III. Team Submissions

We received three systems from two teams. Additional information on each method can be found on the respective technical reports.[5]

Daniotti et al. [4] submitted a variant of the original PARCnet model trained using a novel *Tilt Loss*. This perceptually-motivated L^1-loss function adaptively reweights the frequency axis of the mel-spectrogram error emphasizing the high-frequency range, akin to an upward tilt filter. The proposed PLC system was trained on the *Bach Cello Suite* dataset [8].

Aironi et al. [3] proposed a novel PLC method that uses a bin2bin Generative Adversarial Network (GAN) [9] to generate audio conditioned by the estimate of a linear predictor. The bin2bin model is trained with a linear combination of spectral convergence, log-magnitude STFT loss, and least-square conditional GAN objectives. Aironi and colleagues submitted two systems, a *full* model (54.4 M parameters) and a *lite* model (3.4 M parameters), each trained on an ensemble of three datasets: Medley-solos-DB [6], the Good-sounds.org dataset [10], and 45 hours of MIDI clips from MAESTRO [11] synthesized using SoundFonts.

IV. Evaluation

A. Objective Evaluation

Here, we provide a brief overview of the objective metrics considered as part of the systems evaluation. These metrics have been calculated on all clips in the blind test set. Even if prior studies have observed a statistically significant correlation between some objective metrics and subjective judgments [5], no metric has been definitely proved to work for Music PLC algorithms. For this reason, the final ranking was only determined from the outcome of the MUSHRA test (Section IV-B).

[5]The technical reports are available at https://internetofsounds.net/ieee-is²-2024-music-packet-loss-concealment-challenge

TABLE II
IMPAIRMENT DESCRIPTION FOR PLCMOS AND PEAQ SCORES.

Impairment description	PLCMOS	PEAQ ODG
Imperceptible	5.0	0.0
Perceptible, but not annoying	4.0	−1.0
Slightly annoying	3.0	−2.0
Annoying	2.0	−3.0
Very annoying	1.0	−4.0

Let $y[n]$ and $\hat{y}[n]$ be the N-sample reference and enhanced waveforms, respectively. We also define the vectors $\mathbf{y} = [y[0], ..., y[N]]^T$ and $\hat{\mathbf{y}} = [\hat{y}[0], ..., \hat{y}[N]]^T$. We denote the L^2-norm with $\|\cdot\|_2$.

As far as time-domain metrics are concerned, we compute the Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{n=0}^{N-1}\left(y[n] - \hat{y}[n]\right)^2 , \tag{1}$$

the Signal-to-Distortio Ratio (SDR)

$$\text{SDR} = 10\log_{10}\frac{\|\mathbf{y}\|_2^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2} , \tag{2}$$

and the Scale-Invariant SDR (SI-SDR) [12]

$$\text{SI-SDR} = 10\log_{10}\frac{\|\alpha\mathbf{y}\|_2^2}{\|\alpha\mathbf{y} - \hat{\mathbf{y}}\|_2^2} , \tag{3}$$

where $\alpha = \hat{\mathbf{y}}^T\mathbf{y}/\|\mathbf{y}\|_2^2$.

Additionally, we take into account metrics in the frequency and cepstral domain, respectively. Namely, we compute the Log-Spectral Distance (LSD) [13]

$$\text{LSD} = \frac{1}{M}\sum_{m=0}^{M-1}\sqrt{\frac{1}{K}\sum_{k=0}^{K}\log|Y_m[k]|^2 - \log|\hat{Y}_m[k]|^2} , \tag{4}$$

where $|Y_m[k]|$ is the $(K+1)$-bin magnitude spectrum of the m-th reference signal frame $y_m[n]$ of length 2048 samples, extracted using a Hann window with a hop size of 512. Next, we compute the Mel-Cepstral Distance (MCD) [14]

$$\text{MCD} = \frac{1}{M}\sum_{m=0}^{M-1}\sqrt{\sum_{i=1}^{16}\left(C_m[i] - \hat{C}_m[i]\right)^2} , \tag{5}$$

where $C_m[i]$ is the ith MFCC extracted from the 1024-sample frame $y_m[n]$ integrated over 20 critical bands using a triangular mel-filterbank. Note that the zeroth coefficient was excluded [14].

