Early HARQ using LLR Trend Analysis

Narayan Prasad Kusi
Department of Integrated IT Engineering
Seoul National Univeristy of Science and Technology, Seoul
narayankusi@seoultech.ac.kr

Jiho Kim
Deptartment of IT and Media Engineering
Seoul National Univeristy of Science and Technology, Seoul
kjh2208@seoultech.ac.kr

Prof. Dong Ho Kim*
Department of Integrated IT Engineering
Seoul National Univeristy of Science and Technology,Seoul,
dongho.kim@seoultech.ac.kr

***Abstract*— One of the key usage scenarios in the scope of 5G is ultra-reliable and low-latency communications (URLLC) and HARQ is one of the inherent parts to make the service reliable and efficient. In this work, we proposed a new and reliable metric for the subcode based early HARQ prediction which adopts NMS (Normalized Min Sum) LDPC decoder. It provides flexible iterative decoding with sub-codes of different lengths for early HARQ prediction using the substructure LDPC parity matrix. The proposed Early HARQ prediction analyses the changing trend of LLR values of variable nodes during iterations and predicts feedback of the decodability based on the trend analysis of posterior LLR values after some iterations. The early HARQ prediction using the LLR trend analysis enables us to provide reliable and earlier feedback making faster retransmissions.**

I. INTRODUCTION

With the emergence of 5G NR, higher data rates and low latency services have become key demands from users. Latency and data rate are crucial factors in the context of 5G technology, offering significant advancements over previous generations of wireless communication. The conflicting requirement of 1 millisecond end-to-end latency and 99.999% reliability for URLLC have major impacts, especially on the hybrid automatic repeat request (HARQ) which is used in current wireless standards [1].

HARQ plays major roles in maintaining end-to-end reliability with acceptable latency. HARQ is a physical/MAC layer mechanism that uses feedback to achieve transmission robustness by providing retransmissions based on feedback (ACK, NACK) while maintaining a target BLER[2]. During data transmission, if decoding fails at the receiving end and retransmission becomes necessary, additional latency is incurred for successful data transmission. The retransmitted data is combined with the previous transmission by utilizing the chase combining (CC) or incremental redundancy (IR) [3]. In particular, the HARQ round-trip time (RTT), which is the time interval between reception of the initial transmission and the retransmission, causes a bottleneck in URLLC delivery for low latency services[1]. Reactive HARQ suffers latency issue, while proactive HARQ suffers spectral efficiency and unnecessary retransmission.

II. EARLY HARQ FEEDBACK

The latency in the HARQ depends on the data processing and transmission time at the transmitter denoted by TTI, propagation delay, processing at the receiver, feedback generation, feedback transmission and reception at the transmitter. In case of retransmission, same processes are repeated, and additional delay is added in RTT until the transmission of a transport block is successful. RTT consists of following delay components delay [1]:

$$RTT = \tau + T_{TTI} + T_{LLR} + T_{FB} + T_{A/N} + T_{Tx} \quad (1)$$

Here, "τ" is the propagation delay, T_{TTI} is the transmission time interval, T_{LLR} is the processing time to calculate log likelihood ratio (LLR) from received signal, T_{FB} is feedback generation time, $T_{A/N}$ is the transmission time for ACK/NACK feedback and T_{Tx} is processing time of the feedback at the transmitter. The processing time of the feedback T_{Tx} is assumed fixed and infinitesimal as it contains only a few bits of feedback. So, the RTT can be reduced by shortening the decoding time and providing the feedback or by shortening the feedback generation time T_{FB} as a prediction of codeword decodability.

III. RELATED WORKS TO REDUCE FEEDBACK TIME T_{FB}

Several methods have been proposed to reduce feedback generation time T_{FB} by replacing reactive HARQ with proactive HARQ in the feedback chain. For channel estimation based HARQ prediction, channel estimation is performed based on a reference signal such as (DMRS). A method for receiving and accumulating the quantized received SNR and applying a threshold to predict an ACK or NACK was proposed, where the threshold controls the balance between false positive and false negative errors [4].

