УДК 621.396.1 DOI: 10.31673/2786-8362.2025.021693

Ветошко І.П.; Кравчук С.О., д.т.н.

ALGORITHM FOR ENHANCING VOICE QUALITY IN 5G STANDALONE VIA MOS-ORIENTED DYNAMIC LINK ADAPTATION

Vetoshko I.P., Kravchuk S.O. Algorithm for enhancing voice quality in 5G standalone via MOS-oriented dynamic link adaptation. This paper proposes a MOS-driven, risk-aware dynamic MCS adaptation algorithm for VoNR in 5G Standalone networks. The algorithm enhances voice delivery reliability by combining SNR-uncertainty look-ahead, CVaR-based control of HARQ retransmission tails, and an external VoNR feedback loop for Δ-offset regulation. Unlike OLLA and HARQ-aware/unaware baselines, which optimize BLER without delay prediction, the algorithm evaluates candidate MCS values through analytical BLER curves, IR-combining gain, late-loss probability, and E-model MOS estimation under mobility conditions. Simulation results using CRN-blocked experiments confirm stable performance across BLER∈0.01-0.10: MOS≈4.26-4.27, >98.5% on-time RTP delivery, and Retx≥2≈2.0-2.3%, with no notable goodput degradation. Benchmark schemes (OLLA, Random) generate ≈22–25% deep HARQ chains and MOS<4.0, demonstrating the advantage of MOS-oriented decision policy over BLER-only control. The results show that the proposed algorithmic approach provides deterministic QoE under Doppler-induced CQI drift and is suitable for real-time VoNR deployment in native 5G SA environments.

Keywords: VoNR, 5G Standalone, MCS adaptation, HARQ retransmissions, MOS, QoS/QoE, BLER, dynamic link adaptation, risk-aware scheduling, real-time voice services

Ветошко І.П., Кравчук С.О. Алгоритм покращення якості голосового зв'язку в мережах 5G Standalone на основі MOS-орієнтованої динамічної адаптації радіоканалу. У роботі запропоновано MOS-орієнтований, ризик-чутливий алгоритм динамічної адаптації MCS для VoNR у мережах 5G Standalone. Запропонований метод підвищує надійність доставки голосового трафіку за рахунок поєднання прогнозування SNR з урахуванням невизначеності, CVaR-керування глибокими HARQ-ретрансляціями та зовнішнього VoNR-зворотного контуру для регулювання Δ-offset. На відміну від базових схем OLLA та HARQ-aware/unaware, які оптимізують BLER без передбачення затримок, алгоритм оцінює кандидатні рівні MCS за аналітичними BLER-кривими, приростом IR-комбайнінгу, ймовірністю late-loss та E-model-оцінкою MOS в умовах мобільності. Результати моделювання на основі CRN-блокованих експериментів підтвердили стабільність роботи у діапазоні BLER∈0.01–0.10: MOS≈4.26–4.27, своєчасна доставка RTP-пакетів >98.5%, імовірність Retx≥2≈2.0–2.3% без погіршення goodput. Контрольні підходи (OLLA, Random) формували ≈22–25% глибоких HARQ-ланцюгів та MOS<4.0, що демонструє перевагу MOS-орієнтованої стратегії над BLER-залежним управлінням. Отримані результати показують, що запропонований підхід забезпечує детермінований рівень QoE при Doppler-зумовленому дрейфі CQI та є придатним для реального впровадження VoNR у нативних 5G SA-мережах.

Ключові слова: VoNR, 5G Standalone, адаптація MCS, HARQ-ретрансляції, MOS, QoS/QoE, BLER, динамічне керування каналом, risk-aware scheduling, реальний час голосових сервісів

Introduction

Statement of the problem. Voice over New Radio (VoNR) in native 5G Standalone networks introduces fundamentally new requirements for Quality of Service compared to VoLTE and legacy IMS-based architectures. Voice traffic is highly sensitive to jitter, serial HARQ delays, late-loss delivery and BLER bursts, which may lead to conversational artifacts and MOS degradation even when average link-level metrics remain within nominal limits. However, existing MCS-control approaches — OLLA, HARQ-aware and HARQ-unaware link adaptation — optimize BLER reactively and do not predict retransmission-tail risks, do not evaluate on-time delivery probability, and lack a direct MOS-oriented optimization target [1-3].

The problem intensifies under mobility, where Doppler-induced CQI drift and SNR-prediction errors result in long HARQ chains and unstable perceptual quality. Maintaining acceptable MOS>4.2 while holding Retx≥2 below 3% and ensuring ≥98.5% timely RTP arrival remains an unresolved challenge for real-time VoNR scheduling [4-6]. Therefore, there is a need for an algorithm that considers BLER-curve analytics, IR-combining gain, delay-deadline constraints and E-model voice quality evaluation simultaneously, incorporating risk-aware conditioning rather than relying solely on BLER-target tuning. Addressing this requires a new decision rule that integrates SNR-uncertainty

look-ahead, HARQ-tail suppression and MOS-driven objective formulation within a VoNR control loop.

Analysis of recent studies. Publications [1-3] describe 5G SA and NR-air-interface architecture, highlighting the limitations of conventional OLLA BLER-tracking under HARQ-driven delays. Works [4-5] examine voice-service quality in 5G and discuss CQI/MCS adaptation strategies, yet do not provide mechanisms for tail-risk control or SLA-bounded deadline assurance. Results in [6] evaluate VoNR scheduling and report MOS sensitivity to late-loss spikes, confirming the need for MOS-oriented adaptation rather than raw BLER optimization.

At the same time, studies [7-8] analyze algorithmic MCS selection but remain focused on throughput and long-term BLER trends rather than real-time speech quality metrics. Research [9] evaluates VoNR performance in 5G but lacks CVaR-based tail quantification and does not model serial retransmission penalties. Work [10] presents ML-driven QoS-adaptive scheduling but does not integrate risk-aware HARQ barriers or on-time delivery guarantees.