Furthermore, we compute an Objective Difference Grade (ODG) for each clip in the blind set with the ITU-R BS.1387 Perceptual Evaluation of Audio Quality (PEAQ) [15]. In particular, we use the open-source MATLAB implementation of the PEAQ Basic algorithm by P. Kabal [16]. For this metric, every clip had to be upsampled from 44.1 to 48 kHz.

Finally, we evaluate the latest version[6] of PLCMOS [17], recently released for the ICASSP 2024 Audio Deep Packet Loss Concealment Grand Challenge. Note that this data-driven non-intrusive metric was proposed for and trained with corrupted speech signals. Hence, it is unclear whether it is reliable when it comes to Music PLC. Moreover, PLCMOS is intended for signals at a sampling rate of 16 kHz. Therefore, every clip had to be downsampled accordingly. Here, we report the PLCMOS results for completeness. The impairment descriptions for the PEAQ and PLCMOS scores are reported in Table II.

[6]Available: https://pypi.org/project/speechmos

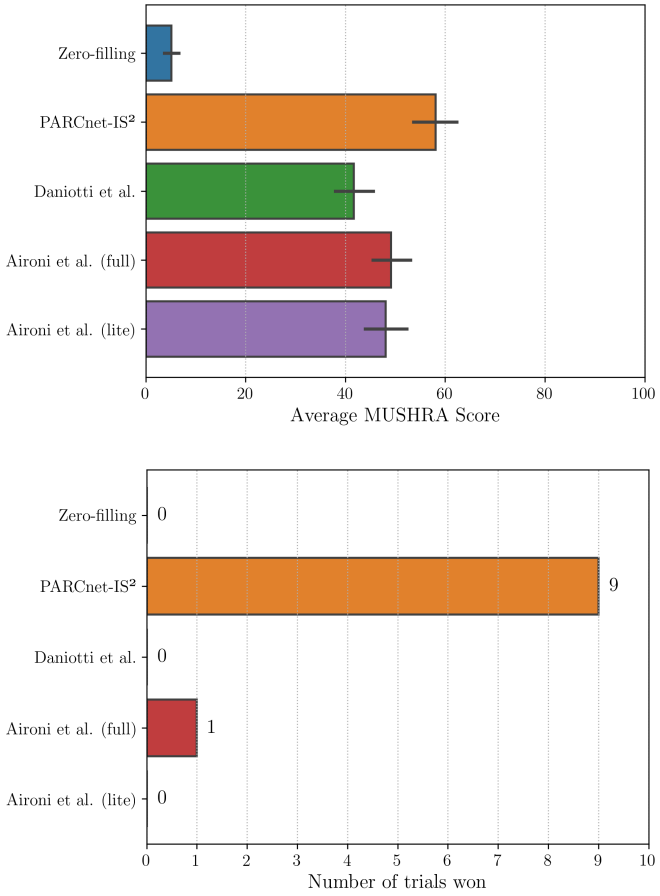


Fig. 1. Results of the MUSHRA test.

B. Subjective Evaluation

The Challenge ranking was determined through a MUSHRA test. To provide a fair test bench, ten signals from the blind test set were handpicked by an expert who only had access to the lossy audio files (zero-filling). In total, two violin excerpts, two cello excerpts, two clarinet excerpts, two double bass excerpts, and two guitar excerpts were selected. All clips are 11.6 seconds long and were presented in full to the assessors. The test was conducted using webMUSHRA [18], a state-of-the-art Web Audio API-based software compliant to the ITU-R Recommendation BS.1534 [19].

For each of the ten excerpts, the (undisclosed) clean audio file was used as Reference, whereas the clip degraded with zero-filling was considered as Anchor. After an initial training

TABLE III

OBJECTIVE METRICS COMPUTED ON THE ENTIRE BLIND SET. MEAN ± STANDARD DEVIATION; BOLD INDICATES THE BEST VALUE FOR EACH METRIC. ↑: HIGHER IS BETTER; ↓: LOWER IS BETTER.

