ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

AN OPEN ARCHITECTURE FOR REAL-TIME QOS MONITORING AND MOS ESTIMATION IN IP-BASED NETWORKS

Eng. Vladimir Vichev
Technical University - Varna
Assoc. Prof. Dr. Eng. Todorka Georgieva
Technical University - Varna

Abstract: *The increasing demand for real-time IP-based services, particularly Voice over IP (VoIP), necessitates advanced and adaptive methods for Quality of Service (QoS) monitoring. This paper presents a modular, open-source architecture for real-time QoS assessment and perceptual quality estimation using Mean Opinion Score (MOS) modeling. The proposed system integrates passive traffic analysis, statistical QoS metrics, and predictive MOS evaluation. A layered design allows separation of concerns and flexible deployment. The system supports real-time alerting, QoS-aware service management, and enables early detection of degradation using normalized jitter, delay, and packet loss.*

Keywords: *QoS monitoring, IP networks, VoIP, MOS estimation, real-time evaluation, packet loss, jitter, delay, SLA*

1. Introduction

Modern IP-based services, especially real-time communications such as VoIP, are highly sensitive to network performance. Minor degradations in metrics such as packet loss, jitter, and delay can significantly impact perceived quality, leading to reduced user satisfaction and Service Level Agreement (SLA) violations [11],[5],[9]. Ensuring the QoS for such applications requires accurate, low-latency monitoring systems capable of interpreting technical metrics in terms of perceptual quality, commonly represented by the MOS [5], [9]. This motivates the development of an open, modular architecture that supports real time MOS estimation and state classification based on empirical thresholds and codec specific behavior [11],[12],[13],[14].

This paper introduces a comprehensive system for real-time QoS monitoring and MOS estimation tailored for VoIP services. The approach combines passive traffic measurement, statistical analysis of delay and loss patterns, codec-aware thresholding, and lightweight regression-based prediction models. A layered architecture separates data acquisition, metric processing, decision logic, and visualization, ensuring both scalability and maintainability. Furthermore, the system is designed with extensibility in mind, allowing integration with machine learning models, RESTful Application Programming Interface (API) and cloud based environments.

2. Related work

QoS monitoring in IP-based networks has been the subject of extensive research over the past two decades, particularly in the context of real-time applications such as VoIP[11][5]. Traditional approaches often rely on periodic polling or SNMP-based monitoring, which

DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

provide only a partial view of service performance and lack sufficient temporal granularity[1]. More advanced methods involve packet-level analysis using tools like NetFlow, sFlow, or Deep Packet Inspection (DPI), which can extract richer metrics but are often resource-intensive and may not scale well in high-throughput environments.

Perceptual quality estimation, most notably using the MOS, has evolved from subjective listening tests to algorithmic models such as the ITU-T E-model [5] and PESQ/POLQA [9]. The E-model provides a parametric estimation of quality based on factors like packet loss, delay, and codec type, but lacks sensitivity to burst loss and dynamic variations. PESQ and POLQA offer more accurate assessments but require access to reference signals, making them unsuitable for passive real-time monitoring [9].

Recent research has introduced machine learning-based MOS prediction, leveraging regression models trained on synthetic or real traffic traces [12],[13],[14]. These methods improve prediction accuracy and support real-time application but often lack interpretability and require large annotated datasets. Other studies have proposed composite indices such as Time over Threshold (ToT) and the Integrated Degradation Index (I-index) to capture multidimensional service degradation and improve state classification accuracy [4], [6], [2]. For example [4] demonstrate regression-based real-time MOS estimation integrating ToT-derived parameters, while [6] and [2] present edge-assisted QoS frameworks utilizing degradation indices for adaptive alerting. Several architectures for QoS monitoring have also been developed, ranging from centralized SNMP-based systems to distributed agents with predictive analytics [6], [7], [8]. Although these solutions provide valuable insights into performance monitoring, most remain proprietary or lack codec-aware adaptability. Centralized platforms such as Zabbix or Prometheus can monitor infrastructure metrics but do not natively support MOS estimation or codec-level adaptation [4]. Commercial solutions (Cisco Prime, VoIPmonitor) offer robust visualization and alerts, but their proprietary nature and high integration cost limit flexibility [1]. Studies such as [4] and [7] demonstrate hybrid models that integrate passive traffic metrics with regression or neural approaches for real-time MOS prediction. Other works [6] and [2] propose edge-assisted QoS analyzers capable of on-device evaluation to reduce latency and monitoring overhead. These studies confirm the trend toward perceptually aware QoS monitoring, though most remain limited in openness, scalability, or codec adaptation. Open-source and modular implementations are still scarce, and interoperability with SLA-driven environments is rarely addressed. The architecture proposed in this paper builds upon these recent findings by offering an open, layered, codec-aware design that supports both real-time QoS assessment and perceptual quality interpretation using lightweight models.