Thus, the literature confirms the relevance of VoNR optimization research, but reveals a gap: existing solutions insufficiently address MOS-driven decision criteria, risk-aware suppression of Retx≥2 events, and prediction of delay-deadline violation under SNR uncertainty. The present study fills this methodological gap by developing a MOS-driven, HARQ-tail-controlled dynamic MCS algorithm for VoNR in 5G SA.

The purpose of this paper is to develop and theoretically justify a MOS-oriented, risk-sensitive dynamic MCS selection algorithm for VoNR in 5G Standalone networks, aimed at ensuring high QoS stability under mobility, reducing HARQ tail events, and improving perceptual voice quality. The proposed algorithmic approach addresses the key limitations of traditional OLLA, HARQ-aware and HARQ-unaware mechanisms, which optimize BLER but do not predict retransmission-related risks or account for MOS-driven performance.

To achieve this goal, the research sets the following core objectives:

- to analyze the impact of delay, jitter and retransmission depth on MOS and identify weaknesses of baseline MCS control strategies;
- to design a dynamic MCS algorithm with MOS-based objective function, CVaR-type risk barriers for Retx≥2 events and SNR-uncertainty-aware look-ahead;
- to implement a VoNR control loop for SLA stabilization;
- to validate the solution through CRN-based simulation against baseline schemes (OLLA, HARQ-aware/unaware, Random) using MOS, on-time delivery, Retx-tails and goodput as primary metrics.

The expected outcome of the study is a robust VoNR-optimized MCS-control mechanism capable of maintaining MOS≈4.26-4.27, ensuring >98.5% on-time delivery, suppressing Retx≥2 to ≈2-2.3%, and outperforming classical adaptation approaches in terms of QoE, resilience to channel variability and suitability for real-time voice transmission.

Presentation of the main research material

In modern Fifth-Generation (5G) mobile networks, particularly in the Standalone (SA) architecture, the deployment of voice services based on Voice over New Radio (VoNR) has become a key direction in the evolution of telecommunications systems. Unlike Voice over LTE (VoLTE), which relies on legacy infrastructures, VoNR operates entirely on the 5G NR radio interface, thereby requiring fundamentally new algorithmic approaches to Quality of Service (QoS) assurance [1]. The main challenge arises from the strong sensitivity of voice traffic to delay, jitter, and packet loss. Even short-term disruptions or excessive HARQ retransmissions may introduce speech artifacts, conversational delay, or a sharp drop in perceived speech quality measured by the Mean Opinion Score (MOS). Traditional approaches to Modulation and Coding Scheme (MCS) adaptation in 5G largely rely on fixed Channel Quality Indicator (CQI) thresholds or on the classical Outer Loop Link Adaptation (OLLA) mechanism [2]. While such algorithms provide reactivity to instantaneous channel conditions, they do not consider the historical behavior of QoS indicators, do not provide

control over the "tail" of HARQ retransmissions, and do not optimize performance with respect to the final perceptual metric (MOS). As a result, the perceived voice quality often becomes unstable, particularly in mobility scenarios with high user velocity (e.g., 60 km/h), where fast Doppler fluctuations lead to frequent SNR prediction errors and late packet deliveries [3].

To address these challenges, this work introduces a HARQ/BLER-oriented dynamic MCS selection algorithm featuring MOS-driven target BLER tuning and explicit control of serial delays and retransmission risks. A key novelty of the algorithmic approach is the integration of classical link-adaptation techniques with risk-aware KPI control: the algorithm predicts the probability of on-time packet delivery and constrains the share of HARQ retransmissions with depth Retx ≥ 2 using barrier-type conditions that limit worst-case degradation [4]. In this way, the proposed algorithm goes beyond the traditional "target BLER ↔ goodput" trade-off and directly optimizes the end-user experience, ensuring a consistently high MOS even under rapidly varying radio conditions. The paper provides the mathematical formulation, implementation details, and simulation-based comparison with baseline algorithms (HARQ-aware, HARQ-unaware, OLLA, and Random). The results demonstrate the ability of the proposed solution to sustain >98.5% on-time delivery, maintain the Retx ≥ 2 tail below 3%, and achieve MOS above 4.25 without compromising average throughput.

Baseline and the Proposed Algorithms. *HARQ-unaware* algorithm selects the highest MCS index that satisfies a predefined target BLER based on the current SNR estimate, applying a small optimistic SNR shift (≈ +0.9 dB). It does not account for transmission history or retransmission-related risks. While it ensures high instantaneous spectral efficiency, it increases the probability of events with two or more retransmissions Retx ≥ 2 and the occurrence of late-loss packets, both of which negatively impact MOS.

HARQ-aware (myopic). This baseline approach is similar to the HARQ-unaware algorithm but uses a conservative SNR shift (≈ –0.3 dB) and a reactive mechanism for controlling retransmission tails. If the fraction of successful transmissions requiring Retx ≥ 2 exceeds thresholds (2.2% or 4%), the MCS index is decreased by one or two steps. Conversely, if this fraction remains consistently below 0.8%, the MCS is increased by one [5]. Although this mitigates retransmission tails locally, it neither enforces strict on-time delivery constraints nor anticipates the delay impact of future HARQ rounds. This limitation is particularly critical for voice traffic, where even modest additional delays may result in noticeable degradation of perceived speech quality.

Trivial Random. This algorithm performs a uniformly random selection of the MCS index from the allowable range. It serves as a lower reference bound for evaluating adaptive strategies, illustrating the consequences of having no adaptation to channel conditions. Since this approach accounts for neither the current SNR state nor any Quality of Service requirements, its results are used exclusively for comparative analysis against more sophisticated algorithms.