	Time-domain			Spectral	Cepstral	Perceptual	
	MSE$\times 10^{-4}$ (↓)	SDR (↑)	SI-SDR (↑)	LSD (↓)	MCD (↓)	PEAQ (↑)	PLCMOS (↑)
Zero-filling	1.837 ± 3.370	11.87 ± 4.07	11.43 ± 4.41	0.599 ± 0.477	5.503 ± 4.710	-3.000 ± 0.866	1.717 ± 0.414
PARCnet-IS² (Baseline)	$\mathbf{0.645 \pm 1.096}$	$\mathbf{16.33 \pm 6.49}$	$\mathbf{17.95 \pm 6.45}$	$\mathbf{0.239 \pm 0.158}$	$\mathbf{2.456 \pm 2.062}$	$\mathbf{-1.832 \pm 1.052}$	$\mathbf{1.953 \pm 0.589}$
Daniotti et al.	8.774 ± 7.004	13.21 ± 8.51	13.34 ± 8.45	0.279 ± 0.211	3.883 ± 4.644	-2.198 ± 1.053	1.935 ± 0.496
Aironi et al. (full)	1.423 ± 2.409	13.32 ± 5.14	13.02 ± 5.40	0.290 ± 0.189	2.991 ± 2.388	-2.048 ± 1.031	1.849 ± 0.510
Aironi et al. (lite)	1.307 ± 2.189	13.04 ± 5.33	13.03 ± 5.35	0.294 ± 0.192	3.100 ± 2.476	-2.069 ± 1.031	1.857 ± 0.513

page where four audio examples were presented, i.e., two pairs of clean and zero-filling clips, participants were tasked to rate the similarity of each test condition with the Reference on a scale of 0 to 100. On each page, six conditions were assessed, including the output of PARCnet-IS², the hidden Reference, and the Anchor. The names of the test conditions were hidden, and the order of the ten trials, as well as the order of the test items within them, was randomized. Volume adjustments were only allowed during the training phase. Then, subjects were asked to keep the level constant for the duration of the test.

A pool of 12 expert assessors with age ranging from 24 to 45 (average: 29.75), none of whom reported hearing impairments, took part in the listening test. The participants, who completed the test in about 20 minutes, self-reported an average of 9.6 years of prior musical training (SD: 5.75). The assessor pool consisted of members of the Image and Sound Processing Lab (ISPL) at Politecnico di Milano, and had previous experience with MUSHRA tests.

V. RESULTS

Figure 1 shows the average scores and 95% confidence intervals obtained across all trials in the MUSHRA test (top) and the number of trials won by each PLC method (bottom). Table III reports the average objective metrics outlined in Section IV-A. Figure 2 depicts the box-and-whisker plots for each metric, whereas Figures 3 and 4 shows the box-and-whisker plots and the average scores of every trial in the listening test, respectively.

Table III indicates that the baseline method, on average, outperforms the submitted PLC systems across all objective metrics. These results appear to be confirmed by the outcome of the MUSHRA test in Figure 1, where PARCnet-IS² won 9 out of 10 trials and Aironi et al. (full) won one trial (Violin #2). This led to the final ranking given in Table I.

ACKNOWLEDGMENT

The organizers wish to thank Luca Turchet and the Organizing Committee of the 2nd IEEE International Workshop on Networked Immersive Audio.

REFERENCES

[1] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "INTERSPEECH 2022 audio deep packet loss concealment challenge," *arXiv preprint arXiv:2204.05222*, 2022.

[2] L. Diener, S. Branets, A. Saabas, and R. Cutler, "The ICASSP 2024 audio deep packet loss concealment grand challenge," *arXiv preprint arXiv:2402.16927*, 2024.

[3] C. Aironi, L. Gabrielli, S. Cornell, and S. Squartini, "Enhancing music packet loss concealment with generative spectrogram inpainting," Università Politecnica delle Marche and Carnegie Mellon University, Tech. Rep., 2024.

[4] F. Daniotti, L. Vignati, and L. Turchet, "Towards perceptual deep learning-based packet loss concealment," University of Trento, Tech. Rep., 2024.

[5] 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*, vol. 5, pp. 266–273, 2024.

[6] V. Lostanlen, C.-E. Cella, R. Bittner, and S. Essid, "Medley-solos-DB: a cross-collection dataset for musical instrument recognition," Zenodo, Sep. 29, 2019. doi: 10.5281/zenodo.3464194.