The mixture of proactive and reactive HARQ protocol to reduce the latency has been studied in [4]. In [5] and [6], Bit Error Rate (BER) estimate based on received LLRs is used to predict the decoding outcome ahead of the actual decoding. They empirically computed a threshold for the BER estimate to predict the decodability. A hybrid solution was proposed in [7]. They combined both LLR and channel-estimation-based feature using logistic regression to decide that received codewords is decodable or not. The proposed logistic regression showed a significant enhancement over other approaches that use only one of both. The authors in [1], [8] and [3] proposed a feedback prediction mechanism that observes the partial decoding behavior of so-called subcodes. Similar to the LLR-based feedback prediction, this approach uses the LLRs as a basis and the posterior bit error rate is calculated after some iteration of the LLRs to generate feedback prediction. However, in contrast to these, subcode-based schemes apply additional processing on the LLRs based on the knowledge of the code structure in [9]. In [2], the authors applied machine learning techniques, i.e. logistic regression, random forests, isolation forests and supervised autoencoders, on the LLR and subcode features.

IV. PROPOSED MODEL

A. Solution Model

We propose a new metric for Subcode-based Early HARQ prediction based on the LLR trend analysis approach. The HARQ system utilizes subcode to predict the decodability of the codework. In this work, LDPC codes are encoded in BG2 parity matrix [10]. Here, *n* represents the code word length, *k* is the number message bits and *m = (n - k)* is the rows of the parity check matrix. The *n* and *m* are different depending on the substructure of the matrix used to generate codeword. The proposed NMS LDPC decoding network uses normalization factor $\alpha = 0.66$ to compensate the performance loss due to min-sum approach in place of Belief Propagation algorithm.

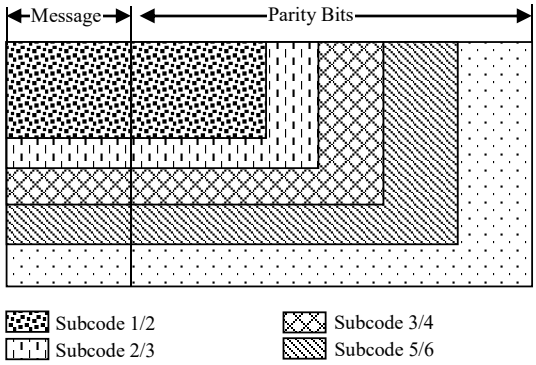

Fig. 1: Illustration of use of Structure of BG2 parity check matrix for different subcode lengths.

Let $L(b_i)$ denote the received channel log-likelihood ratio (LLR) corresponding to i-th bit, and the CN-to-VN messages $L(r_{ji})$ and the VN-to-CN messages $L(q_{ij})$ are propagated and updated iteratively. LLR $L(b_i)$ is given by

$$L(qij) = L(b_i) = log\left(\frac{P(b_i = 0|y)}{p(b_i = 1|y)}\right) = \frac{2y}{\sigma^2} \quad (2)$$

where $L(q_{ij})$ denotes the belief information passing from variable i to its connected check nodes j and σ^2 is the noise variance of AWGN.

$$L^{(l)}(r_{ji}) = \prod_{i' \in vj \setminus i} sgn\left(L^{(l-1)}(q_{i'j})\right) \min_{i' \in Vj \setminus i} \left| L^{(l-1)}(q_{i'j}) \right| \quad (3)$$

Here, the $L^{(l)}(r_{ji})$ is the message sent from j-th check node to i -th variable node message where l is the iteration number and α is the correction factor. $V_{j\setminus i}$ represents neighboring V_j excluding variable node V_i and $C_{i\setminus j}$ represents neighboring C_i excluding check nodes C_j .

$$\Lambda_i = L^{(l)}(q_{ij}) = L(b_i) + \sum_{j' \in C_i \setminus j} \left(L^{(l)}(r_{j'i}) \right) \quad (4)$$

The $L^{(l)}(q_{ij})$ denotes the posterior LLR after l iterations of the variable node information. The LLR values after belief propagation between check nodes and variables nodes are updated in each iteration.