In contrast, this work aims to unify passive traffic analysis with perceptual quality evaluation, codec-aware thresholds, and flexible modular implementation. By synthesizing empirical findings from VoIP QoS research with open-source software tools, this work bridges the gap between network-level measurements and perceptual service quality in real time [11],[13],[10].

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

3. System architecture

The proposed system for QoS monitoring and MOS estimation is designed as a layered architecture, following the principles of modularity, flexibility, and scalability. It separates functional concerns into distinct layers, allowing each component to evolve or be replaced independently while maintaining overall system cohesion. The architecture follows a hierarchical structure, typical for communication and computing systems, where data flows upward from acquisition to interpretation. Each layer depends on the processed results of the one below it, ensuring modular encapsulation and scalability. At the top of the hierarchy is the Visualization and Notification Layer, which serves as the customer and operator interface, presenting the evaluated QoS state, alerts, and reports through dashboards or APIs. The architecture includes five layers: data acquisition; preprocessing; evaluation; aggregation and visualization, allowing modular deployment in both centralized and distributed environments. “Fig. 1” illustrates the full layered architecture, including the toolchain and data flow between layers.

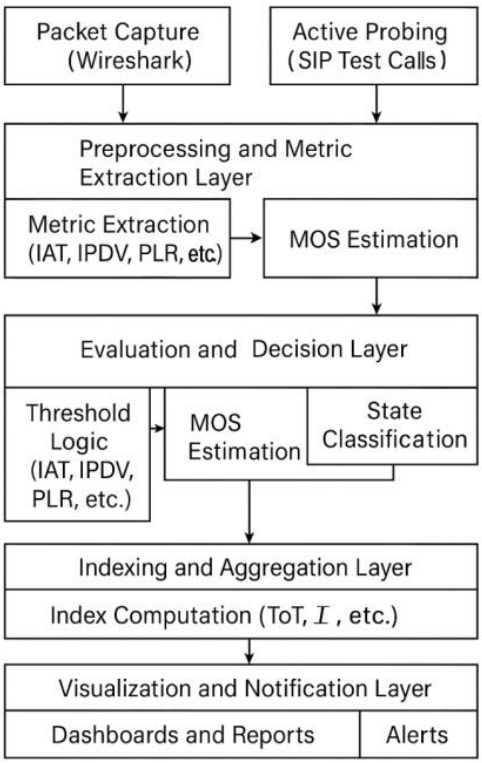

Fig. 1. System architecture

Data Acquisition Layer layer interacts with network traffic via passive capture (using Wireshark/WinEyeQ) and active probing (via Session Initiation Protocol (SIP) test calls with StarTrinity SIP Tester or WinSIP). It supports real-time packet collection or trace-file processing for retrospective analysis. For VoIP traffic, it focuses primarily on Real-time Transport Protocol (RTP) streams and SIP signaling.

Preprocessing and Metric Extraction Layer where, raw packets are parsed into call sessions, and per-stream metrics are computed, including: Inter-arrival time (IAT); Inter-packet delay variation (Inter-Packet Delay Variation (IPDV) / jitter); One-way delay (when

е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

synchronization allows); Packet Loss Rate (PLR); Effective bitrate. Additional preprocessing includes outlier filtering and computation of percentiles (p95), which are used in later normalization.