[7] D. Thery and B. F. Katz, "Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts," in *Proc. of the 23rd International Congress on Acoustics (ICA)*, 2019, pp. 739–746.

[8] C. Chafe, "Bach cello suites data repository," 2020. url: https://ccrma.stanford.edu/~cc/som/bachCello.

[9] C. Aironi, S. Cornell, L. Serafini, and S. Squartini, "A time-frequency generative adversarial based method for audio packet loss concealment," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 121–125.

[10] G. Bandiera, O. Romani Picas, H. Tokuda, W. Hariya, O. Koji, and X. Serra, "Good-sounds.org: A framework to explore goodness in instrumental sounds," in *Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 414–419.

[11] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, A. H. Cheng-Zhi, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," *International Conference on Learning Representations (ICLR)*, 2018.

[12] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[13] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

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

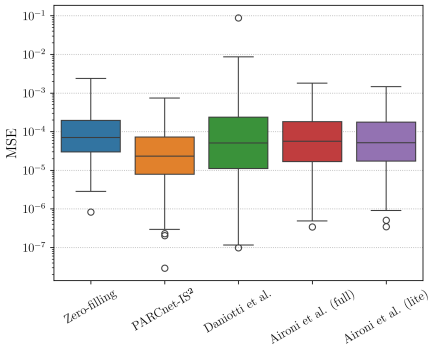
[15] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ–The ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[16] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," McGill, Tech. Rep., 2002.

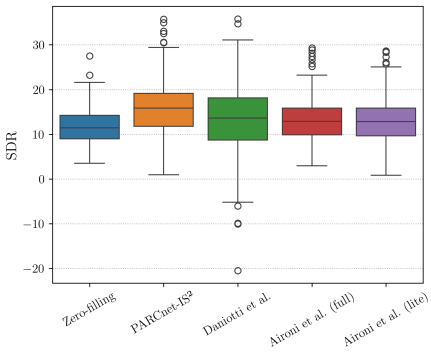
[17] L. Diener, M. Purin, S. Sootla, A. Saabas, R. Aichner, and R. Cutler, "PLCMOS–a data-driven non-intrusive metric for the evaluation of packet loss concealment algorithms," *arXiv preprint arXiv:2305.15127*, 2023.

[18] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, 2018.

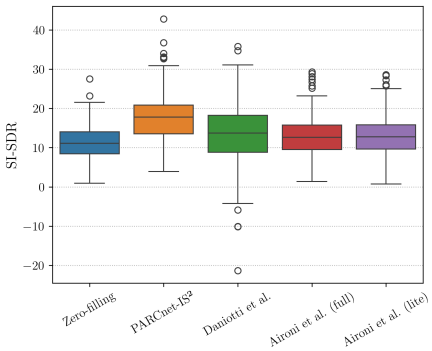
[19] "Method for the subjective assessment of intermediate quality level of audio systems," Rec. ITU-R BS.1534-3, International Telecommunications Union, Geneva, Switzerland, Jun. 2021.