For decodable frames, the trend of the LLR values for each variable node after a few iterations is consistent towards, positive or negative value. In our proposed method, the trend of LLR values is used to predict whether a codeword is decodable or undecodable. For this, the change of direction of posterior LLR values are analyzed and is given by:

$$\Delta L_i^{(l)} = \frac{\partial \Lambda_i^{(l)}}{\partial l} \; for \; l = l_p, l_{p-1} \quad (5)$$

$$\delta_i^l = \begin{cases} +1 \; for \; \Delta L_i^{(l)} \geq 0 \\ -1 \; for \; \Delta L_i^{(l)} < 0 \end{cases} for \; l = l_p, l_{p-1} \quad (6)$$

$$Trend_i = \left(\delta_i^{(l_p)} + \delta_i^{(l_{p-1})} \right) \; for \; i = 1 \; to \; k \quad (7)$$

$$Zero\_Count = k - \frac{1}{2} \sum_i^k |Trend_i| \quad (8)$$

The change of the LLR ΔL for every bit i is calculated on iteration l_{p-1} and iteration l_p. The HARQ feedback prediction is performed after iteration $l = l_p$. In our study, the prediction is done after 5th iterations i.e. $l_p = 5$ and ΔL is calculated in the stage of 3rd-4th and 4th -5th iteration stage. If the SNR is large enough to be decodable by iterative decoding, then after a few iterations, the LLR value should consistently go in either the positive or negative direction, and value of $Trend_i$ in (7) of every bit is non-zero. For any value of variable node i, $Trend_i$ =0 indicates the rise-fall in LLR values during the iterations. So, the probability of the decodable code word is given by:

$$FB_{trend} = \begin{cases} ACK & if \;\; \text{Zero\_Count} \leq th_{trend} \\ NACK & if \;\; \text{Zero\_Count} > th_{trend} \end{cases}. \quad (9)$$

The th_{trend} is a threshold related to how much the LLR of the k variable nodes can be allowed to fluctuate so that the decoder can determine only after l_p iterations whether it will eventually decode successfully after maximum number of iterations. The choice of threshold significantly affects the performance of Early HARQ schemes. The threshold for a false negative rate fn=0.05 over the SNR is shown in figure 2. The threshold values for the SC-HARQ schemes decreases faster than the threshold for the full code scheme used for the prediction. As the SNR increases, the threshold decreases.

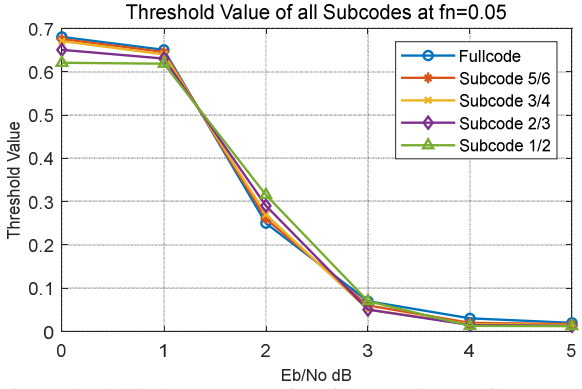

Fig. 2: Threshold values over SNR for a false negative rate (fn) =0.05

Early HARQ enables receiver station to provide feedback at the earlier stage. However, the reduced computational complexity, limited iterations and transmitting only some parts of SC E-HARQ comes at the cost of false positive and false negative predictions. The false positive prediction is worse and corresponds to the increased latency due to retransmission after complete decoding process. Therefore, selecting the appropriate thresholds is critical to the performance of E-HARQ and the operating range should not exceed the threshold values for corresponding E_b/N_0 values.

