Semantic Communication for VR Music Live Streaming With Rate Splitting

Jiaqi Zou, *Graduate Student Member, IEEE*, Lvxin Xu, and Songlin Sun, *Senior Member, IEEE*

***Abstract*—Virtual reality (VR) live streaming has established a remarkable transformation of music performances that facilitates a unique interaction between artists and their audiences within a virtual environment, offering an experience that significantly surpasses the conventional constraints of live music events. This article proposes a novel framework for enhancing VR music live streaming through the integration of semantic communication and rate splitting. The framework aims to improve user experience by efficiently transmitting music and speech components. It utilizes a semantic encoder to separately extract semantic information for music and speech, to capture the unique characteristics of music and speech. After having the extracted feature, we propose a rate-splitting-based algorithm in the transmission of music and speech to enhance user utility by designating music as a common message for all users and speech as a private message targeted to specific users based on their preferences. Simulation results demonstrate significant performance gain compared to the baseline methods.**

***Index Terms*—Music live streaming, rate splitting, semantic communication, virtual reality (VR).**

I. Introduction

In the realm of digital entertainment, virtual reality (VR) live streaming has been regarded as a groundbreaking innovation, offering new possibilities for music performances. This advanced technological modality facilitates a unique interaction between artists and their audiences within a virtual environment, offering an experience that significantly surpasses the conventional constraints of live music events [1]. VR live streaming enables individuals to partake in the dynamic ambiance of concerts without the necessity of physical presence. This technological progression not only challenges the traditional notions of geographical and spatial limitations in live music but also heralds a new era of accessibility, interactivity, and immersion within the music industry [2].

In the domain of VR music live streaming, the minimization of latency is paramount due to its profound impact on user immersion, interactive engagement, audio-visual synchronization, and user comfort. Latency is crucial for maintaining the illusion of presence, a core tenet of VR that fosters a convincing sense of being within the virtual environment [3], [4]. High latency disrupts this seamless integration, leading to a diminished immersive experience. Thus, ensuring low latency is indispensable for the overall quality of VR music live streams, affecting not only the realism and engagement of the virtual experience but also the comfort and satisfaction of the user, which is one of the key challenges in delivering interactive VR music events.

In music live streaming, there are primarily two main types of streams: music as the common content delivered to each receiver and surrounding speech messages that vary based on the interactive group. Rate splitting, a technique that divides a data stream into substreams with different priorities, offers an effective solution for VR music live streaming. By separating universal content (e.g., music) from targeted content (e.g., surrounding speech), it ensures consistent delivery and minimizes delay [5], [6]. Thus, this article proposes to use rate splitting to optimize VR live streaming, enhancing immersion, and interaction by reducing latency for real-time interactions in VR environments.

Semantic communication, by processing source messages to extract their semantics and transmitting only relevant information, holds the potential to significantly reduce data transmission while preserving the original semantics, enabling the provision of the same service quality with lower data transmission [7]. Semantic communication, which emphasizes the transmission of meaningful content over raw data, presents a strategic method for latency reduction in VR music transmission. This approach, by prioritizing the conveyance of significant musical and interactive elements, enables a more efficient data transmission process, thereby mitigating bandwidth demands and facilitating quicker content delivery.

Against this background, this article proposes to utilize semantic communication and rate splitting for VR music live streaming, where both music and speech are transmitted to enhance the experience of the users. Specifically, the semantic information of the music and speech is extracted separately, taking advantage of the semantic encoder. Then, the semantic

Manuscript received 27 March 2024; revised 8 June 2024; accepted 26 July 2024. This work was supported by the National Key Research and Development Program of China under Grant 2023YFF0904600. *(Corresponding author: Songlin Sun.)*

Jiaqi Zou is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: jqzou@mail.tsinghua.edu.cn).

Lvxin Xu and Songlin Sun are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: xulvxxin@bupt.edu.cn; slsun@bupt.edu.cn).

Digital Object Identifier 10.1109/TCSS.2024.3443176

feature of the music is regarded as the common message that is transmitted to all users. The speech information is regarded as a private message that is transmitted to specific users according to their requirements. The main contributions of this article are summarized as follows.

