OPEN ACCESS Freely available online

CC BY

L. Vignati and L. Turchet, "On the Lack of a Perceptually Motivated Evaluation Metric for Packet Loss Concealment in Networked Music Performances," *J. Audio Eng. Soc.*, vol. 73, no. 10, pp. 660–670 (2025 Oct.).
https://doi.org/10.17743/jaes.2022.0227.

On the Lack of a Perceptually Motivated Evaluation Metric for Packet Loss Concealment in Networked Music Performances

LUCA VIGNATI, AND LUCA TURCHET,* *AES Member*
(luca.vignati@unitn.it) (luca.turchet@unitn.it)

Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Packet loss concealment (PLC) is vital in preserving audio quality for networked music performances. Although existing PLC techniques primarily target speech transmission, the unique challenges in music signals, such as complex harmonic structures and diverse timbral ranges, have yet to be adequately addressed. This is in part a result of the fact that a satisfactory objective evaluation metric for music PLC methods is missing. As a first foundational step toward this direction, this paper proposes a novel evaluation metric that leverages insights from music psychoacoustics and uses the constant-Q transform to better quantify glitch audibility induced by unconcealed packet loss (i.e., replaced with zeros) compared with existing metrics. The authors conducted extensive subjective listening tests leading to the creation of a publicly available ground truth data set, mapping objective audio features to human assessments of glitch audibility. Results show that the developed metric outperforms other measures (such as mean squared error and mean absolute error) in predicting perceptual impacts, taking a step toward addressing the need for a specialized metric for PLC in the domain of networked music performances. However, further improvements are needed to match human perceptual accuracy, which calls for further research on the development of a reliable perceptually motivated evaluation metric.

0 INTRODUCTION

Networked music performance (NMP) systems, such as JackTrip [1] or Elk Live [2], aim to enable geographically displaced musicians to play together over the network [3]. Packet loss concealment (PLC) is a critical technique for maintaining audio quality in NMP applications, where packet loss is inevitable because of fluctuations of the network conditions. However, NMPs exhibit distinct missing frame mechanisms compared with traditional packet loss in common applications. Indeed, whereas conventional loss arises from network delivery failures or time to live timeouts, NMP systems primarily discard packets arriving beyond strict 20–30-ms playback deadlines that are necessary for musicians to preserve their ability to synchronize. Additionally, missing audio frames can be a result of clock drift between endpoints with unsynchronized clocks that induce periodic buffer underruns/overruns even on lossless networks (e.g., the receiving endpoint reads packets faster than what the transmitting endpoint delivers). This requires the introduction of mechanisms for maintaining synchronization, which are often addressed by adopting external GPS-derived clocking solutions [4].

Forward error correction and other redundancy-based methods represent some of the most widely adopted techniques in NMP applications to deal with the issue of lost audio frames. Rather than attempting to reconstruct lost audio based on analysis of surrounding frames, these approaches preemptively send redundant information to enable recovery when packets are lost. When sufficient network capacity exists, redundancy-based methods offer the advantage of perfect reconstruction of lost packets without introducing the artifacts common to prediction-based approaches [5]. However, the usage of these schemes entails a larger bandwidth availability in the utilized network link, which is not always possible.

The Opus codec [6] represents a significant development in audio transmission over the network. Its widespread adoption in networked audio applications is a result of its configurable frame sizes (as small as 2.5 ms) and minimal additional latency (approximately 5 ms). The recent advances in Opus (version 5.1, March 2024) have introduced generative PLC capabilities alongside traditional

*To whom correspondence should be addressed, email: luca.turchet@unitn.it

redundancy-based approaches, further enhancing its suitability for maintaining audio quality during packet loss events in musical performances [7]. Despite these advances, the evaluation of Opus's PLC mechanisms specifically for musical content remains an area requiring further investigation. However, the introduction of codecs adds to the overall end-to-end latency because of the times for encoding and decoding. In NMP systems, latency is typically prioritized over reliability; often lossy transmission protocols, such as user datagram protocol, are utilized, and codecs are avoided.

Signal distortion is incurred as the missing audio frames hit playback at the receiver's side, typically manifesting as chopping, popping and clicking. PLC algorithms aim to mask the audible effects of missing audio packets by reconstructing the lost signal in real-time at the receiver. Although PLC has been extensively studied for voice-over–IP applications [8], its use in NMPs presents unique challenges that drastically differ from those encountered in speech PLC. Such challenges are a result of the complex harmonic structures, polyphony, diverse timbres, and wide frequency range of musical signals.

Existing PLC algorithms for music can be broadly categorized into two groups: waveform substitution methods and model-based approaches. Waveform substitution methods reconstruct missing segments by repeating or interpolating portions of previously received audio, often leveraging musical properties, such as pitch and harmonic structure, to improve perceptual quality. Model-based approaches, on the other hand, employ parametric models to interpolate and extrapolate missing segments, with some advanced techniques using machine learning to capture the temporal and spectral characteristics of musical signals [9, 10].

A significant milestone in PLC research for NMPs was reached with the first edition of the IEEE-IS^2 2024 Music Packet Loss Concealment Challenge, held as part of the 5th IEEE International Symposium on the Internet of Sounds in Erlangen, Germany [11]. This challenge aimed to promote research on PLC specifically for NMP applications. Notably, owing to the lack of an accurate metric to objectively evaluate PLC algorithms for music, the organizers of the challenge had to rely on human subjects participating in listening tests to determine the winners. Nevertheless, listening tests are an expensive and time-consuming process that in the case of music PLC requires expert listeners (typically musicians). This underscores the critical need for a reliable, perceptually motivated objective metric specifically designed to automate the evaluation of PLC methods in musical contexts.