B. Latency Analysis of Early HARQ:

According to [1], the latency of the HARQ system is categorized to T_1 and T_2, where T_1 is the time from the start of the subcode transmission to feedback calculation, and T_2 is the sum of processing time, feedback transmission time, and the time until the entire transmission is received. Since data transmission is not complete without acknowledgement of the data, we modified the time required to send feedback to the transmitter to include in both T_1 and T_2, given by

$$T_1 = 2\tau + rsc.T_{TTI} + T_{LLR} + T_{FB} + T_{A/N} \quad (10)$$

$$T_2 = 2\tau + rsc.T_{TTI} + T_{LLR} + T_{FB} + T_{A/N} \quad (11)$$

where *rsc* is the ratio of subcode size and the total transmission size.

In 3GPP Release 16, 120kHz is the maximum SCS (sub-carrier spacing) allowed for data. Hence, the lowest

achievable slot duration (TTI) is 125µs. We use 60KHz SCS for the TTI calculation, which corresponds to 250 µs as per the 3GPP Standard. $T_{A/N}$ is assumed 25% of TTI. In [1], the decoding latency of a min-sum LDPC decoder is given by:

$$T_{FB}, LDPC = \frac{N \cdot d_v}{Z \cdot f_c} I \quad (12)$$

T_{FB} depends on the algorithm of the LDPC decoding, where N is the number of variable nodes; d_v is the average variable node degree; Z is the lifting factor; f_c is the clock frequency assumed to be 1GHz and I is the number of Iterations. The latency is calculated based on the assumptions presented in Table I.

Table I: RTT Timing Assumptions for Latency Evaluation

Timing Parameter	Regular HARQ	SC (1/2)	SC (2/3)	SC (3/4)	SC (5/6)
τ(1Km)	0.003	0.003	0.003	0.003	0.003
rsc	1.000	0.500	0.667	0.750	0.833
T_{TTI}	0.250	0.250	0.250	0.250	0.250
T_{LLR}	0.100	0.100	0.100	0.100	0.100
T_{FB}	0.004	0.000	0.001	0.001	0.001
$T_{A/N}$	0.063	0.063	0.063	0.063	0.063
T_1(ms)	0.422	0.294	0.336	0.357	0.378
T_2(ms)	0.422	0.294	0.336	0.357	0.378
Total delay with retransmission	0.845	0.588	0.672	0.713	0.755
Delay %	100.00%	69.59%	79.49%	84.44%	89.40%

V. PERFORMANCE EVALUATION

Based on the latency calculation, the URLLC service with timing constraint of 1ms for RTT allows at most one retransmission for all subcodes except for subcode ½, if the codeword is not decodable. The subcode ½ is allowed to retransmit up to 2 times upon failure within 1ms interval. The fig 3. shows the effective BLER performance for *fn*=0.05 with a single retransmission when predicting NACK and decoding failure, along with chase combining at receiver. The observed BLER performance using the subcode 5/6 is better than subcode SC- 3/4, SC-2/3 and SC 1/2 and is comparable with the full code performance at 20th iteration.

Table II: Simulation Parameters

Transport Block	360
Channel Code	LDPC
Parity check matrix, Lifting Size	BG2, Z=36
Check constrains for prediction	1-576(1/2), 1-828 (2/3), 1-1008(3/4), 1-1224(5/6)
Modulation order	BPSK
Decoder Type	Normalized Min-Sum
Decoding Iterations	20
Iterations for estimation	5
Max. false negative	0.05

The effective BLER in Figure 3 shows that the proposed scheme provides a reliability of BLER 10^{-5} at Eb/No =3.5 dB for subcodes 5/6 and at around Eb/No =3.8 dB for other subcodes. Therefore, the tactile internet service including URLLC can be operated with this approach for latency efficient and reliable communication.