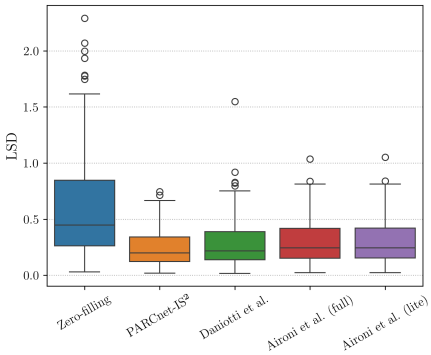
(a) MSE



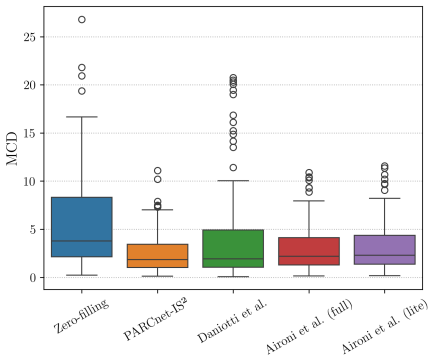
(b) SDR



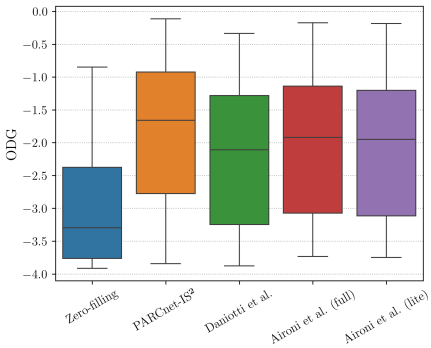
(c) SI-SDR



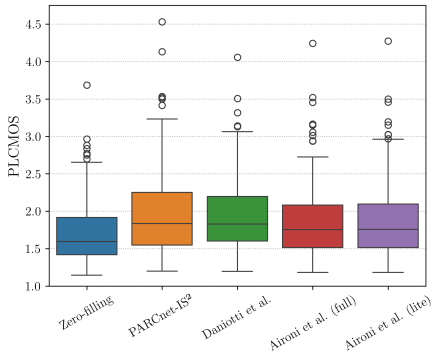
(d) LSD



(e) MCD

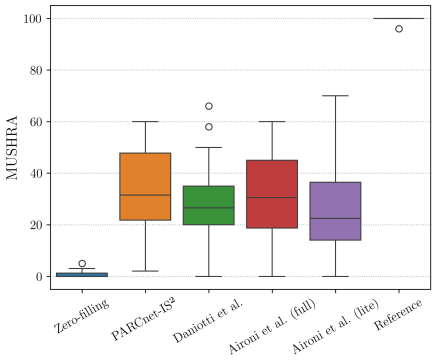


(f) PEAQ

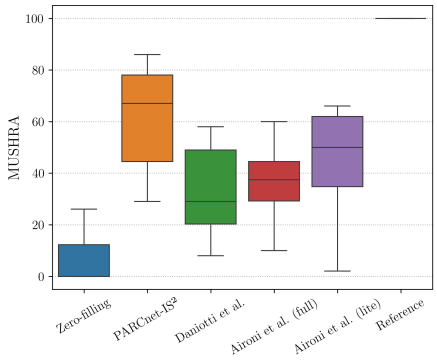


(g) PLCMOS

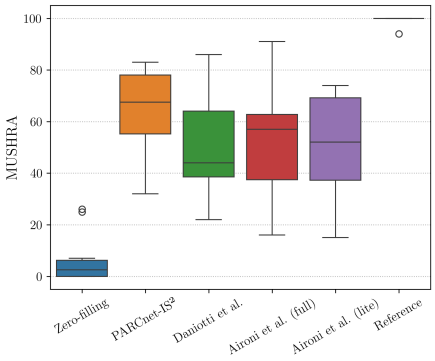
Fig. 2. Objective metrics computed on the entire blind test set.