OLLA (classical outer loop). The classical OLLA algorithm aims to maintain a predefined average BLER by dynamically adjusting an SNR offset (Δ), as shown below:

$$\Delta \leftarrow \begin{cases} \Delta + \varepsilon\downarrow, & ACK, \\ \Delta - \varepsilon\uparrow, & NACK, \end{cases} \qquad m = \arg\max_{m \in \mathcal{M}} \{ \text{BLER}\,(\hat{\gamma} + \Delta, m) \leq \text{target} \} \qquad (1)$$

where Δ∈[−12, +6]. After a successful transmission (ACK), the offset increases by a small step ε↓ (0.05 dB) while after a failed attempt (NACK), it decreases by a larger step ε↓ (0.6 dB) with saturation within the interval [−12, +6] dB. The MCS index is defined as the maximum of those for which BLER $(\hat{\gamma} + \Delta, m) \leq$ target-BLER, where $\hat{\gamma}$ denotes the estimated SNR. Although OLLA effectively stabilizes the long-term average BLER, it does not guarantee that packet deliveries meet a strict delay deadline (D_max) under rapid channel fluctuations. Moreover, OLLA does not control the tail of the retransmission distribution Retx ≥ 2, which results in increased jitter and a higher fraction of late-loss deliveries – both of which are critical for voice services.

The proposed algorithm. MOS-driven, risk-aware look-ahead with VoNR control loop is specifically designed for VoNR and selects the MCS so as to maximize the expected speech quality (MOS) while simultaneously satisfying strict on-time delivery requirements (SLA) and controlling risk-related "tails" (events with Retx ≥ 2 and late-loss packets). Unlike baseline approaches that primarily optimize BLER, the proposed algorithm focuses on the direct impact on MOS,

incorporating the effects of delay and jitter (through on-time delivery and BurstR), packet losses (hard + late), and the burstiness of degradation (BurstR). The decision process relies on an effective channel estimate γ_{eff}, which combines the instantaneous SNR estimate $\hat{\gamma}$, a dynamic offset Δ (controlled by an external loop), and an aging penalty $k_{\text{age}} \cdot q_{\text{wait}}$ accumulated while the packet waits in the queue:

$$\gamma_{\text{eff}} = \hat{\gamma} + \Delta - k_{\text{age}} \cdot q_{\text{wait}} \quad (2)$$

where q_{wait} is the predicted queueing delay (before the first transmission attempt), and k_{age} is a coefficient reflecting the degradation of the relevance of $\hat{\gamma}$ during waiting (≈1.4 dB per 10 ms in the simulation). Based on γ_{eff}, a base MCS m_0 that satisfies the target BLER is selected, after which an adaptive candidate window is constructed around m_0. The radius of this window adapts to the observed volatility of decisions: it contracts under stable conditions and expands when disturbances occur [5]. This reduces the risks of both excessive aggressiveness and staying "locked'' in overly conservative settings.

For each candidate, a short look-ahead is performed that accounts for SNR estimation uncertainty and the HARQ structure. The uncertainty is modeled using a three-point approximation near a normal error distribution (RMS ≈2.1 dB): for the three SNR nodes $\gamma = \gamma_{\text{eff}} \pm \sigma_{\text{est}}$, γ_{eff} , the algorithm computes HARQ per-round success probabilities with IR combining, the delay distribution across HARQ rounds, the probability of meeting the deadline D, the expected number of retransmissions, and the probability of an event Retx ≥ 2. Using these estimates, the conditional mean and standard deviation of delay for on-time packets are calculated, while late deliveries are treated as losses [6]. These quantities are then used to compute the expected MOS via the E-model, with the BurstR parameter capturing the burstiness of losses. A key component of the algorithm is risk-oriented tail control. A separate "worst admissible'' SNR node is considered:

$$\gamma_{\text{risk}} = \gamma_{\text{eff}} - k_\sigma \sigma_{\text{est}} \quad (3)$$

where σ_{est} is the standard deviation of the SNR estimation error (RMS estimation error ≈2.1 dB), and k_σ is a quantile multiplier used for tail analysis ($k_\sigma \approx 2.05$ corresponds approximately to the 2nd percentile of the SNR distribution). At this node, the algorithm verifies whether the required on-time delivery probability (SLA $\rho_{\text{req}} \geq 98.5\%$) is preserved and whether the estimated probability of events with two or more retransmissions does not exceed the configured threshold ($\rho_{\geq 2} \approx 3\ \%$). Candidates that do not satisfy these conditions are discarded from further consideration. For the remaining candidates, a unified selection criterion is applied, giving preference to options with higher expected MOS while penalizing a high expected number of retransmissions and closeness to SLA boundaries. Mathematically, this criterion can be expressed as:

$$\text{Score(m)} = \text{E[MOS(m)]} - \lambda_{\text{retx}}\text{E[\#Retx(m)]} - \Lambda\big(\rho_{\text{req}} - \text{E}[\text{P}_{\text{on}}(\text{m})]\big) - \Lambda\Big(\rho_{\text{req}} - \text{P}_{\text{on}}^{\text{risk}}(\text{m})\Big) - \Lambda\big(\text{P}_{\text{Retx}\geq 2}^{\text{risk}}(\text{m}) - \rho_{\geq 2}\big) \quad (4)$$

where $\mathbf{MOS(m)}$ is the E-model estimate of speech quality for candidate m, incorporating the mean delay and jitter of on-time packets, as well as total losses (hard + late) adjusted for burstiness via BurstR; $\mathbf{E[\cdot]}$ denotes averaging across the three SNR uncertainty nodes; $\mathbf{\#Retx(m)}$ is the number of HARQ retransmissions (0, 1, 2,…); $\boldsymbol{\lambda_{\text{Retx}}}$ is the penalty weight for retransmissions; $\mathbf{P_{on}^{risk}(m)}$ is the on-time delivery probability at the risk node; and $\mathbf{P_{Retx\geq 2,}^{risk}(m)}$ is the probability of the event Retx ≥ 2 at the risk node.

