MU-HARQ Prediction using Deep Learning and Optimization for URLLC

Narayan Prasad Kusi, Abdulvosit, Dong Ho Kim*

Seoultech Univ., Seoultech Univ., *Seoultech Univ.

narayankusi@seoultech.ac.kr, abdulbosit@seoultech.ac.kr, *Dongho.kim@seoultech.ac.kr

URLLC 를 위한 딥러닝 및 최적화를 이용한 MU-HARQ 예측

나라얀쿠시, 압둘보싯, 김동호*

서울과학기술대학교, 서울과학기술대학교, *서울과학기술대학교

Abstract

This paper introduces the Multiuser-HARQ based on the adaptive HARQ prediction model in uplink scenario and allocates the resources to the users for HARQ retransmissions based on the decoding reliability prediction with an objective to meet the latency requirement for URLLC and latency sensitive applications.

Ⅰ. Introduction

In the uplink of an OFDMA system, users are allocated with orthogonal resource elements (REs), effectively mitigating the intra-cell interference. As a result, the received signal-to-noise ratio (SNR) at the base station primarily depends on the user's transmission power, path loss between the UE and the base station, frequency, and the effects of channel fading. Thus, one of the major factors in decoding reliability of the codeword is the received SNR of the signal. In the multiuser HARQ, the number of retransmissions and resources required for successful decoding of the codewords depends on the channel environment and user distributions in the cell. For URLLC, it is essential to implement hybrid automatic repeat request (HARQ) approach along with advanced error correction techniques, but it inevitably increases latency. To minimize latency in multiuser-HARQ [1], it is proposed with a group-based reallocation method where users cooperatively utilize IR-HARQ feedback to share resources without collisions. The scheme considers sequential RV transmissions upon decoding failure and shares the available bandwidth among the users. To utilize the feedback idle time in MU-HARQ, PL-HARQ and PL-MU-HARQ are introduced in [1]. In these schemes, UEs transmit sequentially in a pipelined manner, each occupying the full bandwidth during its turn. Additionally, users are grouped into subgroups, enabling full-bandwidth transmission per subgroup and pipelined decoding across groups for M transmission times.

Ⅱ. System Model

We assume N uplink users are synchronized and transmit with RV0 simultaneously in the same slot time. For each i_{th} UE, the number of retransmissions $x_i \in \{0,1,2,3\}$ is modeled as a random variable, with the set represented by vector $X = \{x_1, x_2, \ldots, x_N\}$. Building on the deep learning-based HARQ prediction approach in [2], we propose a multi-user HARQ framework such that the BS predicts the required number of RVs for each UEs. The receiver (BS) uses features [2] such as estimated BER, LLR trends, fluctuation, squared LLRs, estimated SNR, and soft CSI—to predict decoding reliability class among the classes {0, 1, 2, 3} which depicts the number of RV retransmissions required. The intelligent predictions assist BS to estimate and perform efficient resource optimization and scheduling for HARQ retransmissions to meet strict URLLC latency constraints. The DL based prediction model estimates x_i per user, and the base station allocates time-frequency resources across slots or across the bandwidth for one time retransmission of RVs. The initial goal is to minimize totals slots N_s for N UEs. The total predicted resource required, in terms of REs, for N UEs is given by $R_{pred} = \sum_{i=1}^{N} Lx_i$ where L denotes the number of REs required for an RV. We assume total subcarriers, $S_f = 12N_{RB}$ where N_{RB} is total resource blocks and each UE is assigned initially with U_{RB} resource blocks for the RV0 transmissions. Subsequent RV transmissions are dynamically scheduled by the base station (BS) based on the prediction and available resources. In this framework, the DCI should be extended to include a decoding reliability class, indicating the predicted number of RVs from the set {0,2,3,1} along with resource assignment information as feedback to the UEs. The decodability of the codeword depends on received signal quality, which is influenced by location of UEs, log-normal shadowing, and small-scale fading. We use CDL-C to model the channel impairment by fading in link level simulations [2]. The link budget is calculated based on the user distance, d, and model log-normal shadowing as $X\sigma \sim$

$\mathcal{N}(0,\sigma^2)$[dB]. We perform simulations with shadowing variance of 1, 3 and 6 dB to represent different degrees of shadow fading, corresponding to user dispersion in various clutter environments.