As a first foundational step toward addressing the lack of an evaluation metric specifically designed for PLC algorithms in NMPs, the present study focuses on measuring the audibility of unconcealed packet loss. By focusing on scenarios where missing packets are simply replaced with zeros (which produces the most obvious artifacts), the authors establish a baseline for audibility assessment that can later be extended to evaluate actual PLC methods that produce more subtle artifacts. The proposed metric leverages insights from music psychoacoustics to quantify the audibility of packet loss–induced artifacts as perceived by musicians and listeners in NMP contexts. Through a set of listening tests, this study establishes a ground truth data set that maps objective audio features to subjective assessments of glitch audibility in musical signals. The proposed metric is evaluated against existing metrics and demonstrates its superior performance in predicting the perceptual impact of packet loss on musical signals. Although this study does not directly address the PLC quality rating, given the fact that it focuses on the audibility of the audio dropout signal distortion, it discusses how the proposed metric can be used to guide the development of new PLC algorithms specific to NMPs.

1 RELATED WORK

1.1 Packet Loss Concealment in Audio Transmission

Early work on audio-based PLC focused primarily on speech transmission. Lin et al. proposed a hybrid PLC algorithm for MDCT-based audio codecs, which employed different strategies for tone-dominant and noise-dominant audio segments [12]. This approach demonstrated the importance of adapting PLC techniques to the specific characteristics of the audio content. In the realm of music transmission, Sacchetto et al. investigated the use of autoregressive models for real-time PLC in NMP applications [13]. Their low-complexity prediction method showed promise in filling audio gaps caused by missing packets, highlighting the potential of statistical models in PLC for music.

The advent of deep learning has opened new avenues for music PLC research. Mezza et al. introduced a hybrid PLC method for real-time networked music applications, combining traditional signal processing techniques with neural networks [9]. Their approach, named PARCnet, utilizes a feedforward neural network to estimate the time-domain residual signal of a parallel linear predictor, demonstrating improved performance over traditional methods.

1.2 Audio Quality Ratings

The challenges in the field of audio PLC stem from the need to maintain high audio quality while dealing with the unpredictable nature of packet loss in network transmissions. Various methods have been proposed to evaluate the audio quality, including in NMP contexts [5, 14].

The International Telecommunication Union (ITU) recommendation ITU-R BS.1534-3 Multistimulus test with hidden reference and anchor (MUSHRA) is a subjective evaluation method encompassing a scale ranging between 0 (bad quality) and 100 (excellent quality), designed to assess intermediate audio quality levels (i.e., not perfect but also not severely degraded) [15]. It is commonly used to test technologies such as audio codecs, speech codecs for telecommunications, and playback systems such as headphones.

Perceptual evaluation of audio quality (PEAQ), defined by the ITU in the recommendation ITU-R BS.1387, models

perceptual characteristics of human hearing and combines multiple psychoacoustic model outputs into a single metric known as the objective difference grade. The objective difference grade score ranges from 0 (no perceptible difference from reference) to −4 (very annoying degradation) [16]. PEAQ was originally developed to assess the impairment introduced by lossy audio codec.

Recent research has shown that basic audio quality ratings (as per ITU-R BS.1534-3) and annoyance ratings (as per ITU-R BS.1387) may differ [17]. Other researchers have questioned the reliability of PEAQ for evaluating audio quality in streaming applications over lossy networks, noting that it may not accurately capture the perceptual effects of packet loss [18].

1.3 Audio Quality Prediction Systems

Signal quality rating prediction systems aim to estimate perceived audio quality using a combination of signal analysis, perceptual modeling, and machine learning. These systems typically follow a common architecture (see Fig. 1 in [16] for a reference). The approaches include:

- Time-frequency signal representation: Audio inputs are transformed into time-frequency representations (e.g., via filter banks or fast Fourier transform), emulating the resolution characteristics of the human auditory system;
- Human hearing modeling: Psychoacoustic models are applied to simulate perceptual phenomena such as masking, loudness, and temporal integration;
- Neural network–based rating estimation: Increasingly, neural networks are trained on subjective quality data (e.g., BS.1387 or MUSHRA scores) to learn mappings from perceptual features to quality ratings.

Recent systems have begun to incorporate neural networks trained explicitly in music quality ratings, demonstrating improved sensitivity to the perceptual impact of transmission-induced artifacts. A notable example is the study by Organisciak et al. [19], which proposes an architecture trained directly on subjective ratings of music affected by PLC. This approach has shown promising results, outperforming traditional models on data sets with packet loss and concealment artifacts. In general, state-of-the-art methods have been applied to the cases of 1) mapping from codec artifacts to audio quality rating and 2) mapping from PLC artifacts to audio quality rating. However, such methods have not been applied to the NMP-specific case presented in this paper, namely the mapping from audio dropout artifacts to artifact audibility rating.

1.4 Evaluation Metrics for Music Packet Loss Concealment

Objective metrics commonly used in the literature [9, 13] for the automatic assessment of music PLC algorithms, such as mean squared error (MSE) and mean absolute error (MAE), may fail to capture the perceptual impact of packet loss on music quality as perceived by human listeners. Even metrics developed for evaluating broadband audio quality, such as PEAQ [16], may not accurately reflect the perceived quality of reconstructed musical signals [18].

Audio artifacts from missing frames in NMP typically manifest as chopping, popping and clicking, essentially representing broadband spectral flux. This suggests that spectral analysis approaches may be particularly suitable for their detection and measurement.

1.5 Challenges in Networked Music Performances

The specific requirements of NMP applications present unique challenges for PLC algorithms. Turchet et al. explored the potential of 5G-enabled Internet of Musical Things architectures for remote immersive musical practices, emphasizing the need for low-latency, high-quality audio transmission [20]. However, it should be noted that achieving the sub–30 ms latency required for real-time musician interaction using 5G remains largely theoretical in practical network deployments. Recent empirical studies have documented significant challenges in achieving such low latencies over public 5G infrastructures, with actual implementations often exceeding the acceptable thresholds for synchronous musical performance [21, 22]. These studies underscore the importance of developing PLC techniques that can operate within the strict time constraints of real-time music performance.

Dürre et al. highlighted the critical need for specialized metrics to evaluate network-based packet loss in NMP scenarios [21]. They observed that conventional metrics, such as PEAQ, may be inadequate for assessing the impact of packet loss on audio quality in NMPs because such metrics were not designed to reflect the specific characteristics of transmission systems. Their work emphasizes that the temporal distribution of packet loss (not merely statistical percentages) significantly affects perceived quality. In a different vein, although redundancy-based methods such as forward error correction offer excellent audio quality, they require additional bandwidth, creating a tradeoff between audio fidelity and network resource utilization that must be carefully managed in NMP contexts.