In the case of equal Score(m) values, priority is given to the lower MCS index, which corresponds to a conservative policy aimed at reducing the risk of long HARQ chains.

On top of the look-ahead mechanism, an outer VoNR control loop operates, dynamically adjusting the offset Δ and stabilizing short-term quality indicators. After receiving a negative acknowledgment (NACK), the offset is immediately decreased, with a stronger reduction when events with two or more retransmissions are observed, which allows the algorithm to quickly damp overly aggressive decisions [7]. Within a sliding window of ≈0.8 seconds, the algorithm tracks key quality indicators: the fraction of late-loss packets, hard losses, and events with two or more

retransmissions. If any of these indicators exceeds its configured soft threshold (≈0.8%, 0.4%, and 3%, respectively), the offset Δ is reduced. If all indicators consistently remain within "excellent'' ranges (late-loss packets below 0.2%, Retx ≥ 2 below 2%), the offset Δ increases slowly. The offset is saturated within a safe interval and smoothed to prevent oscillations, thereby preserving the predictive behavior of the look-ahead mechanism.

As a result, the proposed algorithm exhibits predictable behavior under mobility conditions: the aging penalty prevents overestimating the relevance of the SNR estimate while the queue builds up, the risk-oriented check filters out overly aggressive MCS choices in unfavorable channel states, and the VoNR control loop keeps key quality indicators within tight bounds without sacrificing useful data rate. The computational complexity of the algorithm is suitable for real-time implementation: for each packet, only a few MCS candidates (typically 3-5), three SNR uncertainty nodes, and up to four HARQ rounds are evaluated. All calculations are based on analytical BLER curves and probability summations. The parameters of the risk-oriented checks – such as the required on-time delivery probability (98.5%), the SNR error amplitude, penalty weights, and thresholds for events with two or more retransmissions – were calibrated in the simulator and can be adapted to a specific radio-network profile.

Experimental Design and Statistics. The study is based on a blocked experimental design over random seeds with pairwise comparisons, which provides high statistical power and reduces the impact of environmental randomness. The primary blocking unit is the random number generator seed (Common Random Numbers, CRN), within which all algorithms experience the same channel trajectory, the same sequence of random HARQ events, and the same channel estimation errors. This Common Random Numbers approach reduces the variance of differences between algorithms and ensures objective comparison, since any performance gap is driven solely by differences in MCS selection logic.

The experimental factor space covers five target BLER levels: 0.01, 0.03, 0.05, 0.08, and 0.10. At each point, five MCS selection strategies are compared: the proposed risk-aware MOS-driven algorithm, the myopic HARQ-aware scheme, the HARQ-unaware variant, the classical OLLA algorithm, and a trivial random strategy. For each combination of algorithm and target BLER, twenty independent runs are executed (seeds 1-20). Each run consists of 4000 RTP packets, corresponding to ≈80 seconds of voice traffic with a packetization interval of 20 ms. The nominal VoNR scenario assumes five simultaneous users, a user speed of 60 km/h (representing urban/suburban mobility), a carrier frequency of 3.5 GHz (band n78), up to three HARQ retransmissions with a nominal inter-round spacing of 8 ms, and a delivery deadline of 32 ms. For these parameters, the Doppler shift is about 194 Hz, and the channel coherence time is estimated at 2.2 ms. The effective interval between HARQ rounds increases with the multiplexing of multiple users, reflecting the real cost of additional retransmission in terms of deadline [8]. The channel is modeled as a combination of slow variations, short-term correlation, and shadow fading. The BLER curves follow a logistic shape with monotonically improving performance as the MCS index increases. The HARQ combining gain is represented as an increase in effective SNR from round to round in incremental redundancy (IR) mode.

The unit of statistical analysis is the run-seed pair. At the packet level, we record events of successful delivery, deadline miss, and the number of HARQ retransmissions. These events are then aggregated within each seed separately for every scheme and every target BLER value. The fraction of on-time deliveries is defined as the fraction of packets that arrive before the deadline. Late deliveries are treated as losses, since for speech quality they are equivalent to the absence of useful signal at the playback instant. Hard losses are cases in which a packet is not recovered after the maximum number of retransmissions. The probability of events with two or more retransmissions (Retx ≥ 2) is considered separately as an indicator of the retransmission tail, which has the strongest impact on deadline violation risk. For packets that meet the deadline, the mean one-way delay and jitter are computed. These packets form the signal actually perceived by the user; therefore, analyzing their temporal variability is most relevant from the speech-quality perspective. The MOS is calculated using an adapted version of the E-model that accounts for loss burstiness (BurstR) and a reduced jitter

buffer of 10 ms, which increases sensitivity to short, bursty delay fluctuations characteristic of VoNR. For interpreting throughput, we additionally report the useful rate (goodput), defined as the codec bitrate multiplied by the fraction of packets without loss or lateness.

Aggregation of results at each point of the factor space is performed as the mean over twenty seeds with 95% confidence intervals. The intervals are constructed from the standard error of the mean under a normal approximation. Owing to the blocked design with common random numbers, these intervals remain narrow even for tail metrics such as Retx ≥ 2. All summary estimates are reported at the level of individual runs (seeds), which is consistent with the chosen observational unit and does not rely on an assumption of independence between packets within a single run.