1) We introduce a new framework that combines semantic communication with rate splitting specifically for VR music live streaming.
2) Our approach involves utilizing semantic communication to transmit compact semantic information extracted from the original large-scale data. Furthermore, we employ two distinct bandwidths to extract semantic features separately, taking into account the different frequency characteristics of music and speech.
3) Furthermore, we introduce the application of rate splitting to the transmission of music and speech, aiming to enhance the efficiency and quality of data delivery in VR music live streaming environments.

The remainder of this article is organized as follows. Section II introduces the related works, including rate splitting and semantic communication. Section III introduces the system model, including the semantic communication lightweight encodec model and the rate splitting model. The proposed framework, including the framework overview, the semantic encoder, the quantizer and decoder, and the precoder optimization, is given in Section IV. Numerical results are provided in Section V. Finally, we conclude the article in Section VI.

Notations: $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and the conjugate transpose of a matrix, separately; $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L1 norm and L2 norm of a matrix, respectively; $\mathrm{Tr}(\cdot)$ denotes the trace of a square matrix; $\log_2(\cdot)$ is the base-2 logarithm function; $\mathbb{C}^{m\times n}$ stands for an $m \times n$ complex matrix; $\mathbf{x} \sim \mathcal{CN}(\mathbf{\Lambda}, \mathbf{\Delta})$ represents the circularly symmetric complex Gaussian vector having a mean vector of $\mathbf{\Lambda}$ and covariance matrix of $\mathbf{\Delta}$; and $\mathrm{Re}(a)$ and $\mathrm{Im}(a)$ denote the real and imaginary parts of a complex number a, respectively.

II. Related Works

A. Rate Splitting

The concept of rate splitting in rate splitting multiple access (RSMA) is proposed as a novel, versatile, and robust framework for designing and optimizing future wireless networks' nonorthogonal transmission, multiple access (MA), and interference management strategies [8]. RSMA provides a soft bridge between two extreme interference management strategies: fully decoded interference and interference treated as noise, by splitting user messages and enabling nonorthogonal transmission of common messages for multiuser decoding and private messages for individual user decoding.

RSMA offers a more appealing solution in terms of performance and complexity, retaining the benefits of space division multiple access (SDMA) and nonorthogonal multiple access (NOMA) while addressing their inherent limitations [9]. RSMA encompasses SDMA and NOMA as special cases, transitioning to SDMA if channel strengths are similar and orthogonal and to NOMA if channels exhibit diverse strengths and alignment. The use of RS is influenced by multiuser interference from imperfect channel state information at the transmitter (CSIT) in multiantenna deployments [10], [11]. RSMA's rate performance surpasses that of SDMA and NOMA, optimally exploiting both spatial dimensions and CSIT availability, even in scenarios of perfect or imperfect CSIT [12], [13]. RSMA is robust to inaccurate channel state information (CSI) and resilient to hybrid quality of service (QoS) requirements, performing efficiently in nonorthogonal, misaligned, or similar user channels, regardless of perfect or imperfect transmitter knowledge of CSI [14].

Previous research has shown the feasibility of supporting enhanced VR performance using RSMA technology based on semantic communication. Huong Giang et al. [15] studied the maximum total rate of downlink RSMA systems, framing the optimization problem as a Markov decision process and employing deep reinforcement learning algorithms to handle the stochastic network environment. In [16], RSMA's private message portion is explored for semantic information transmission, achieving ultrareliable and low-latency communication (URLLC). Huang et al. [17] proposed an intelligent reflecting surface (IRS)-assisted RS VR streaming system, leveraging users' common interests in VR streaming, while IRS supports high-resolution 360-degree video transmission using the deep deterministic policy gradient with imitation learning (DeepGRAIL) algorithm to optimize various parameters.