As the field of NMP continues to evolve, there is a growing need for PLC techniques that can adapt to diverse musical genres, instrument types, and network conditions. Turchet and Krstulović's work on "DSP as a Service" points to future directions where cloud-based processing could enhance PLC capabilities for NMPs [23]. This approach could potentially leverage more complex models and larger data sets to improve PLC performance while maintaining low latency. Although promising for various applications, it is important to clarify that server-based digital signal processing for low-latency NMP currently faces significant challenges in practical implementations. As demonstrated by Carôt et al., in current network environments, these technologies primarily serve as audience distribution solutions rather than enabling real-time musician interactions that require stricter latency constraints [24].

In conclusion, although significant progress has been made in PLC for audio transmission, there remains a need for specialized techniques that address the unique challenges of NMP. Thus, future research should focus on developing adaptive, low-latency PLC algorithms that can maintain high audio quality across a wide range of musical content and network conditions.

2 METHODOLOGY

The methodology of the present study was based on the following steps:

1. **Defining the ground truth**: The authors began by establishing the ground truth, which involves specifying a target and the metrics to be measured. The long-term goal is to accurately measure any glitches that occur in audio signals. However, because of constraints in time and resources, it was necessary to narrow the scope of this study to manageable yet relevant aspects. This study chooses to limit its focus on specific types of glitches by imposing constraints on packet loss distribution and still incorporating various audio types without limiting the locations of the glitches. The aim is to assess the audibility of these glitches as perceived by human listeners.
2. **Quantitative measurement**: With the ground truth defined, quantitative measures are implemented to evaluate the audibility of audio glitches as perceived by humans. To this end, listening tests are conducted in which participants provide their feedback and insights.
3. **Evaluation of existing metrics**: Once the ground truth data are acquired, existing metrics are evaluated (typically employed in music PLC research [9, 13]) by applying them to the same audio clips that participants rated during the listening tests. The outcomes of these evaluations are subsequently compared against the ground truth to derive scores for these metrics.
4. **Algorithm development**: A new algorithm is proposed based on insights gathered from the authors' expertise in the PLC field. This algorithm undergoes an iterative development cycle where it is tested against the ground truth, refined, and retested to maximize its performance.

2.1 Ground Truth

For this initial study, the aim is to establish a simple and well-defined ground truth. This study has chosen to use short audio clips of 3 s, which are affected by a single packet loss of fixed size. The location of the packet loss in a given audio file is randomized. The audio clip is then extracted from the audio file. Specifically, the clip is centered around the packet loss to ensure that the relative position of the packet loss within the audio clip remains consistent.

Two audio recordings were selected that feature instruments with radically different timbres: an electric guitar and a cello. For each recording, using the PLC Test Bench[1] software reported in [25], a lost packet distribution was generated based on the following criteria: 1) two consecutive packets are lost every 1.5 s, and 2) the packet size is set to 64 samples at a sampling rate of 44.1 kHz.

It is important to note that the choice of two consecutive packets lost represents a simplified scenario compared with the more challenging burst packet loss conditions often encountered in real-world NMPs. Although redundancy-based PLC methods can typically handle two consecutive missing packets effectively, longer bursts of packet loss, where redundancy mechanisms are more likely to fail, remain outside the scope of this initial study. This simplified approach was selected to establish a controlled environment for developing and validating this study's perceptually motivated metric, serving as a foundation for future extensions to more complex packet loss patterns.

To maximize the audibility of the glitches and enhance the effectiveness of this initial study, the authors chose to leave packet losses unconcealed by simply replacing missing packets with zeros without any crossfade or transition. Although this approach does not constitute actual PLC, it provides a clear baseline case with maximally audible artifacts, serving as a necessary first step before evaluating more sophisticated concealment methods that produce more subtle artifacts.

The selection of electric guitar and cello represents a controlled comparison of contrasting timbral profiles (bright versus dark, sustained versus transient rich) rather than comprehensive musical coverage. This focused approach enables clear demonstration of metric improvements over existing methods and establishes methodology for future expansion to broader instrumentation.

2.2 Listening Tests

To address the challenge of algorithmically measuring glitch audibility, a listening test was designed to extract audibility scores from human listeners. Twelve expert musicians (all males, aged between 25 and 41 y, mean age $= 30.3$ y, SD $= 5.2$) took part in the experiment. These musicians had at least 15 years of musical experience. All musicians used a pair of Beyerdynamic DT 770 PRO 80-ohm headphones.

Using pyMUSHRA and webMUSHRA [26, 27], a custom listening test was created by modifying the MUSHRA template. Notably, this test is not a standard MUSHRA test but rather a specialized data set–labeling task. For each instrument, two separate listening tests were conducted to increase the number of rated audio clips and reduce fatigue (thus enhancing results reliability) by having subjects take the tests on two different days. Specifically, the test was structured as follows:

1. The test begins with a page featuring hand-picked audio clips as references.

[1] https://github.com/CIMIL/plc-testbench

2. Subjects then listen to batches of 12 audio clips and provide audibility scores for each.
3. The rating scale is continuous from 0 to 100 and divided into five regions: "Not at all," "Barely," "A bit," "Much," and "Very much," answering the question "How much can you hear the glitch?"
4. Two hand-picked clips serve as hidden reference and anchor to ensure full utilization of the rating scale in each batch of audio clips.
5. Two pages of 12 audio clips are presented, followed by a 5-min break, then repeated.
6. The test is blind, with shuffled audio clips within pages and randomized page order.

The choice of a continuous rating scale from 0 to 100 with descriptive anchors aligns with methodologies validated in previous NMP research by Tsioutas et al. [14], who demonstrated that similar rating scales effectively capture musicians' perceptions of audio artifacts. This approach allows subjects to precisely quantify their perception of glitch audibility while maintaining a clear conceptual framework through the labeled regions.