Evaluation and Decision Layer is the core analytical component. And it: Performs MOS estimation using codec-specific regression models; Applies empirical thresholds (for PLR, pps, bitrate); Classifies service state into OK / Warning / Fail, based on MOS and critical metrics; Supports real-time evaluation and adaptive thresholding.

Indexing and Aggregation Layer. This layer aggregates per-session evaluations and derives secondary indicators: ToT, duration when metrics exceed safe bounds; I-Index, weighted score summarizing quality loss; Concordance checks: comparing passive metrics against perceptual scores; SLA logic: validation of thresholds per service class or customer profile.

Visualization and Notification Layer is responsible for: Real-time dashboards using specific tool, or custom web interfaces; Alerts via SNMP, email, triggered by state transitions; Export of historical reports in CSV, PDF, or other formats; Optional API for integration with external monitoring platforms.

4. QoS evaluation algorithm

To enable real-time assessment of service quality and predictive MOS estimation, the proposed system implements a lightweight yet robust evaluation algorithm. This algorithm translates raw QoS metrics into perceptual service states using codec-aware thresholds, statistical normalization, and regression-based MOS prediction. It also calculates quality degradation indices to support SLA decisions and state transitions.

The algorithm operates on a minimal but sufficient set of QoS parameters extracted from RTP flows:

- PLR – calculated over RTP sequence gaps.
- IAT – used for jitter and burstiness detection.
- IPDV – a measure of jitter.
- One-way Delay – when measurable via synchronized clocks or timestamps.
- Bitrate – measured over active call periods.
- Codec ID – determines packetization time and reference thresholds.

Each parameter is aggregated using statistical descriptors such as median, mean, and p95, which offer better resilience to outliers in bursty traffic.

QoS metrics are normalized to a unified scale to improve comparability and facilitate index computation [12],[5]. For example, the normalized inter-arrival time percentile is computed as:

$$nIAT95 = \frac{IAT_{95\%} - IAT_{base}}{IAT_{tol}}, \quad (1)$$

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

Where IATbase and IATtol are codec-specific baseline and tolerance thresholds, respectively. Similar normalization is applied to: PLR (vs codec-specific PLR baseline), IPDV (vs maximum jitter tolerance) and Bitrate (vs nominal media rate).

The core MOS prediction is handled through a nonlinear regression model, trained on measurement data obtained through synthetic load conditions and controlled degradations [12], [13]. The model outputs a predicted MOS value in the ITU-T scale (1 to 5) [5].

$$MOS = f(nPLR, nIAT95, nIPDV, nBitrate)\,, \tag{2}$$

The model is derived from empirical data using linear or polynomial regression per codec and scenario. The MOS estimation enables interpretation of technical degradation in perceptual terms, aligned with user experience. Two auxiliary indices are computed to capture degradation severity and duration:

ToT: an index for each degradation metric, reflecting the percentage of time the metric exceeds its tolerance.

$$ToT_{PLR} = \frac{time_{PLR>PLR_{tol}}}{total\ time}\,, \tag{3}$$

I-Index : a weighted sum of normalized metrics.

$$I = w_1.nPLR + w_2.nIAT95 + w_3.nIPDV + w_4.nBitrate\,, \tag{4}$$

Weights wi are calibrated to match the codec's relative sensitivity to each metric, obtained from training data [12].

Based on MOS and thresholds, the algorithm classifies the session quality into three service states shown on “Table 1”.

Table 1

State classification and alerts

MOS Range	Service State	Description
≥ 4.0	OK	Acceptable user quality
3.0 – 4.0	Warning	Potential degradation
< 3.0	Fail	Unacceptable quality

5. Implementation

To validate the proposed architecture and QoS evaluation algorithm, a full implementation was developed using accessible open-source tools and tested in a controlled environment. The goal was to demonstrate that real-time perceptual quality monitoring and

DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

predictive alerting can be achieved without reliance on proprietary hardware or software platforms [11],[12],[13].