Algorithm comparison is performed at the operating point with target BLER 0.05. Three hypotheses are formulated at this point: the proposed algorithm ensures a higher proportion of on-time deliveries, a higher MOS, and a lower proportion of events Retx ≥ 2 compared to the baseline schemes. The present study reports descriptive statistical estimates with confidence intervals, whereas formal hypothesis testing is not the primary focus. However, the seed-level primary data allow paired nonparametric tests to be carried out on the differences "proposed − baseline". For example, you can apply the Wilcoxon criterion with a two-tailed test, the median shift estimate according to Hodges–Lehmann, and the dimensionless effect size indicator $r = \frac{|Z|}{n}$, where Z is the normalized test statistic and n = 20 is the number of pairs. The value of r reflects the relative magnitude of the difference between algorithms: the larger r is, the stronger the systematic shift in favor of one of them, independent of sample size. These procedures are fully compatible with the chosen seed-based blocking and can be used for further analysis when the raw table is available. The experimental data are stored in CSV format, which facilitates reproducibility and additional analyses. The full summary table contains mean values and 95% confidence intervals for all algorithms and all target BLER levels. A separate subset is formed for the main algorithms used in the primary comparison. The seed-level primary data are provided as a raw dataset, enabling formal statistical testing and further investigations.

The sensitivity of the algorithms to changes in environmental conditions was evaluated separately. We considered variations in HARQ combining gain, in the variance of the SNR estimation error, and in the TTI duration, which determines the effective interval between HARQ rounds and, consequently, the cost of an additional retransmission in terms of meeting the delivery deadline. All scenarios were constructed using the same blocked protocol with common random numbers and the same aggregation procedure: means over 20 seeds with 95% confidence intervals. An important element of the study is the ability to tune the algorithm parameters to specific network conditions. The penalty weights for retransmissions, the threshold rules for controlling tail behavior, and the parameters of the outer VoNR loop can be adapted to user speed, the degree of multiplexing, and the required Quality of Service. This adaptability makes it possible to deliberately balance MOS, the fraction of on-time deliveries, and the frequency of Retx ≥ 2 events at different operating points. For completeness, channel and scheduling parameters were also varied: changes induced by user multiplexing, which affect the effective interval between HARQ rounds, as well as the levels of SNR estimation noise, which determine the accuracy of MCS selection and, ultimately, the achieved service quality. All scenarios were evaluated using the same key metrics – MOS, the fraction of on-time deliveries, Retx ≥ 2, and useful rate (goodput) – with the same observational unit, namely at the run–seed level [8].

Overall, the chosen experimental design and the consistent use of descriptive statistics (means over seeds and 95% confidence intervals) ensure reliable and reproducible results. The possibility of parametric tuning makes the proposed approach flexible and suitable for a wide range of scenarios, allowing not only to confirm the effectiveness of the algorithm under nominal conditions, but also to assess its robustness to environmental changes and to determine optimal settings for specific network configurations.

Results (Figures and Quantitative Findings). The study of MCS selection algorithms for VoNR in 5G SA has revealed substantial differences in the effectiveness of various approaches in

maintaining voice quality. The proposed risk-aware, MOS-driven MCS selection algorithm demonstrates a consistent advantage across the entire range of target BLER levels, BLER ∈ {0.01, 0.03, 0.05, 0.08, 0.10}. Owing to the use of a look-ahead mechanism that accounts for SNR estimation uncertainty and retransmission-related risks, the algorithm provides high Quality of Service even under challenging channel conditions. Under identical channel traces and HARQ event realizations (enabled by blocking with CRN), the proposed algorithm consistently satisfies the on-time delivery requirement. The fraction of on-time deliveries remains at ≈98.5-98.9%, with narrow 95% confidence intervals. This directly translates into end-user perceived quality: MOS stays at a high level of ≈4.26-4.27 without noticeable degradation when moving between different operating points. The retransmission tail at Retx ≥ 2 remains controlled and low (≈2.0-2.4%), and the fraction of late-loss packets is ≈1.5-1.7%. The useful data rate (goodput) for the proposed approach at fixed target BLER = 0.05 is approximately 12.2-12.4 kb/s, which is comparable to the best-performing baseline schemes.

The behavior of the baseline approaches differs markedly. The myopic HARQ-aware scheme gradually loses on-time delivery performance as target BLER increases: the on-time fraction drops to ≈96-97%, late-loss events grow (to ≈2.0-3.5%), and Retx ≥ 2 rises to ≈3.2-3.4% for HARQ-aware and ≈4-5.5% for HARQ-unaware. The HARQ-unaware scheme, which ignores the risk associated with future HARQ rounds, is even more aggressive in its MCS choices and, at higher target BLER levels, accumulates more late-loss deliveries. As a result, the on-time delivery fraction decreases to ≈95.5-96.0%, and MOS drops to ≈4.22-4.24.

The classical OLLA algorithm and the random strategy produce long HARQ chains and substantial late-loss rates: the on-time delivery fraction is only ≈78-83%, MOS is in the range of ≈3.96-4.02, and Retx ≥ 2 is ≈22-25%. This indicates that optimizing only the average BLER, or reacting locally to the retransmission tail, does not guarantee compliance with playout SLA constraints and a stable MOS in a mobile VoNR scenario. A slight increase in the mean delay among on-time packets for the proposed algorithm (≈14.7-14.8 ms versus ≈13.7-14.2 ms for the myopic baselines) is a natural consequence of correctly prioritizing the deadline. Borderline packets that baseline schemes effectively push into late-loss are, under the proposed algorithm, more likely to arrive by the deadline and thus enter the “on-time” sample, where they raise the mean. Since late packets are counted as losses in MOS evaluation, this strategy improves subjective quality without sacrificing goodput. The distributions of used MCS indices support this interpretation: the proposed approach avoids unnecessary excursions to the highest MCS indices, keeping most probability mass in the “operational middle’’ of the range. In contrast, the HARQ-unaware scheme more frequently selects very high MCS values, which increases HARQ chain length and the fraction of late-loss deliveries [6]. Taken together, these findings show that the combination of MOS-based look-ahead with barriers on on-time delivery and Retx ≥ 2 provides robust voice quality under mobility, keeping tail risks under control without losing throughput [10].