2.3 Metrics Evaluation

In the field of PLC applied to NMPs, evaluation metrics are algorithms that quantify the difference between an original signal and its reconstructed version. These metrics take both signals as input and compute a score representing their dissimilarity. This score is computed as a distance between the two signals, typically just a sample-wise difference. The distance used for the proposed metric is the sum of the differences between the bins of the time-frequency representations of the two signals.

The following metrics commonly used in PLC for NMPs were evaluated:

- **MSE** measures the average squared difference between the original and reconstructed signals in the time domain.
- **MAE** calculates the average absolute difference between the original and reconstructed signals in the time domain.
- **PEAQ** was originally developed for evaluating audio codec quality and has become the de facto standard for assessing PLC algorithms in the literature, despite debates about its accuracy in this context [18]. GstPEAQ, an open-source implementation of the PEAQ algorithm, was employed as a GSTREAMER plug-in [28], testing both "basic" and "advanced" operation modes.

Metric selection followed established practices in NMP-focused PLC research, prioritizing measures validated for music reconstruction tasks. MSE and MAE provide granular error quantification across time-frequency domains, whereas PEAQ aligns with ITU standards for perceptual quality assessment under impairment conditions analogous to packet loss. Alternative psychoacoustic metrics were excluded because of 1) focus on timbral attributes (sharpness, roughness) rather than packet loss artifacts, 2) optimization for speech codecs rather than music PLC, and 3) insufficient validation in NMP contexts.

2.3.1 Quantifying the Distance

To validate the accuracy of each metric, their results were compared with the established ground truth. Each metric was applied to all audio clips used in the listening tests and the resulting scores were normalized to match the ground truth range (0–100). This normalization allows for direct comparison between metric scores and ground truth values. The following methods were employed to quantify the distance between metric predictions and ground truth:

- **MAE-D** calculates the average of absolute differences between predicted and observed values, giving equal weight to all individual differences.
- **MSE-D** computes the average of squared differences between predicted and observed values, also weighting all individual differences equally.
- **Spearman's rank correlation coefficient (SRCC)-D** is a nonparametric measure that assesses the strength and direction of association between two ranked variables. It evaluates how well the relationship between two variables can be described using a monotonic function.

2.4 Algorithm Development

The methodology described thus far facilitates an iterative approach to developing a new metric. With the ground truth data available, new ideas can be swiftly implemented and tested against existing metrics using Testbench software [25]. The study's development process followed these key steps:

1. **Empirical observations**: A multimodal analyses of audio clips was conducted, including auditory evaluation and examination of time-frequency representations. The availability of ground truth values compared with metric values allowed for investigation of why certain glitches were correctly reported by the metric and others were not.
2. **Intuition**: Educated guesses were made about the metric's performance, and potential improvements were made based on the observations.
3. **Research**: Relevant literature was explored to provide theoretical grounding for the intuitions, transforming them into actionable hypotheses.
4. **Implementation**: These hypotheses were implemented in the algorithm.
5. **Testing**: The new algorithm was evaluated against the ground truth data and other metrics, quantifying the distance as explained in SEC. 2.3.1.

This iterative process continued until satisfactory results were achieved or other exit criteria were met. Notable iterations focused on the following areas.

2.4.1 Selection of Time-Frequency Representation

Early in the research for this work, it became evident that waveform-only metrics were inadequate for the assessment of audibility of raw audio dropouts (and in general for PLC evaluation). Therefore, various time-frequency representations were explored:

- **Short-time Fourier transform** is a method for analyzing frequency content over time, implemented using librosa [29].
- **Constant-Q transform (CQT)** is a variation of the Fourier transform with logarithmically spaced frequency bins, implemented using the Essentia Python module [30].
- **Dynamic compressive gammachirp filter** is an auditory filterbank simulating cochlear frequency analysis and compression, implemented using the brian2Hears Python module [31].

Contrary to initial expectations regarding dynamic compressive gammachirp filter, the CQT emerged as the best approximation of human perception, and all subsequent analyses were conducted using this representation.

Specifically, for CQT a single-value score was derived from the time-frequency representation across the sample by summing all the residuals that are the result of the spectral difference. A normalization of the single values was then applied to make them fit into the 0–100 range of the audibility rating scores. The metric's implementation is available online.[1]

2.4.2 Testing the Masking Hypothesis

Time-frequency masking is a complex psychoacoustic phenomenon according to which low-energy signals become inaudible when higher-energy signals are present nearby in the time-frequency plane. Why this happens in the ear has a lot to do with how the basilar membrane works (for details see [32]).

The empirical observations found led to the formulation of a *masking hypothesis*, suggesting that portions of the glitch might be masked by the surrounding signal. This phenomenon was investigated using data from Necciari et al. [33], who measured simultaneous time and frequency-based masking using compact stimuli in the time-frequency domain. A version of the metric that computed masking thresholds in the time-frequency region surrounding the glitch was implemented, and these were used to isolate unmasked portions before quantifying audibility. However, test results indicated that this approach did not yield significant improvements.

2.4.3 Fine Tuning Constant-Q Transform Parameters

The iterative methodology proved highly effective in optimizing the CQT parameters, as described in Table 1. A grid search was performed on individual parameters, starting from Essentia's default values. For each iteration, the value that minimized the distance between the authors' metric and the ground truth was selected. The final optimized parameters are presented in SEC. 3.2.

Table 1. Parameters of the CQT.

Parameter	Description
Intorno length	Size in seconds of the audio clip centered around the packet loss
Linear/decibel magnitude	Choice between working with linear amplitude or converting it to decibels
Bins per octave	Number of frequency bins per octave
Minimum frequency	The minimum frequency considered
Minimum window	The minimum size allowed for the windows

3 RESULTS

3.1 Ground Truth Data

To evaluate the reliability and consistency of subjects' ratings, the intraclass correlation coefficient (ICC) was calculated. The ICC measures inter-rater agreement on a scale from 0 to 1, with higher values indicating stronger agreement. Table 2 presents the overall ICC values for the listening test data, including the ICC for both single and average raters, considering absolute agreement and treating raters as random or fixed effects. The ICC(1) value of 0.745 indicates substantial agreement for single raters. For average ratings, the ICC(1,k) reaches 0.972, demonstrating excellent agreement. The high ICC values and narrow 95% confidence intervals suggest that subjects provided consistent and reliable ratings throughout the listening tests.