Improvement schemes of RSMA have also been extensively studied. The uplink RSMA communication problem aimed at maximizing the total wireless user rate was investigated in [18], introducing a user-pair-based algorithm that enables each user pair to utilize RSMA and allocates orthogonal frequencies to users in different pairs. Yang et al. [19] proposed a successive convex approximation algorithm for multiantenna base station RSMA to obtain suboptimal solutions maximizing the transmission power of common messages. Besides, the resource allocation problem in a reconfigurable intelligent surface (RIS)-assisted wireless communication system with RSMA was investigated in [20], proposing an iterative algorithm to address phase optimization and beamforming optimization subproblems iteratively.

B. Semantic Communication

Yang et al. provide a comprehensive survey on semantic communication in 6G, categorizing it into semantics-oriented, goal-directed, and semantic-aware communication [21]. Semantic communication reduces bandwidth usage, enhances reliability, and meets future network demands for intelligence and simplicity, making it crucial for 6G networks [22]. The advancement of AI has shown immense potential in wireless communication, enabling semantic encoding tasks. Researchers use deep learning models to model semantic features of information sources, achieving significant results. For text sources, models such as GPT [23] and BERT [24] excel in natural language processing tasks. Guo et al. proposed a semantic importance-aware communication (SIAC) scheme using pretrained models such as ChatGPT and BERT [25]. For image sources, Ren and Wu introduced an asymmetric semantic communication network

using a diffusion model for image transmission and recovery, outperforming GAN-based models [26]. For audio sources, Encodec, a high-fidelity neural audio compression model, achieves up to 40.

Recent surveys have addressed different aspects of semantic communication. Lan et al. presented a machine intelligence semantic communication framework for human-to-human (H2H), human-to-machine (H2M), and machine-to-machine (M2M) communication [27]. Qin et al. provided an overview of the theory, frameworks, system designs, and performance metrics of semantic communication [28]. Iyer et al. investigated technological trends in semantic communication in intelligent wireless networks, discussing cross-layer interactions, goal-oriented communication applications, and challenges [29]. Liu et al. reviewed semantic communication applications in UAV communication, remote image perception, intelligent transportation, and healthcare [30]. Li et al. surveyed technologies such as AI, spatiotemporal data representation, semantic IoT (SIoT), and semantic-enhanced digital twins (SDTs), presenting use cases in the ubiquitous semantic metaverse [31].

Researchers have proposed feasible architectures for semantic communication in edge distributed network architecture. A metanetwork proposed in [32] can exceed Shannon's limit by leveraging multifaceted information and intelligent collaboration among distributed entities. Shi et al. proposed an architecture based on federated edge intelligence, allowing users to offload semantic encoding and decoding tasks to edge servers, supporting resource-efficient semantic-aware networks [33]. The potential technological applications of semantic communication have been extensively studied. Rezaei et al. developed software for automatic transmission of semantically segmented map images via BPSK channels [34]. Chen et al. designed cross-modal semantic fusion and similarity evaluation methods for multimodal data transmission [35]. Wu et al. presented cross-task semantic transfer, a transfer learning approach for object detection training with limited labels [36]. Tang et al. proposed combining semantic features from direct and relay links to estimate information recovery, introducing a metric for balancing recovery and energy consumption [37]. Sheng et al. introduced a BERT-based multitext task communication system [38].

III. System Model and Problem Formulation

As depicted in Fig. 1, we consider a multiuser VR transmission system, where the BS equipped with M transmit antennas serves K single-antenna VR users for communication with $K \leq M$. Let $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$ denote the VR user set. In the context of live music performances experienced through VR, the transmission of music is a universal requirement for all users. Beyond this foundational aspect, the BS additionally transmits the speech content, encompassing interactions among users. This latter form of transmission is contingent upon specific requests by the users, indicating a customized approach to content delivery based on individual user needs or preferences. This dual-faceted transmission strategy underscores the importance of a flexible and responsive communication infrastructure within VR music performance environments, catering both to the collective experience of music and to the personalized interactive experiences among participants.