The implementation leverages the following tools:

• WinSIP– used for generating synthetic VoIP traffic and simulating various codec configurations [15].

• StarTrinity Network Tester – for active traffic injection and quality measurement under varying network loads [10].

• WinEyeQ/Wireshark – for passive packet capture and inspection of RTP streams and SIP signaling [15], [16].

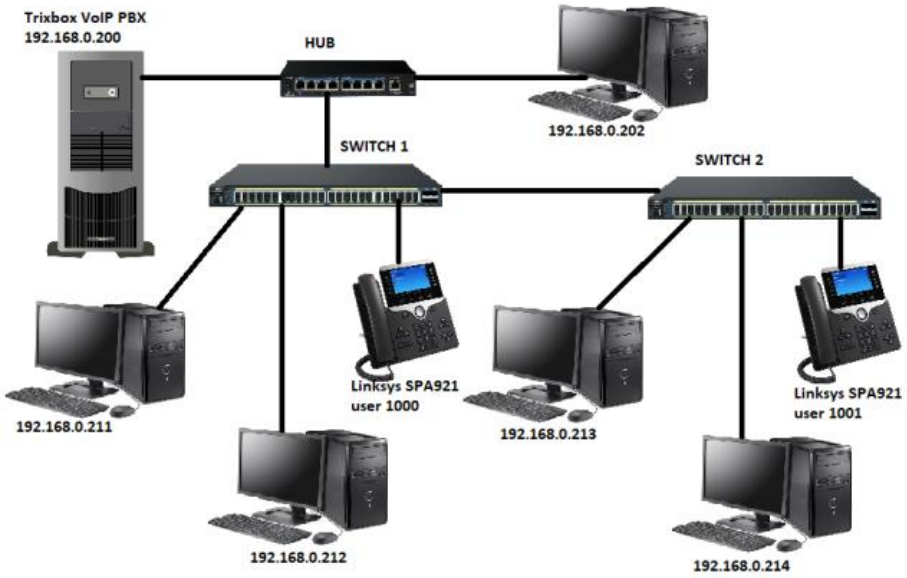

Fig. 2. Experimental setup

The system was deployed in a lab environment, with simulated traffic routed through configurable bottlenecks to model real-world congestion and degradation. “Fig. 2” shows the experimental setup used in validation. It includes: a SIP client/server and packet capture and inspection (192.168.0.211 and 192.168.0.214); traffic generator establishing varying network loads (192.168.0.212 and 192.168.0.213); low level packet analyzer on 192.168.0.202.

The experiments were conducted under controlled network conditions to assess how the proposed architecture performs in real time when faced with variable network degradation. Each test scenario was defined by a codec (G.711, G.729, iLBC) and a degradation pattern (uniform loss, burst loss, delay spikes, and bandwidth limitation). “Fig. 3” illustrates the real-time IAT evolution during a sample call session. The observed variations correspond to jitter dynamics, where the algorithm accurately detects deviation peaks beyond 20 ms, triggering a warning state. “Fig. 4” presents bandwidth utilization across multiple simulated calls. The data confirms that bitrate adaptation remains stable within codec-specific ranges (e.g., 64 kbps for G.711) under moderate load and degrades predictably under congestion, validating the reliability of the metric extraction layer. “Fig. 5” shows the correlation between predicted MOS and R-factor in real time. The system maintains consistent MOS estimation within ±0.2 of the

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

reference model during transient degradation, demonstrating the responsiveness of the evaluation algorithm. “Fig. 6” compares aggregated QoS metrics for the iLBC codec, revealing its higher resilience to packet loss (PLR ≤ 2%) and smaller MOS variance compared to G.711. These findings align with ITU-T E-model predictions and confirm codec-specific accuracy of the regression-based MOS estimation.

Overall, the experimental validation shows that the architecture successfully detects performance degradation at an early stage and provides interpretable quality indicators suitable for SLA monitoring and alerting. The results confirm the feasibility of real-time, codec-aware MOS prediction using open-source tools without proprietary dependencies.