3.2 Optimization of Constant-Q Transform Parameters

As introduced in SEC. 2.4.3, a comprehensive grid search was conducted to determine the optimal parameters for the CQT in the context of the glitch audibility metric. This optimization process aimed to find the parameter set that best aligns the CQT-based metric with human perception of audio glitches. Table 3 presents the optimal parameters resulting from this search.

These parameters were determined through an iterative process, evaluating the performance of the CQT-based metric against the ground truth data for each parameter combination. The final set represents the configuration that achieved the closest alignment between the metric's output and human-perceived glitch audibility. Beyond their empirical performance, each parameter has psychoacoustic relevance:

- **Intorno length (IL = 3,000 ms)**: This value represents the optimal temporal context needed to evaluate glitch audibility. Beyond 3,000 ms, diminishing returns in performance improvement were observed and computational complexity continued to increase.

Table 2. ICC values.

Type	Description	ICC ↑	95% Confidence Interval
ICC(1)	Single raters, absolute	0.745	(0.68, 0.81)
ICC(2)	Single random raters	0.746	(0.68, 0.81)
ICC(3)	Single fixed raters	0.770	(0.71, 0.83)
ICC(1,k)	Average raters, absolute	0.972	(0.96, 0.98)
ICC(2,k)	Average random raters	0.972	(0.96, 0.98)
ICC(3,k)	Average fixed raters	0.976	(0.97, 0.98)

Table 3. Optimal parameters of the CQT.

Parameter	Optimal Value
Intorno length	3,000
Linear/decibel magnitude	decibels
Bins per octave	48
Minimum frequency	10 Hz
Minimum window	2,048

Note. The intorno length (analysis window length around the lost packet) is measured in milliseconds. The minimum window is measured in samples.

Table 4. Comparison of metric accuracy

Metric	MAE-D	MSE-D	SRCC-D
MSE	32.24	1,479.96	0.37
MAE	24.06	929.06	0.36
PEAQ (basic)	21.85	760.1	0.7
PEAQ (advanced)	33.79	1,650.23	0.16
Perceptual	**13.7**	**283.05**	**0.79**

Note. Lower absolute error values (MAE-D and MSE-D) and higher correlation values (SRCC-D) correspond to better perceptual alignment. Bold values indicate the best results achieved by the proposed metric.

- **Magnitude representation (dB)**: The logarithmic nature of decibel scaling aligns with human auditory perception, which naturally processes intensity on a logarithmic scale. This enables the metric to weight intensity differences in a perceptually meaningful way.
- **Bins per octave (BPO = 48)**: Human listeners can detect frequency differences as small as 0.2–1% in the midfrequency range. With 48 bins per octave, each bin represents approximately a 1.5% frequency change, adequately capturing most perceivable frequency differences and maintaining computational efficiency.
- **Minimum frequency (MF = 10 Hz)**: By placing the minimum frequency at 10 Hz, just below the average human hearing threshold (approximately 20 Hz), it was ensured that no audible portion of the signal is discarded from analysis.
- **Minimum window size (MW = 2,048 samples)**: This parameter governs the crucial trade-off between temporal resolution (ability to locate events in time) and frequency resolution (ability to distinguish between close frequencies). The chosen value provides sufficient temporal resolution to detect brief glitches while maintaining adequate frequency discrimination.

It is important to note that, although these parameters performed optimally in this current experimental setup, the preliminary nature of this study (with limited diversity in participants, instruments, and PLC techniques) means they should not be considered definitive. Future work with more extensive and diverse data sets may reveal different optimal values or demonstrate varying sensitivities to parameter adjustments. A comprehensive robustness analysis of parameter sensitivity would be valuable in future studies with larger, more diverse data sets.

3.3 Metrics Comparison

The present study compared several metrics for evaluating glitch audibility in audio signals. Fig. 1 illustrates the absolute error between each metric and the ground truth sorted in ascending order. The perceptual metric consistently demonstrates smaller or equal errors compared with other metrics. Notably, the maximum absolute error for the perceptual metric is about 42, whereas MSE, MAE, and PEAQ "advanced" exhibit worst-case absolute errors of about 80, and PEAQ "basic" shows a maximum error of about 70. Table 4 provides a quantitative comparison of the metrics using MAE-D, MSE-D, and SRCC-D.

Additionally, the Friedman test and, subsequently, pairwise comparisons using the Wilcoxon signed-rank test were performed on the data series shown in Fig. 1 (see Fig. 2). Such nonparametric tests were chosen because the error distributions were non-normal.

There was a statistically significant difference in the absolute error depending on the metric ($\chi^2(4) = 73.2$; $p < 0.001$). Wilcoxon signed-rank tests with Bonferroni correction quantified pairwise differences. Effect sizes were calculated using matched pairs rank-biserial correlation (r), where values ≥ 0.5 indicate large practical significance. The Wilcoxon signed-rank test indicated that metric used here exhibited significantly lower absolute errors than PEAQ basic ($W = 714$, $p < 0.001$, $r = 0.62$), MAE ($W = 796$, $p < 0.001$, $r = 0.58$), MSE ($W = 347$, $p < 0.001$, $r = 0.71$), and PEAQ advanced ($W = 248$, $p < 0.001$, $r = 0.75$).