Table 1 presents a comparative analysis of the algorithms in terms of key Quality of Service metrics at the nominal operating point, where the target BLER is set to 0.05. This operating point was chosen because it is particularly informative for contrasting on-time delivery, MOS, the fraction of second- and higher-order retransmissions (Retx ≥ 2), and the useful data rate (goodput). The metrics are reported in the format “mean value ± half-width of the 95% confidence interval”, obtained from 20 independent simulation runs using CRN. The symbol “±” denotes the half-width of the confidence interval, characterizing variability around the mean. The mean delay reported in the table is computed only for packets that were successfully delivered within the deadline. The useful data rate (goodput) is defined as the codec bitrate multiplied by the fraction of packets that experienced neither loss nor lateness [10]. This reporting format provides a consistent basis for interpreting differences between algorithms in terms of voice-service quality in the mobile VoNR scenario.

Figure 1 shows how the aggregate MOS varies with the target BLER for the different MCS selection algorithms. The Proposed Algorithm achieves the highest subjective speech quality, maintaining MOS values ≈4.27 regardless of target BLER in the range 0.01–0.10. This indicates the stability of the algorithm and its ability to preserve high speech quality even as the block error rate increases. In contrast, the HARQ-aware and HARQ-unaware schemes exhibit a gradual decrease in

MOS as BLER grows, reflecting the accumulation of jitter and packet losses due to a less conservative MCS selection. For example, at target BLER = 0.1, MOS for HARQ-aware drops to ≈4.25, and for HARQ-unaware to ≈4.22, which may lead to a perceptible degradation of voice quality for end users.

Table 1

Comparison of algorithms in terms of QoS metrics at the nominal operating point (target BLER = 0.05)

Algorithm	MOS	On-time deliveries, %	Late-loss, %	Retx ≥ 2, %	Mean delay (on-time packets), ms	Useful rate (goodput), kbit/s
Proposed	4.27±0.01	98.7±0.1	1.3±0.1	2.3±0.1	14.75±0.05	12.48±0.06
HARQ-aware	4.25±0.01	96.6±0.2	2.6±0.1	3.0±0.2	13.70±0.05	12.27±0.06
HARQ-unaware	4.24±0.01	96.3±0.2	2.8±0.1	4.3±0.2	14.05±0.06	12.20±0.06
OLLA	3.98±0.01	80.6±0.3	13.6±0.3	24.0±0.2	15.50±0.05	10.15±0.1
Random	3.97±0.01	78.4±0.2	4.9±0.1	23.1±0.2	14.60±0.05	9.90±0.1

The classical OLLA loop and Random (a trivial scheme with random MCS selection) produce the lowest performance, keeping MOS below 4.0 because of large delay variability and an increased frequency of retransmissions [9]. This highlights their inability to adapt effectively to dynamic channel conditions. The stability of the Proposed Algorithm curve is explained by the fact that the optimization objective accounts for the expected MOS, an SNR-based risk quantile, and a penalty for the probability of Retx ≥ 2. This yields an optimal balance between reliability and modulation efficiency.

Even a seemingly small MOS improvement – for example, from 4.20 to 4.27 – has a substantial impact on voice service quality. On the E-model scale, a 0.07-point difference can correspond to a shift from the middle of the "Good" category toward its upper bound, or even into "Excellent" for some users. This is especially important for real-time services such as VoIP or VoNR, where stability and predictability of quality are critical for comfortable speech perception. Consequently, even when the target BLER is relaxed to 0.1, speech quality remains in the "Good" category (above 4.2 points on the E-model scale), which confirms the robustness of MOS-oriented control to channel fluctuations and the effectiveness of the proposed algorithm under realistic network conditions.

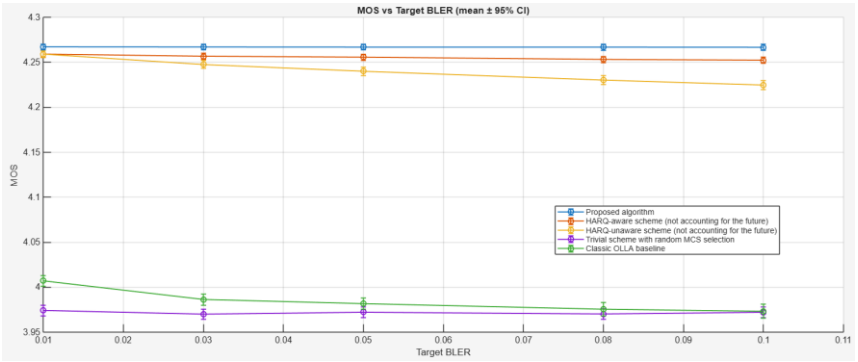

Fig. 1. MOS versus target BLER for different MCS selection algorithms

Figure 2 shows the dependence of the fraction of on-time delivered packets on the target BLER. The Proposed Algorithm achieves the highest fraction of on-time deliveries, consistently exceeding 98.5%, fully meeting the Req on-time requirement. This demonstrates the high reliability of the algorithm and its ability to maintain stable packet delivery even under varying block error rates. For comparison, the HARQ-aware and HARQ-unaware schemes exhibit a gradual decrease in the fraction of on-time deliveries as target BLER increases. For example, at target BLER = 0.1, the on-time delivery fraction for HARQ-aware drops to ≈97%, and for HARQ-unaware to ≈96%. This

degradation is caused by the lack of look-ahead and risk penalties, which leads to delay fluctuations due to HARQ retransmissions.

The classical OLLA loop and Random (a trivial scheme with random MCS selection) perform significantly worse, keeping the fraction of on-time delivered packets at only ≈78-80%. This indicates their inability to adapt effectively to dynamic channel conditions, resulting in substantial delays and packet losses.

The strong stability of the Proposed Algorithm curve is due to the optimization objective, which includes a barrier on the probability of deadline violation and an enhanced penalty on a risk SNR quantile (CVaR). This makes it possible to effectively suppress delay tails even under harsh radio conditions, thereby ensuring high QoS. Consequently, even when the target BLER is increased to 0.1, the fraction of on-time deliveries remains above 98.5%, confirming the robustness of the algorithm to channel fluctuations and its effectiveness in realistic networks.