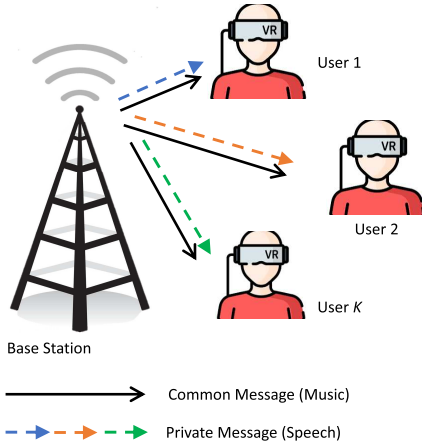

Fig. 1. Illustration of a VR music performance live streaming. The base station transmits both music and speech to multiple users according to their preferences.

A. Semantic Coding Model

In VR music live streaming, semantic communication is developed for music and speech transmission. Due to distinct frequency distributions, high-bandwidth and low-bandwidth compression with a lightweight audio coding model [39] is utilized, ensuring end-to-end audio signal transmission.

An audio signal of duration d is represented as a sequence $\boldsymbol{y} \in [-1, 1]^{C \times T}$ with C representing the number of audio channels, $T = d \cdot f_{\text{sr}}$ for the number of audio samples at a given sample rate f_{sr}. The lightweight encoder–decoder model is mainly composed of four components.

1) The encoder module employs a 1-D convolution layer with C channels and a kernel size of 7, succeeded by four 2-D convolution blocks, as shown in Fig. 2. The residual unit contains two convolutions with a kernel size of 3 and a skip-connection. Subsequently, the convolution blocks are followed by a transformer layer for sequence modeling and one 1-D convolution layer.
2) The quantizer module utilizes residual vector quantization (RVQ) to quantize the encoder's output. Vector quantization involves mapping an input vector to the nearest entry in a specified-size codebook. RVQ enhances this process by calculating the residual postquantization, subsequently subjecting it to further quantization using a secondary codebook and repeating as necessary.
3) The decoder module mirrors the encoder module, outputting the final mono or stereo audio.
4) Balanced loss functions: The reconstruction loss consists of both time and frequency domain components and the VQ commitment loss. In the time domain, we minimize the L1 distance between the target $\boldsymbol{y}$ and compressed audio $\hat{\boldsymbol{y}}$, denoted as

$$\ell_t(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1. \tag{1}$$

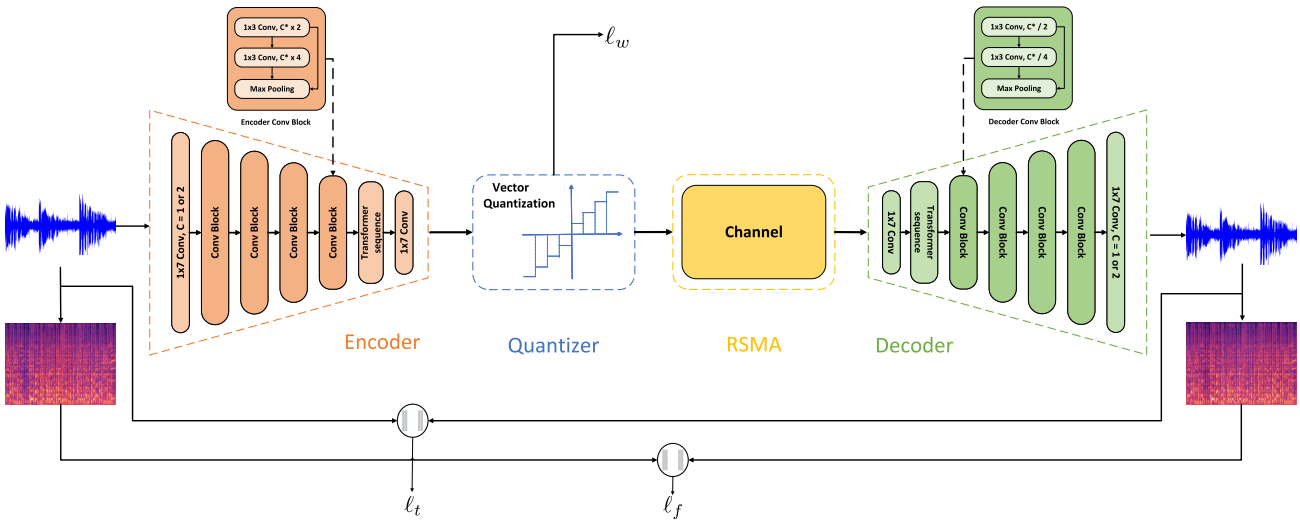

Fig. 2. Illustration of the lightweight end-to-end audio coding model.