Key observations from the authors' analysis include the following:

1. **MSE and MAE**: As anticipated, these traditional metrics perform poorly across all evaluation criteria, indicating their limited ability to capture perceptual aspects of glitch audibility;

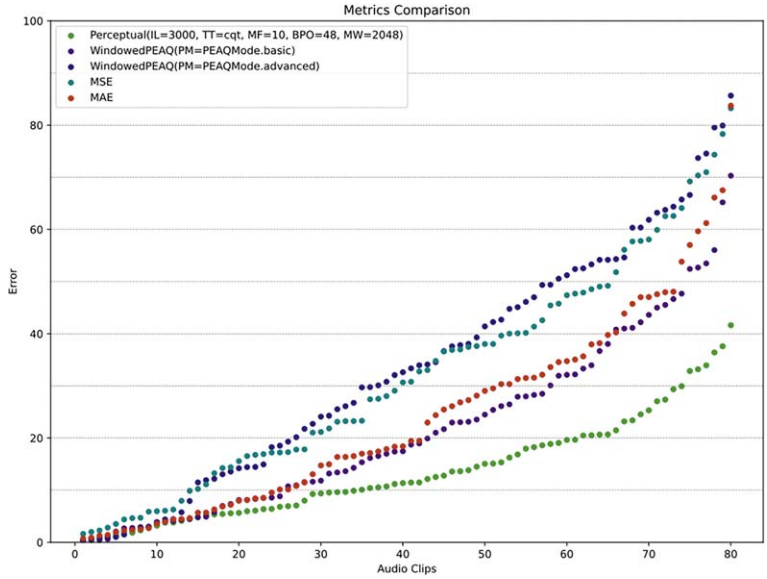

Fig. 1. The absolute error between each metric and the ground truth for each audio clip, ordered by absolute error.

2. **PEAQ (basic)**: This metric achieves a high SRCC value (0.7) despite relatively high MAE and MSE values. This suggests that PEAQ (basic) captures a nonlinear relationship with the ground truth that is reflected in the SRCC but not in the linear error measures;
3. **PEAQ (advanced)**: Surprisingly, this version of PEAQ performs the worst across all evaluation criteria, with a particularly low SRCC of 0.17. This unexpected result warrants further investigation into the suitability of PEAQ (advanced) for glitch audibility assessment. Although the reason why this happens is beyond the scope of this study, the fact that PEAQ (advanced) outperforms PEAQ (basic) outlines how unstable and unreliable PEAQ is when used to evaluate PLC artifacts, as already stated in [18];
4. **Perceptual Metric**: This study's proposed metric outperforms all the others in all three evaluation criteria. It demonstrates superior performance in both linear (MAE-D, MSE-D) and nonlinear (SRCC-D) relationships with the ground truth.

The proposed perceptual metric achieves statistically significant improvements over all other metrics, with an MAE-D of 13.7, an MSE-D of 283.05, and an SRCC-D of 0.79. These results underscore the effectiveness of this study's approach in accurately quantifying glitch audibility.

4 DISCUSSION

In this paper, the authors have presented a novel perceptually motivated metric for evaluating the audibility of glitches caused by unconcealed packet loss in musical signals transmitted over networks. Although specifically focused on the case where missing packets are replaced with zeros (representing no concealment) as a first methodological step, the insights gained and metric developed provide a foundation for future work in evaluating actual PLC methods. The artifacts produced by sophisticated PLC al-

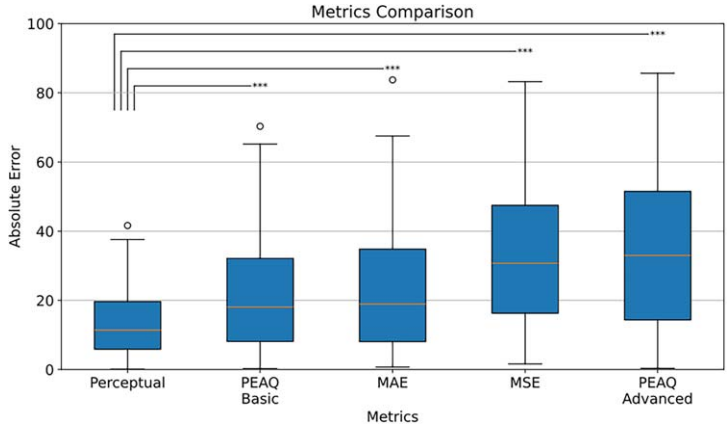

Fig. 2. Comparison between metrics. The orange line represents the 50th percentile, box edges indicate the 25th and 75th percentiles, and whiskers show the range of ratings. Outliers are marked with circles. Significance is only shown between each metric and the perceptual metric. The asterisks (***) indicate $p < 0.001$.

gorithms would be similar in nature to those studied here, albeit more subtle and therefore more challenging to detect and measure.

The results reported in SEC. 3.3 demonstrate that the CQT-based metric outperforms existing metrics in terms of alignment with the perceptual ground truth established through the listening tests. This metric achieves a significantly lower MAE and MSE and a higher SRCC, indicating its superior ability to predict the subjective audibility of glitches. The conducted comparative analysis highlights the limitations of traditional metrics and existing perceptual models in assessing glitch audibility. It also demonstrates the potential of the proposed perceptual metric to more accurately reflect human perception of audio glitches in various contexts.

The development of this metric has strong implications for the NMP field. It provides the basis for developing improved tools for assessing the perceptual impact of packet loss on musical signals compared with existing metrics, potentially aiding in the development and optimization of PLC algorithms that prioritize perceptual quality. The work in this study focused specifically on measuring the audibility of unconcealed packet loss as a necessary first step toward developing metrics for evaluating actual PLC methods. This approach allowed for the establishment of a baseline with the most obvious artifacts before addressing the more subtle artifacts produced by sophisticated concealment algorithms in future work. A more reliable objective metric for evaluating the audibility of PLC artifacts would allow researchers in this field to avoid the assessment via subjective listening tests, which are typically time consuming and require the involvement of specialized listeners (i.e., people with an extensive musical background). However, the accuracy of the proposed metric still falls short of human perception, and further refinements are needed.

Future work will focus on validating the generalizability of this metric across a wider range of musical genres, instrument types, and network conditions. Expanding the ground truth data will likely allow for finding more general CQT parameters and, in general, enabling deeper insights into what this metric lacks to achieve better accuracy. Furthermore, the parameters of the CQT were optimized through an iterative process guided by the listening test data. The resulting parameters provide a strong foundation for applying the CQT in perceptual audio evaluation contexts. However, the robustness of these parameters to small adjustments and their generalizability across different musical contexts remains an open question. Future work with larger and more diverse data sets will enable more comprehensive analyses of these parameters.