VI. CONCLUSION

In this paper, we proposed a new simple and reliable Early HARQ prediction metric based on a posterior LLR trend analysis with threshold calculations for the range of operations. The latency calculations meet the constraint of 1ms RTT with one time retransmission for URLLC Service. The effective BLER performance shows that it can provide the reliability with effective BLER 10^{-5} at Eb/No =3.5 dB for subcodes 5/6 and around Eb/No =3.8 dB for other subcodes which is suitable for URLLC service that requires both low latency and reliability.

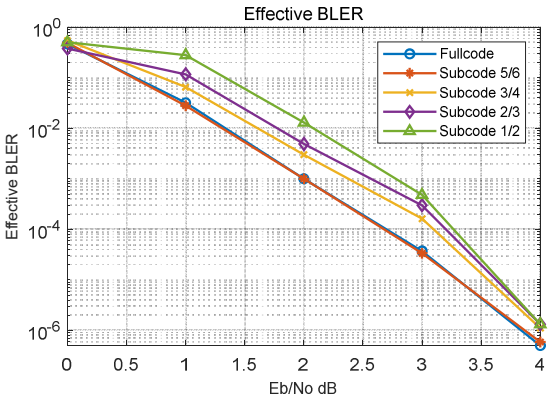

Fig.3: Achieved Effective BLER Over SNR for a false negative fn=0.05

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00368-003, Development of the 6G Service Targeted AI/ML-based autonomous-Regulating Medium Access Control (6G STAR-MAC), 50%) and (No. RS-2023-00229330, 3D Digital Media Streaming Service Technology, 50%).

REFERENCES

[1] Göktepe, Baris & Faehse, Stephan & Thiele, Lars & Schierl, Thomas & Hellge, Cornelius. (2018). Subcode-Based Early HARQ for 5G. 1-6. 10.1109/ICCW.2018.8403491.

[2] Strodthoff, N., Göktepe, B., Schierl, T., Hellge, C., & Samek, W. (2019). Enhanced machine learning techniques for early HARQ feedback prediction in 5G. IEEE Journal on Selected Areas in Communications, 37(11), 2573-2587.

[3] Wu, Yue, and Hongwen Yang. "Optimising energy efficiency of LDPC coded chase combining HARQ system." Electronics Letters 51, no. 6 (2015): 490-492.

[4] Makki, B., Svensson, T., Caire, G., & Zorzi, M. (2018). Fast HARQ over finite blocklength codes: A technique for low-latency reliable communication. IEEE Transactions on Wireless Communications, 18(1), 194-209.

[5] Berardinelli, G., Khosravirad, S. R., Pedersen, K. I., Frederiksen, F., & Mogensen, P. (2016, May). Enabling early HARQ feedback in 5G networks. In 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring) (pp. 1-5). IEEE.

[6] Berardinelli, G., Khosravirad, S.R., Pedersen, K.I., Frederiksen, F. and Mogensen, P., 2016, September. On the benefits of early HARQ feedback with non-ideal prediction in 5G networks. In 2016 International Symposium on Wireless Communication Systems (ISWCS) (pp. 11-15). IEEE.

[7] AlMarshed, Saleh, Dionysia Triantafyllopoulou, and Klaus Moessner. "Supervised learning for enhanced early HARQ feedback prediction in URLLC." 2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat). IEEE, 2020.

[8] Göktepe, B., Rykova, T., Fehrenbach, T., Schierl, T., & Hellge, C. (2020, December). Feedback prediction for proactive harq in the context of industrial internet of things. In GLOBECOM 2020-2020 IEEE Global Communications Conference (pp. 1-7). IEEE.

[9] Göktepe, B., Hellge, C., Schierl, T., & Stanczak, S. (2023). Distributed Machine-Learning for Early HARQ Feedback Prediction in Cloud RANs. IEEE Transactions on Wireless Communications.

[10] WF on LDPC Parity Check Matrices, document R1-1711982, 3GPP TSG RAN WG1 NR AH #2, Qingdao, China, Jun. 2017. [Online]. Available:http://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_AH/NR_AH_1706/Docs/R1-1711982.zip