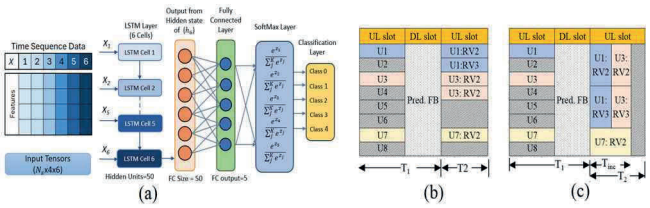

Fig 1(a) HARQ prediction Model, (b) Slot wise and (c) incremental resource allocation during retransmission

We consider the system with fixed radio resources of N_{RB} to be shared among the UEs for uplink. RV retransmissions are dynamically scheduled in both time and frequency domains. If it is scheduled to perform retransmission for N_s uplink slots, the total available resources are $T_{Res} = N_s S_f T_s$, where T_s is the number of OFDM symbols per slot. Since RV requirements differ across users, slot-wise resource allocation as shown in figure 1(b) is inefficient. To reduce resource wastage, we propose to assign symbol period wise incremental resource allocation for retransmission across the bandwidth as shown in figure 1(c), ultimately minimizing the HARQ latency.

$$\begin{aligned} T_1 &= TTI + 2\tau + T_{LLR} + T_{FB} + T_{AN} + \delta \\ T_{inc} &= \eta(TTI_{sym}) + 2\tau + T_{LLR} + T_{dec} + T_{AN} \\ T_2 &= N_s TTI + 2\tau + T_{LLR} + T_{dec} + T_{AN} \end{aligned} \quad (1)$$

Here, $T_{RTT} = T1 + T2$ and $T_{eRTT} = T1 + T_{inc}$ represent the effective round-trip time (RTT) for MU-HARQ under slot-wise and incremental resource allocation strategies for RV transmissions, respectively. T1 accounts for the initial RV0 transmission and feedback processing. T_{inc} denotes the additional delay in incremental resource allocation, while T_2 represents the latency incurred during slot-wise allocation for retransmissions across N_s uplink slots, where $N_s = \left\lceil \frac{R_{pred}}{T_s} \right\rceil$ and $\eta = \left\lceil \frac{R_{pred}}{S_f} \right\rceil$ is optimum symbols period required. To satisfy the URLLC latency constraint $T_{RTT} \leq 1ms$, parameters such as $SCS = 60kHz$, $T_{FB} = 0.0006ms$, $\tau = 0.003\,ms$, $T_{dec} = 0.0032\,ms$, $T_{LLR} = 0.1\,ms$, $T_{AN} = 0.0179ms$ and $TTI = 0.25\,ms$ and the model prediction time $\delta = 0.015ms$ are used. This yields $T_1 = 0.3895\,ms$ Through simple numerical search, we find that $\eta \leq 28$ i.e. $N_s \leq 2$ to meet the $1\,ms$ constraint. Under strong fading environments, many UEs are predicted to require additional RVs for reliable decoding, which may exceed the available resources ($R_{pred} > T_{Res}$). In such cases, MU-HARQ cannot satisfy reliability and latency simultaneously. To address resource limitations, we propose increasing the retransmission power for RVs for selected UEs, so lesser retransmissions are sufficient, thereby improving the SNR after soft combining with RVs in IR-HARQ which assists to meet latency constraint.

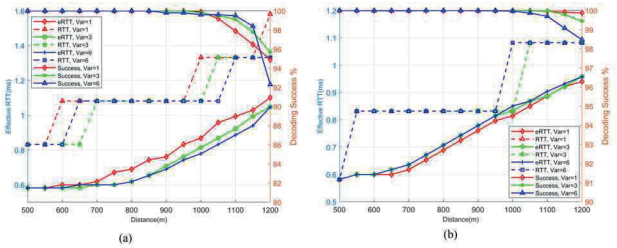

Fig. 3: Effective latency and success (a)based on HARQ Prediction (b) based on HARQ prediction with increased power in retransmission

The simulation result illustrates the performance of the overall decoding success rate, effective RTT versus distance of UEs from the base station, under different shadowing variances, according to DL based HARQ prediction are presented in Fig. 3(a). These results assume a nominal transmit power of 16 dBm and allow up to 4 RV transmissions. The effective RTT shows higher latency in this case. In contrast, Fig. 3(b) illustrates the performance when RV0 is transmitted at nominal power and retransmissions are performed at 23 dBm. In this case, retransmissions are limited to 2 RVs for UEs predicted to require more RVs under nominal power and the effective latency is greatly reduced.

Ⅲ. Conclusion

The simulation results demonstrate that incremental resource allocation strategies yield lower latency compared to slot-wise transmission in prediction-based IR-HARQ. Additionally, increasing the retransmission power significantly enhances decoding reliability and reduces latency, enabling the system to better meet the stringent requirements of URLLC. These findings highlight the effectiveness of the proposed prediction assisted MU-HARQ framework in optimizing resource utilization while adhering to reliability and latency constraints. In future work, we plan to extend this work to optimize the latency, power, and reliability extensively and analytically.

ACKNOWLEDGMENT

This work was supported by the National Research Foun dation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00278925).

REFERENCES

[1] R. Santos, D. Castanheira, A. Silva and A. Gameiro," Pipelined Multi-User IR-HARQ Scheme for Improved Latency Performance in URLLC," in IEEE Access, vol. 12, pp. 33473-33485, 2024.

[2] N. P. Kusi, S. H. Ahn and D. H. Kim," Hybrid ARQ for URLLC Using Deep Learning," in 2024 15th International Conference on Information and Communication Technology Convergence (ICTC), pp. 179-181, 2024.