It is worth noticing that although PEAQ is often used in PLC research [9], prior work has shown that its scores do not correspond with PLC quality evaluation scores [18]. This is likely because PEAQ is specifically trained on distortions introduced by perceptual audio codecs and targets the BS.1387 rating scale. As a result, its applicability to network-impaired audio (resulting from packet loss) may be limited. Therefore, the findings concerning the difference between audio quality scores predicted by PEAQ and the artifact audibility scores by listeners were somewhat expected.

Although this metric demonstrates significant improvement over traditional measures through strong inter-rater reliability among expert musicians, its validation on only two instrument types limits claims of generalizability. The participant pool consisted exclusively of male expert musicians aged 25–41, which may further constrain the applicability of findings, given that perceptual judgments of glitch audibility can vary significantly across demographic groups. Factors such as gender, age, level of musical training, cultural background, and hearing sensitivity could influence how individuals perceive and evaluate audio glitches. For example, older listeners might show altered thresholds for high-frequency degradation, and cultural background might shape aesthetic judgments of glitch audibility.

This study's choice of focusing solely on specific instruments reflects an emphasis on methodological validity rather than universal applicability. Future work should not only expand testing to percussive, polyphonic, and vocal content to verify broader effectiveness but also investigate whether the performance of the proposed metric remains consistent across more diverse demographic profiles, including participants of varying musical expertise, age groups, and gender diversity. Such efforts would enhance both the ecological and perceptual validity of the metric for applications beyond professional music performance contexts.

5 CONCLUSION

In conclusion, this paper provided the following key contributions:

1. The authors proposed a novel, perceptually motivated metric for evaluating audibility of signal distortion from audio dropout that outperforms other tested metrics applied to musical signals.
2. The authors established the first publicly available ground truth data set[2] mapping objective audio features to subjective assessments of glitch audibility in NMPs, focusing on musical content.
3. A case study was presented that demonstrates how the proposed metric can drive the development of perceptually optimized PLC algorithms for musical signals in NMPs.
4. More broadly, the absence of a psychoacoustic perceptual metric specific to PLC in NMP is discussed, aiming to raise awareness of its importance for advancing the field and to spark discussion within the research community.

Although this study did not directly address the assessment of the quality of PLC methods, the proposed perceptually motivated metric for evaluating glitch audibility in

[2]https://github.com/CIMIL/Isolated-Packet-Loss-Audibility-Dataset

network-impaired music signals represents a first but necessary step toward the achievement of such a goal. In future work, the plan is to apply and improve the proposed metric for the case of PLC methods assessment.

However, further work is still needed to close the gap between the accuracy of the proposed metric and human perception. Moreover, the metric validation needs to be expanded to diverse instruments and musical contexts. It is the authors' hope that the results reported in this study could contribute to the development of more robust and perceptually aware systems for real-time musical interaction over networks.

6 ACKNOWLEDGMENT

The authors acknowledge the support of the Musical Metaverse made in Europe (MUSMET) project funded by the European Union's European Innovation Council (EIC) Pathfinder Open (Grant Number 101184379). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the EIC. Neither the European Union nor the EIC can be held responsible for them.

7 REFERENCES

[1] J.-P. Cáceres and C. Chafe, "Jacktrip: Under the Hood of an Engine for Network Audio," *J.New Music Res.*, vol. 39, no. 3, pp. 183–187 (2010 Nov.). https://doi.org/10.1080/09298215.2010.481361.

[2] L. Turchet and C. Fischione, "Elk Audio OS: An Open Source Operating System for the Internet of Musical Things," *ACM Trans. Internet Things*, vol. 2, no. 2, paper 12 (2021 May). https://doi.org/10.1145/3446393.

[3] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823–8843 (2016 Dec.). http://doi.org/10.1109/ACCESS.2016.2628440.

[4] P. Ferguson, C. Chafe, and S. Gapp, "Trans-Europe Express Audio: Testing 1000 Mile Low-Latency Uncompressed Audio Between Edinburgh and Berlin Using GPS-Derived Word Clock, First With Jacktrip Then With Dante." presented at the *148th Convention of the Audio Engineering Society* (2020 Jun.), paper 605.

[5] M. Bosi, A. Servetti, C. Chafe, and C. Rottondi, "Experiencing Remote Classical Music Performance Over Long Distance: A Jacktrip Concert Between Two Continents During the Pandemic," *J. Audio Eng. Soc.*, vol. 69, no. 12, pp. 934–945 (2021 Dec.). https://doi.org/10.17743/jaes.2021.0056.

[6] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-Quality, Low-Delay Music Coding in the Opus Codec," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), paper 8942.

[7] J.-M. Valin, J. Büthe, A. Mustafa, and M. Klingbeil, "DRED: Deep REDundancy Coding of Speech Using a Rate-Distortion-Optimized Variational Autoencoder," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 8, pp. 1441–1447 (2024 Dec.). http://doi.org/10.1109/JSTSP.2024.3482972.

[8] 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).

[9] A. I. Mezza, M. Amerena, A. Bernardini, and A. Sarti, "Hybrid Packet Loss Concealment for Real-Time Networked Music Applications," *IEEE Open J. Signal Process.*, vol. 5, pp. 266–273 (2023 Dec.). http://doi.org/10.1109/OJSP.2023.3343318.

[10] C. Aironi, L. Gabrielli, S. Cornell, and S. Squartini, "Enhancing Music Packet Loss Concealment With Generative Spectrogram Inpainting," *Universita Politecnica delle Marche and Carnegie Mellon University*, Tech. Rep. (2024).

[11] A. I. Mezza and A. Bernardini, "The IEEE-IS2 2024 Music Packet Loss Concealment Challenge," *arXiv preprint arXiv:2409.18564* (2024).

[12] Z. Lin, J. Lu, and X. Qiu, "An Effective Hybrid Low Delay Packet Loss Concealment Algorithm for MDCT-Based Audio Codec," *Appl. Acoust.*, vol. 154, pp. 170–175 (2019 Nov.). https://doi.org/10.1016/j.apacoust.2019.04.036.