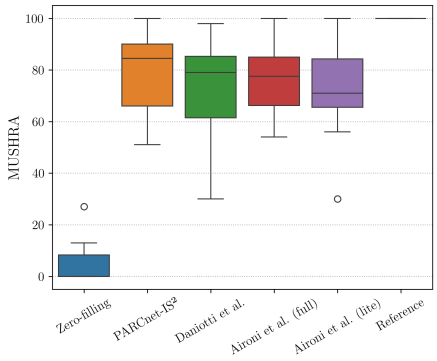
(a) Cello #1



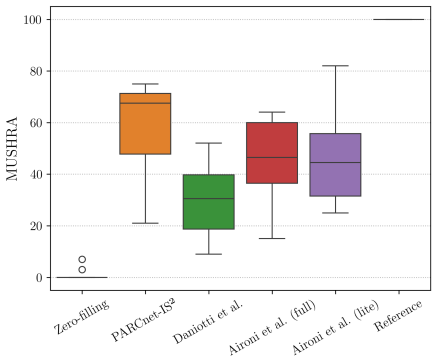
(b) Cello #2



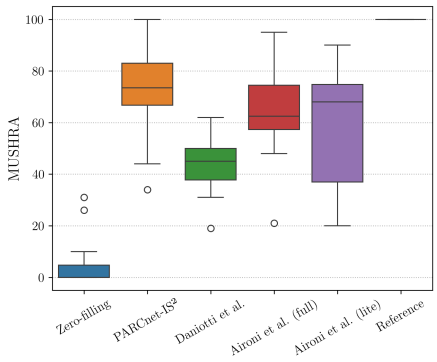
(c) Clarinet #1



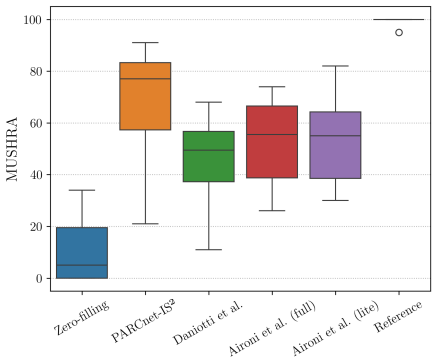
(d) Clarinet #2



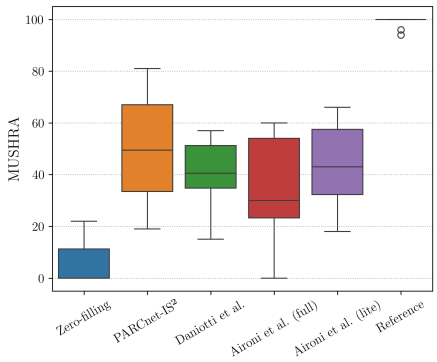
(e) Double Bass #1



(f) Double Bass #2

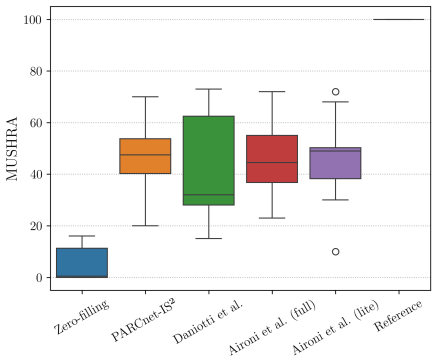


(g) Guitar #1

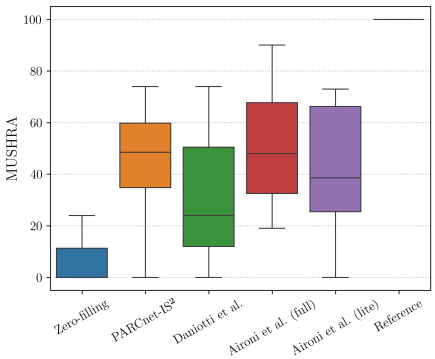


(h) Guitar #2

Fig. 3. Box-and-whisker plots of the individual trials in the MUSHRA test.

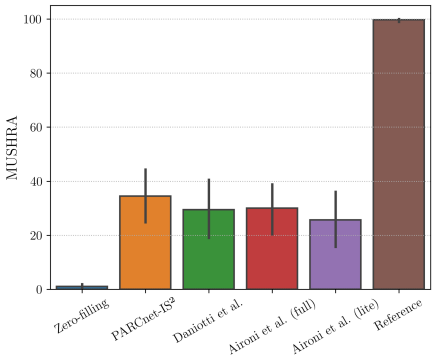


(i) Violin #1

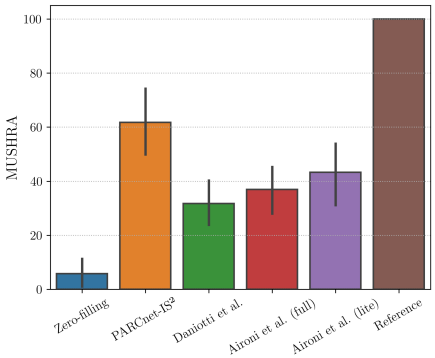


(j) Violin #2

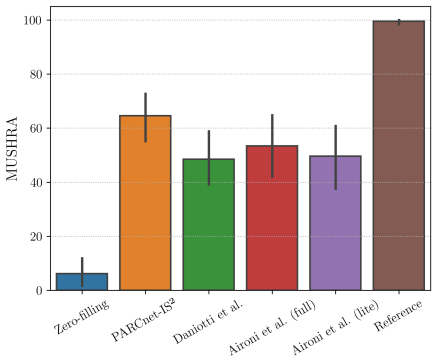
Fig. 3. Box-and-whisker plots of the individual trials in the MUSHRA test (cont.)