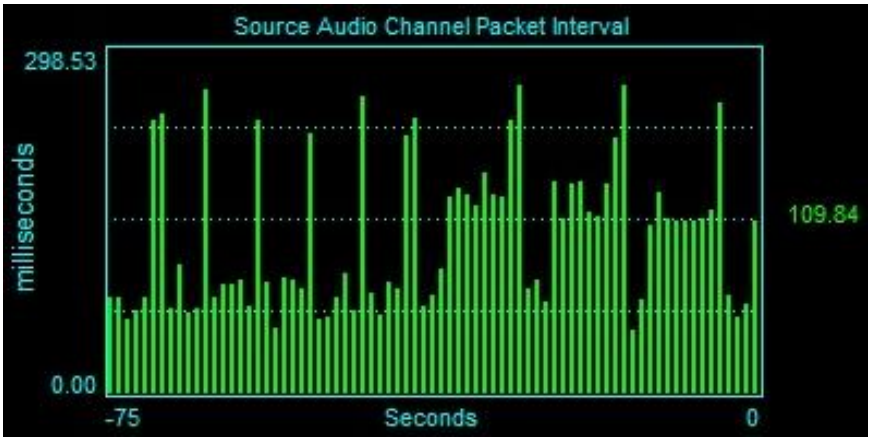

Fig. 3. Real-time packet interval

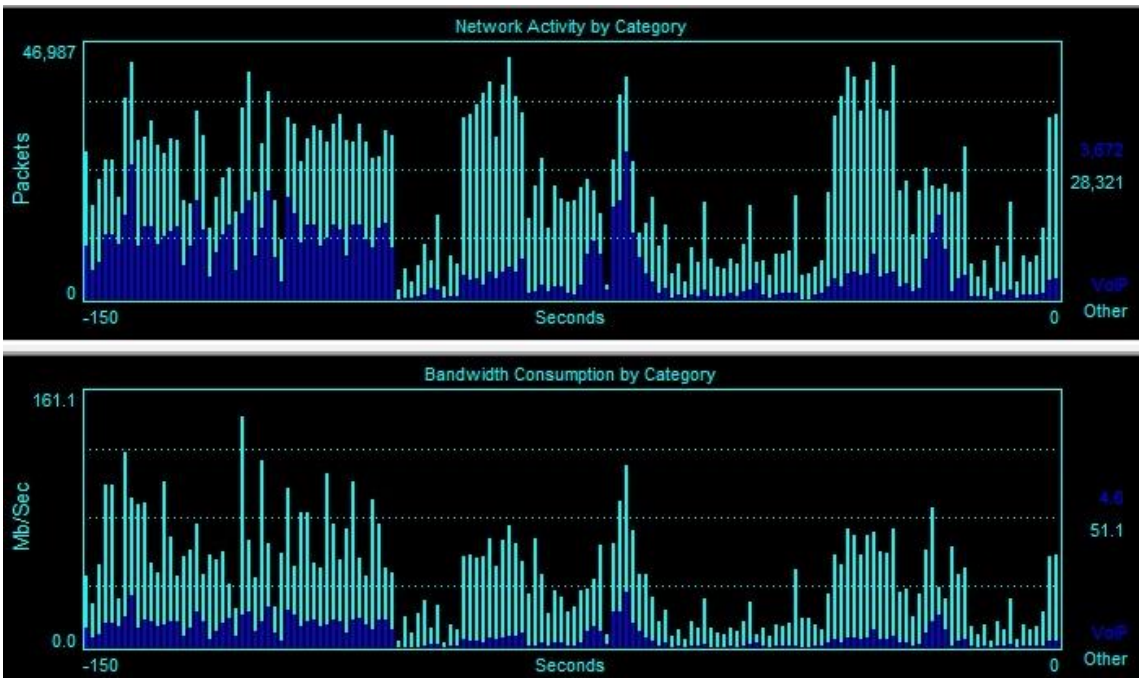

Fig. 4. Real-time bandwidth utilization

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

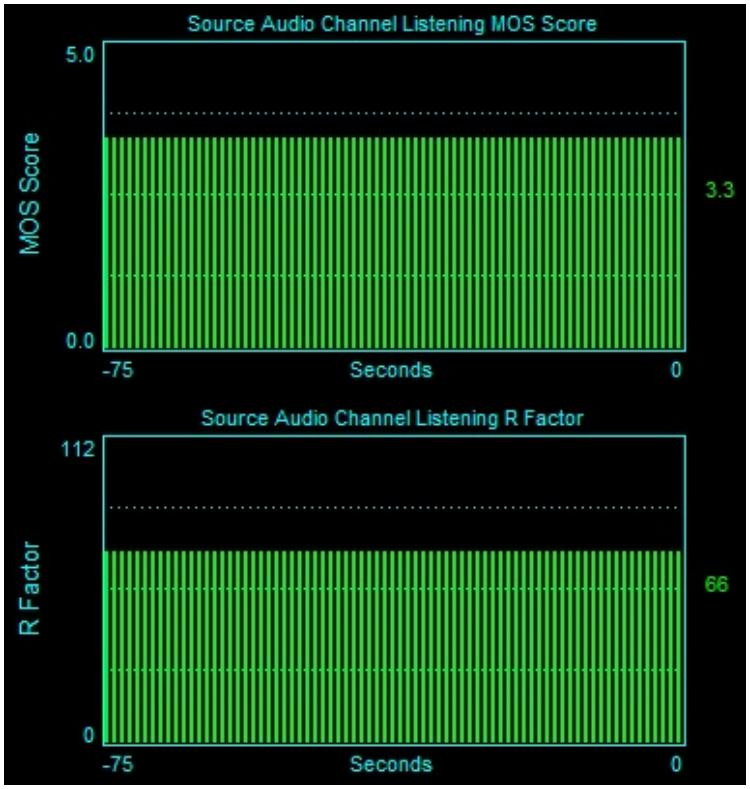

Fig. 5. Real-time MOS and R-factor

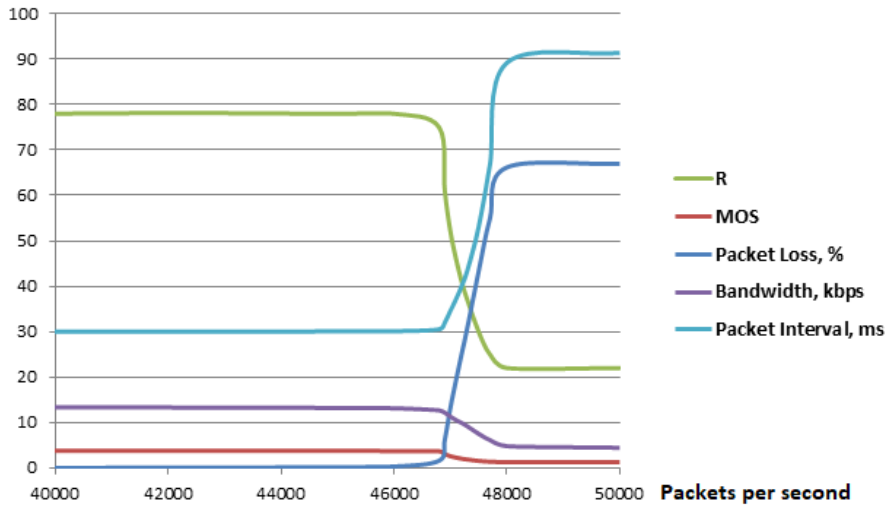

Fig. 6. Obtained QoS metrics, iLBC codec

6. Conclusion

The proposed architecture introduces a five-layer model that separates traffic acquisition, metric processing, service evaluation, index computation, and visualization. A lightweight algorithm is implemented to estimate perceptual quality (MOS) using normalized QoS parameters such as packet loss, jitter, delay, and bitrate. The system was validated in a controlled testbed using open tools such as StarTrinity SIP Tester, Wireshark, and WinEyeQ. Results demonstrate accurate service state classification and timely alert generation, even under dynamic network degradation. Comparative analysis with recent works [4], [7], [6], [2], [3], [8]

DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

shows that the proposed system achieves performance comparable to state-of-the-art QoS monitoring architectures. For instance, the average deviation between predicted and reference MOS values in our implementation is within ±0.2, which is similar to the regression-based models reported in [4] and [2]. Unlike AI-intensive approaches such as [3], this solution maintains lower computational complexity while preserving high estimation accuracy. Furthermore, in contrast to the edge-assisted architectures described in [6], the proposed system emphasizes modularity and open-source deployment rather than hardware-bound optimization.

7. References

1. Blake S., Black D., Carlson M., Davies E., Wang Z., Weiss W., An Architecture for Differentiated Services, IETF RFC 2475, Dec. 1998. Available: https://datatracker.ietf.org/doc/html/rfc2475.
2. Garcia-Torres M., Pinto-Roa D., Castilo C., Quinonez B., Vazques G., Allogretti M., Diaz M., Feature selection applied to QoS/QoE modeling on video and web-based mobile data services: An ordinal approach, Computer Communications, vol. 217, pp. 230–245, 2024.
3. Hamidou H., Kouraogo J., Tapsoba D., Sie O., Machine learning based Quality of Experience (QoE) Prediction Approach in Enterprise Multimedia Networks, Computer Science Research Days, 2022.
4. Hu Z., Ren Y., Tao Y., Haijun G., Qing L., Evaluating QoE in VoIP networks with QoS mapping and machine learning algorithms, Neurocomputing, vol. 386, pp. 63–83, 2020.
5. ITU-T Recommendation G.107, The E-model: A computational model for use in transmission planning, Int. Telecommun. Union, Geneva, Switzerland, Apr. 2015.
6. Leonte S., Pastrav A., Zamfirescu C., Puschita E., Voice Quality Evaluation in a Mobile Cellular Network: In Situ Mean Opinion Score Measurements, IEEE Sensors, vol. 24, no. 20, 2024.
7. Mauro M. Di, Calatro G., Postiglione F., Song W., Liotta A., Multivariate time series characterization and forecasting of VoIP traffic in real mobile networks, IEEE Transactions on Network and Service Management, 2023.
8. Moreira R., Cunha H., Moreire L., Silva F., VINEVI: A Virtualized Network Vision Architecture for Smart Monitoring of Heterogeneous Applications and Infrastructures, IEEE Access, vol. 12, 2024.
9. Rix A., Beerends J., Hollier M., Hekstra A., Perceptual Evaluation of Speech Quality (PESQ)—A New Method for Speech Quality Assessment of Telephone Networks and Codecs, IEEE Int. Conf. Acoustic, Speech, Signal Process, 2001.
10. StarTrinity SIP Tester at: https://startrinity.com/VoIP/SIPTester
11. Vichev V., Georgieva T., Behavior of Voice Codecs QoS Metrics in the IP Network Overload Zone, International Conference AUTOMATICS AND INFORMATICS 2024, IEEE, 2024.

DOI : https://doi.org/10.29114/ejtuv.vol1.iss2.38

12. Vichev V., Georgieva T., Behavior of VoIP Traffic QoS Metrics in Loaded Networks, 32nd National Conference with International Participation "Telecom 2024", IEEE, 2024.
13. Vichev V., Georgieva T., IP Network Performance Analysis in VoIP Environment, International Conference AUTOMATICS AND INFORMATICS 2024, IEEE, 2024.
14. Vichev V., Nonlinear Relationships Analysis of QoS Metrics Under Dynamic Network Conditions, 33th National Conference with International Participation "Telecom 2025", IEEE, 2025.
15. WinEyeQ, WinSIP at: https://touchstone-inc.com/wineyeq.html.
16. Wireshark Network Protocol Analyzer at: https://wireshark.org.