[13] M. Sacchetto, Y. Huang, A. Bianco, and C. Rottondi, "Using Autoregressive Models for Real-Time Packet Loss Concealment in Networked Music Performance Applications," in *Proceedings of the 17th International Audio Mostly Conference*, pp. 203–210 (St. Polten, Austria) (2022 Sep.). https://doi.org/10.1145/3561212.3561226.

[14] K. Tsioutas, G. Xylomenos, I. Doumanis, and C. Angelou, "Quality of Musicians' Experience in Network Music Performance: A Subjective Evaluation," presented at the *148th Convention of the Audio Engineering Society* (2020 Jun.), paper 10357.

[15] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

[16] T. Thiede, W. C. Treurniet, R. Bitto, et al., "PEAQ-The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000 Jan./Feb.).

[17] D. K. Goot, H. Chaubey, T. Y. Hsu, and W. S. Deal, "A Perceptual Evaluation of Music Real-Time Communication Applications," *IEEE Access*, vol. 11, pp. 46860–46870 (2023 Apr.). https://doi.org/10.1109/ACCESS.2023.3271525.

[18] A. F. Khalifeh, A.-K. Al-Tamimi, and K. A. Darabkh, "Perceptual Evaluation of Audio Quality Under Lossy Networks," in *Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking*, pp. 939–943 (Chennai, India) (2017 Mar.). http://doi.org/10.1109/WiSPNET.2017.8299900.

[19] K. Organiściak, and J. Borkowski, "Single-Ended Quality Measurement of a Music Content via Convolutional Recurrent Neural Networks," *Metrol. Meas. Syst.*, vol. 27, no. 4, pp. 721–733 (2020 Jan.) http://doi.org/10.24425/mms.2020.134849.

[20] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-Enabled Internet of Musical Things Architectures for Remote Immersive Musical Practices," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 4691–4709 (2024 May). http://doi.org/10.1109/OJCOMS.2024.3407708.

[21] J. Dürre, N. Werner, S. Hämäläinen, et al., "In-Depth Latency and Reliability Analysis of a Networked Music Performance Over Public 5G Infrastructure," presented at the *153rd Convention of the Audio Engineering Society* (2022 Oct.), paper 10621.

[22] L. Vignati, G. Nardini, M. Centenaro, P. Casari, S. Lagén, B. Bojovic, S. Zambon, and L. Turchet, "Is Music in the Air? Evaluating 4G and 5G Support for the Internet of Musical Things," *IEEE Access*, vol. 12, pp. 38081–38101 (2024 Mar.). https://doi.org/10.1109/ACCESS.2024.3374641.

[23] L. Turchet and S. Krstulović, "DSP as a Service: Foundations and Directions," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 6212–6226 (2024 Sep.). http://doi.org/10.1109/OJCOMS.2024.3464696.

[24] A. Carôt, F. Sardis, M. Dohler, S. Saunders, N. Uniyal, and R. Cornock, "Creation of a Hyper-Realistic Remote Music Session With Professional Musicians and Public Audiences Using 5G Commodity Hardware," in *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6 (London, UK) (2020 Jul.). http://doi.org/10.1109/ICMEW46912.2020.9105995.

[25] L. Vignati, S. Dallona, and L. Turchet, "PLC Testbench: A Modular Tool for the Study and Comparison of Audio Packet Loss Concealment Algorithms," in *Proceedings of the 4th International Symposium on the Internet of Sounds*, pp. 1–8 (Pisa, Italy) (2023 Oct.). http://doi.org/10.1109/IEEECONF59510.2023.10335357.

[26] M. Schoeffler, S. Bartoschek, F.-R. Stöter, et al., "webMUSHRA—A Comprehensive Framework for Web-Based Listening Tests," *J. Open Res. Softw.*, vol. 6, no. 1, paper 8 (2018 Feb.). https://doi.org/10.5334/jors.187.

[27] N. Werner, "pyMUSHRA," https://github.com/nils-werner/pymushra (accessed May 1, 2025).

[28] M. Holters and U. Zölzer, "GstPEAQ–An Open Source Implementation of the PEAQ Algorithm," in *Proceedings of the 18th International Conference on Digital Audio Effects* (Trondheim, Norway) (2015 Nov./Dec.).

[29] B. McFee, M. McVicar, D. Faronbi, et al., "librosa/librosa: 0.10.1," (2023 Aug.). https://doi.org/10.5281/zenodo.8252662.

[30] D. Bogdanov, N. Wack, E. Gómez, et al., "ESSENTIA: An Audio Analysis Library for Music Information Retrieval," in *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 493–498 (Curitiba, Brazil) (2013 Nov.)

[31] M. Stimberg, R. Brette, and D. F. Goodman, "Brian 2, An Intuitive and Efficient Neural Simulator," *eLife*, vol. 8, p. e47314 (2019 Oct.). https://doi.org/10.7554/eLife.47314.

[32] T. Necciari, B. Laback, S. Savel, et al., "Auditory Time-Frequency Masking: Psychoacoustical Data and Application to Audio Representations," in *Proceedings of the 8th International Conference on Speech, Sound and Music Processing: Embracing Research in India*, pp. 146–171 (Bhubaneswar, India) (2012 Mar.).

[33] T. Necciari, B. Laback, S. Savel, et al., "Auditory Time-Frequency Masking for Spectrally and Temporally Maximally-Compact Stimuli," *PloS One*, vol. 11, no. 11, p. e0166937 (2016 Nov.). https://doi.org/10.1371/journal.pone.0166937.

THE AUTHORS

Luca Vignati

Luca Vignati received a B.Sc. degree in computer and electronic engineering from the University of Pavia, Italy, and an M.Sc. degree in computer science and engineering from Politecnico di Milano, Italy. He received his Ph.D. in industrial innovation from the Department of Information Engineering and Computer Science of the University of Trento and EIT Digital.

•

Luca Turchet

Luca Turchet is an associate professor in the Department of Information Engineering and Computer Science of the University of Trento, Italy. He is the chair of the Institute of Electrical and Electronics Engineers (IEEE) Emerging Technology Initiative on the Internet of Sounds and the founding president of the Internet of Sounds Research Network.