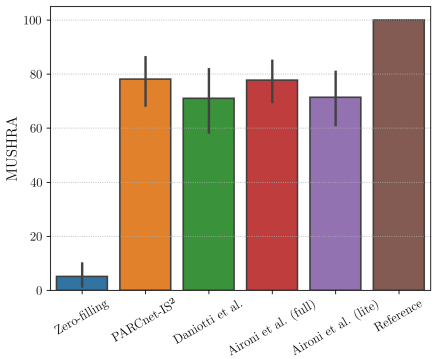
(a) Cello #1



(b) Cello #2

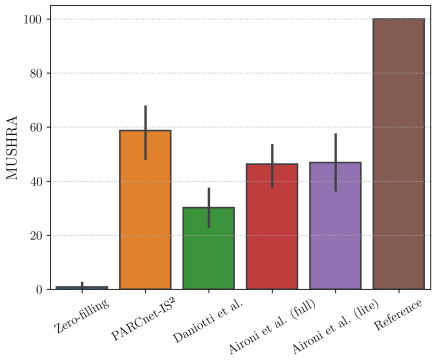


(c) Clarinet #1

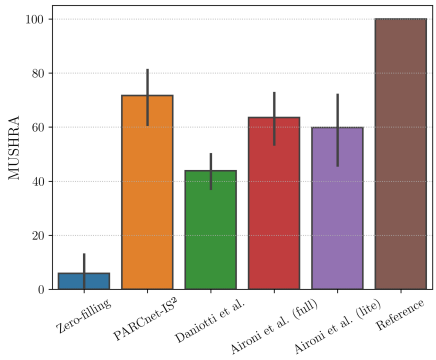


(d) Clarinet #2

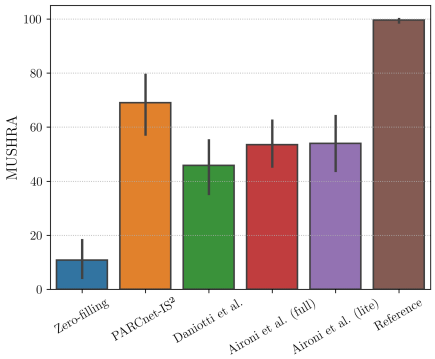
Fig. 4. Average scores and 95% confidence intervals of the individual trials in the MUSHRA test.



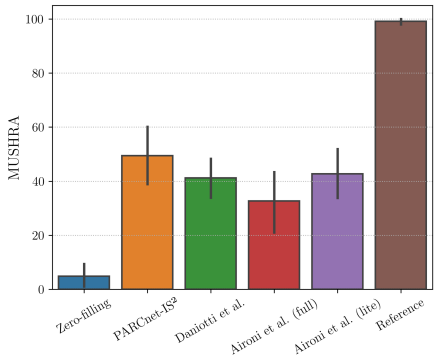
(e) Double Bass #1



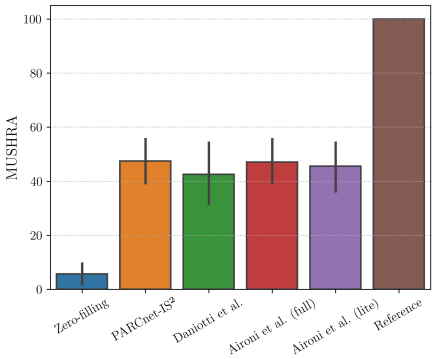
(f) Double Bass #2



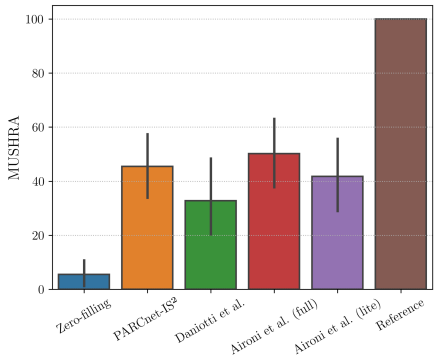
(g) Guitar #1



(h) Guitar #2



(i) Violin #1



(j) Violin #2

Fig. 4. Average scores and 95% confidence intervals of the individual trials in the MUSHRA test (cont.)