Additionally, in the frequency domain, we employ the L2 losses over the mel-spectrogram, integrating multiple time scales, denoted as

$$\ell_f(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{s} \sum_{i \in e} \|\mathcal{S}_i(\boldsymbol{y}) - \mathcal{S}_i(\hat{\boldsymbol{y}})\|_1 + \|\mathcal{S}_i(\boldsymbol{y}) - \mathcal{S}_i(\hat{\boldsymbol{y}})\|_2 \quad (2)$$

where s is normalized parameter, $\mathcal{S}_i(\cdot)$ is a 64-bins mel-spectrogram function using a normalized STFT with window size of 2^i and hop length of $2^i/4$, $e = 5, \ldots, 11$, is the set of scales. For each residual step $n \in \{1, \ldots N\}$ (with N depending on the bandwidth target for the current batch), noting z_c the current residual and $q_c(z_c)$ the nearest entry in the corresponding codebook, we define VQ commitment loss l_w as

$$l_w = \sum_{n=1}^{N} \|\boldsymbol{z}_c - q_c(\boldsymbol{z}_c)\|_2^2. \quad (3)$$

Overall, the generator is trained to optimize the following loss, summed over the batch:

$$L_G = \lambda_t \cdot \ell_t(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \lambda_f \cdot \ell_f(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \lambda_w \cdot \ell_w \quad (4)$$

where λ_t, λ_f, and λ_w are the scalar coefficients to balance between the terms.

The entire system undergoes end-to-end training to minimize a reconstruction loss spanning both the temporal and frequency domains. Additionally, a VQ commitment loss is incorporated, operating at varying resolutions. A visual depiction is provided in Fig. 2 for clarity.

B. RS-Based Downlink Transmission

The music message is encoded into a common stream s_c using a codebook shared by both users. Thus, s_c is a common stream required to be decoded by both users. The speech message required by the kth user is encoded into the private stream $s_{p,k}$. Hence, the overall data streams to be transmitted based on RS are $\mathbf{s} = [s_c, s_{p,1}, s_{p,2}, \ldots, s_{p,K}]$ and $\mathrm{tr}(\mathbf{s}^H\mathbf{s}) = \mathbf{I}$.

The data streams are linearly precoded with the beamforming matrix $\mathbf{W} = [\mathbf{w}_c, \mathbf{w}_{p,1}, \mathbf{w}_{p,2}, \ldots, \mathbf{w}_{p,K}]$, where $\mathbf{w}_c$ is the precoder for the common stream s_c and $\{\mathbf{w}_{p,k}\}_{k=1}^K$ is the precoder for the private stream $\{\mathbf{s}_{p,k}\}_{k=1}^K$. Then, the transmitted signal vector of the BS is given by $\mathbf{x} = \mathbf{W}\mathbf{s}$. We denote the channel from the BS to the kth user as $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$ and $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$. Then, the received signal at the kth user is given as follows:

$$y_k = \mathbf{h}_k^H \mathbf{w}_c s_c + \sum_{k=1}^{K} \mathbf{h}_k^H \mathbf{w}_k s_{p,k} + z_k \quad (5)$$

where z_k denotes the additive white Gaussian noise (AWGN) received at the kth user, $z_k \sim \mathcal{CN}(0, \sigma^2)$. The kth user first decodes the common message by treating the private messages of all users as interference. The SINR of the common message at the kth user is given by

$$\gamma_{c,k} = \frac{|\mathbf{h}_k^H \mathbf{w}_c|^2}{\sum_{j=1}^{K} |\mathbf{h}_k^H \mathbf{w}_{p,j}|^2 + \sigma^2}. \quad (6)$$