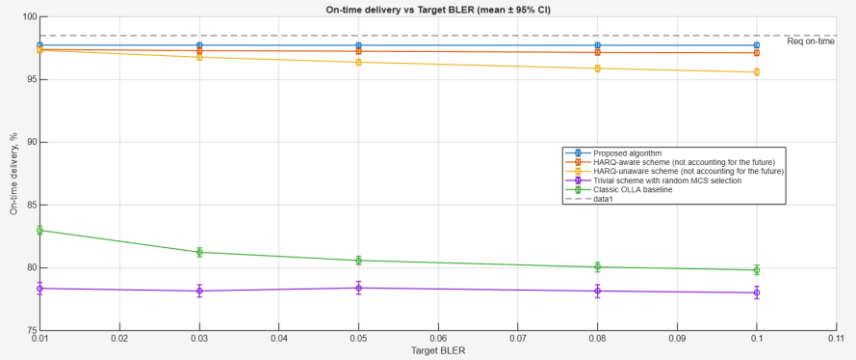

Fig. 2. On-time delivery (fraction of packets delivered by the deadline) versus target BLER

Figure 3 shows how the fraction of late-loss packets (packets that arrive after the 32 ms deadline) varies with the target BLER. The Proposed Algorithm exhibits exceptional performance, maintaining a minimal late-loss fraction of ≈1.5% regardless of the target BLER value in the range 0.01-0.10. This is a factor of about 1.5-2 better than the HARQ-aware and HARQ-unaware algorithms, which show late-loss fractions in the range of 2.5-3.5%. This advantage confirms the effectiveness of the built-in VoNR control loop, which dynamically adjusts the OLLA offset and regulates the MCS based on the actual queue state and retransmission behavior. As a result, it prevents delay accumulation and reduces the probability of missing the deadline.

For the classical OLLA loop, a substantial increase in late-loss is observed, up to ≈13-14% as BLER grows. This indicates the limited effectiveness of one-step adaptation without MOS feedback, which leads to significant delays and packet losses. The trivial scheme with random MCS selection also exhibits high late-loss values – around 5%, which further confirms its inability to react effectively to changing channel conditions.

The obtained results highlight the robustness of the Proposed Algorithm to channel errors. The algorithm maintains a balance between low delay and stable QoE at a level above 4.25 MOS, while simultaneously reducing late-loss to ≈1%. This demonstrates the high effectiveness of the algorithm in real networks, where channel fluctuations can significantly affect service quality. Thus, the Proposed Algorithm not only delivers high-quality voice transmission but also preserves system stability even under substantial channel variations, which is critically important for modern communication systems such as VoNR [10].

Figure 4 shows how the fraction of second- and higher-order retransmissions ("Retx"≥2) depends on the target BLER for different MCS selection algorithms. The Proposed Algorithm achieves the lowest fraction of events with two or more retransmissions, maintaining values at ≈2.0-2.3% across the entire target BLER range from 0.01 to 0.10. This indicates the effectiveness of the algorithm in preventing serial retransmissions, which can lead to substantial delays and degradation of QoE. For comparison, the HARQ-aware and HARQ-unaware schemes exhibit a gradual increase in the fraction of second- and higher-order retransmissions as target BLER grows. For example, at target BLER = 0.1, the fraction of "Retx"≥2 events for HARQ-aware reaches ≈3.2-3.4%, and for HARQ-unaware

≈4.0-5.5%. This increase is caused by the absence of explicit risk prediction for retransmissions and insufficient control of the distribution tails, which results in delay accumulation and a higher probability of serial retransmissions.

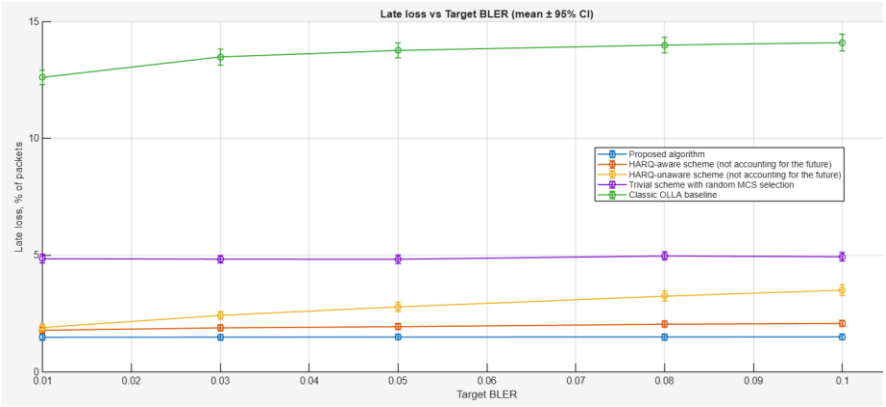

Fig. 3. Late-loss fraction versus target BLER

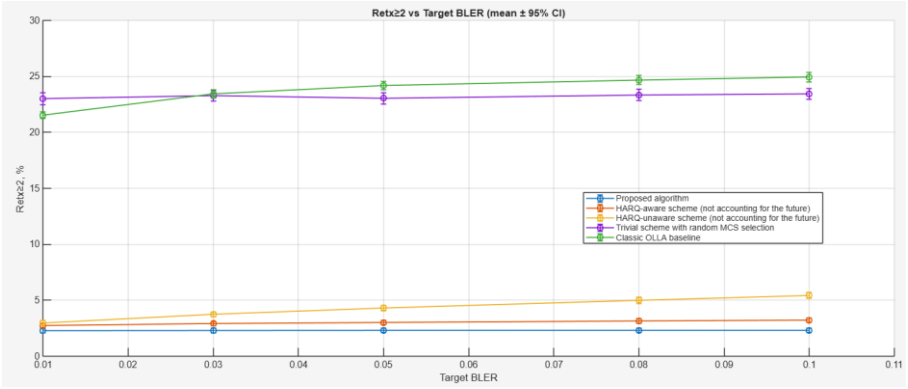

Fig. 4. Fraction of Retx ≥ 2 retransmissions versus target BLER

The classical OLLA loop and Random (a trivial scheme with random MCS selection) perform significantly worse, keeping the fraction of second- and higher-order retransmissions at ≈22–25%. This demonstrates their inability to control retransmission risk effectively and to adapt to dynamic channel conditions, which in turn leads to substantial delays and packet losses.

The effectiveness of the Proposed Algorithm in controlling second- and higher-order retransmissions is driven by its risk-aware design, which accounts for SNR estimation uncertainty and anticipates the probability of serial retransmission events. The algorithm imposes barrier constraints on the probability of Retx ≥ 2 events, allowing it to effectively suppress the tails of the retransmission distribution even under challenging radio conditions. Consequently, even when the target BLER is increased to 0.1, the fraction of second- and higher-order retransmissions for the Proposed Algorithm remains at ≈2.0-2.3%, which confirms the robustness of the algorithm to channel fluctuations and its effectiveness in real-world networks. This is particularly important for real-time services such as VoNR, where stability and predictability of quality are critical for comfortable speech perception.

Conclusions

The presented study demonstrates that in 5G Standalone networks with VoNR support, classical MCS algorithms fail to ensure sufficient QoS stability due to the lack of serial retransmission risk prediction, insensitivity to RTP delivery timing, and the absence of a direct MOS-driven optimization target. The proposed algorithm overcomes these limitations by introducing a MOS-oriented objective function, CVaR-like risk barriers for Retx ≥ 2 events, and a look-ahead evaluation mechanism that accounts for SNR uncertainty. Additionally, an external VoNR control loop maintains service indicators within SLA thresholds even under high channel variability.

Simulation results confirm the superiority of the approach: across the entire BLER range ∈[0.01-0.10], the algorithm sustains MOS≈4.26-4.27, on-time delivery above 98.5%, and lowers the probability of Retx ≥ 2 to 2-2.3%, while matching baseline schemes in terms of goodput. In contrast, OLLA and Random produce second-and-higher order HARQ chains of 22-25%, with MOS dropping below 4.0. The proposed mechanism therefore delivers predictable conversational quality under mobility conditions (60 km/h), which is particularly critical for real-time voice services. These results show that prioritizing MOS and controlling tail risks is a more effective strategy than BLER-only optimization or single-step reactive adaptation. Thus, the work establishes a robust algorithmic foundation for VoNR-quality solutions in future commercial 5G deployments and sets a direction for further research in ML-driven MCS optimization, interpretable QoE metrics, and energy-aware HARQ control policies.

References

1. 3GPP TS 38.300 version 17.2.0 Release 17. NR; NR and NG-RAN Overall Description. Effective from 2023-05-18. Official edition. FRANCE: 650 Route des Lucioles, F-06921 Sophia Antipolis Cedex, 2023. 152 p.

2. Ветошко І.П., Кравчук С.О. Пріоритизація голосового трафіку в 5G: роль планувальників у забезпеченні QoS. XIX Міжнародна науково-технічна конференція «Перспективи телекомунікацій 2025», Київ, 14–18 квітня 2025, с. 189–192.

3. 3GPP TS 38.331 version 17.6.0 Release 17. NR; Radio Resource Control (RRC) Protocol specification. Effective from 2024-02-11. Official edition. FRANCE: 650 Route des Lucioles, F-06921 Sophia Antipolis Cedex, 2024. 421 p.

4. Vetoshko I.P., Kravchuk S.O. Opportunities to Improve the Quality of Voice Services in 5G Networks // 2023 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), ISBN: 979-8-3503-4848-4, 13-18 November 2023, Kyiv, Ukraine. https://doi.org/10.1109/UkrMiCo61577.2023.10380376.

5. 3GPP TS 23.501 version 16.7.0 Release 16. System architecture for the 5G System (5GS). Effective from 2021-01-21. FRANCE: Sophia Antipolis Cedex, 2021. 451 p.

6. 3GPP TR 38.802 V14.2.0 Release 14. Study on new radio (NR) access technology. Effective from 2017-09-28. Official edition. FRANCE: Sophia Antipolis Cedex, 2017. 142 p.

7. Vetoshko I.P., Kravchuk S.O. Possibilities of improving the voice services quality in 5G networks // Information and Telecommunication Sciences. – 2023. – Vol.14, No 2. – P. 9-16, https://doi.org/10.20535/2411-2976.22023.9-16

8. Kwon J., Park J., Rhee W. A QoS-aware adaptive resource allocation scheme using CQI and MCS for LTE-A networks. *Wireless Personal Communications*. – 2014. – Vol. 76. – P. 193–207.

9. Vora P. Voice over 5G (Vo5G): Technical Challenges and Performance Evaluation. *IEEE Communications Magazine*. 2021. Vol. 59, No. 10. P. 16–22.

10. Vetoshko I. P., Kravchuk S. O. Integration of machine learning-based prediction and dynamic QoS optimization into adaptive VONR scheduling in 5G standalone networks: a simulation-based approach. Scientific notes of Taurida National V.I. Vernadsky University. Series: Technical Sciences. 2025. Vol. 1, no. 3. P. 43–56. URL: https://doi.org/10.32782/2663-5941/2025.3.1/07

Автори статті

Ветошко Іван – аспірант, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

ORCID: 0000-0002-0009-7610

Кравчук Сергій – доктор технічних наук, професор, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

ORCID: 0000-0002-4118-0226

Authors of the article

Vetoshko Ivan – postgraduate, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0000-0002-0009-7610

Kravchuk Serhiy – Doctor of Sciences (technical), Professor, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0000-0002-4118-0226