Then, the achievable rate for the common message at the kth is given by

$$R_{c,k} = \log_2(1 + \gamma_{c,k}). \quad (7)$$

Let R_c denotes the transmission rate of the common message. All users need to decode the common message first and then remove it from their respective received signal to decode their private message. To ensure the successful decoding of the common message for all users, we have the following constraint:

$$R_c = \min\{R_{c,1}, R_{c,2}, \ldots, R_{c,K}\}. \quad (8)$$

After decoding the common message, user n removes the signal corresponding to the common message from y_k using SIC and decodes its private message by treating the private messages of other users as interference. Thus, the SINR of the private message at the kth user is given by

$$\gamma_{p,k} = \frac{|\mathbf{h}_k^H \mathbf{w}_{p,k}|^2}{\sum_{j=1, j\neq k}^{K} |\mathbf{h}_k^H \mathbf{w}_{p,j}|^2 + \sigma^2}. \quad (9)$$

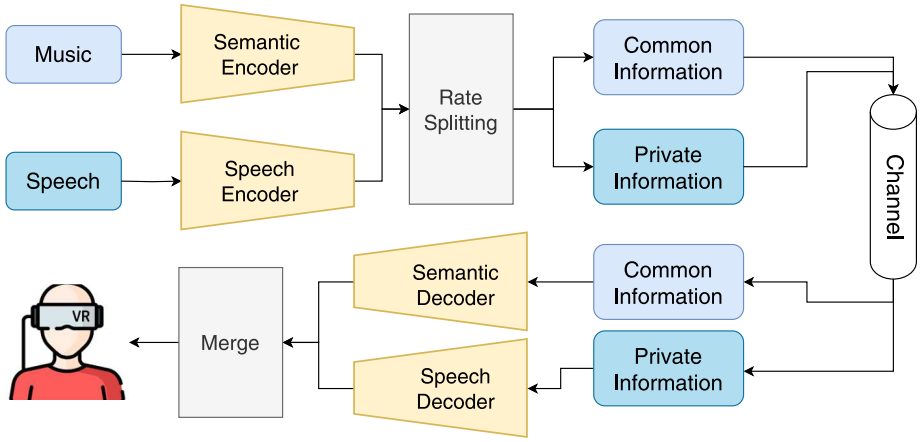

Fig. 3. Illustration of a VR music performance live streaming. The base station transmits both music and speech to multiple users according to their preferences.

Then, the achievable rate of the private message of the kth user is given by

$$R_{p,k} = \log_2(1 + \gamma_{p,k}). \tag{10}$$

Following the RS structure described above, the total achievable rate of the kth user can be represented by

$$R_k = R_{c,k} + R_{p,k}. \tag{11}$$

Considering the different preferment of the users, we define the utility of the users as u_k, which can be given by

$$u_k = \mu_{c,k} R_{c,k} + \mu_{p,k} R_{p,k}. \tag{12}$$

C. Problem Formulation

Based on the discussion above, the problem can be formulated as follows:

$$\max_{\mathbf{W}} \sum_{k=1}^{K} u_k \tag{13a}$$

$$\text{s.t. } \| \mathbf{W}\|_F^2 \leq P_{\max} \tag{13b}$$

$$R_c = \min\{R_{c,1}, R_{c,2}, \ldots, R_{c,K}\} \tag{13c}$$

$$R_c \geq R_{c,\text{th}} \tag{13d}$$

$$R_{p,k} \geq R_{\text{th},k}, \forall k \in K. \tag{13e}$$

Our objective is to maximize the aggregate utility of all users, as given in the objective function. Equation (13b) specifies the power limitation, with $P_{\max}$ representing the overall budget for transmission power. Equation (13d) outlines the requirement for the transmission of common messages to ensure the fidelity of music transmission, wherein $R_{c,\text{th}}$ signifies the threshold rate for such common messages. Similarly, (13e) establishes the criteria for the transmission of private messages, a requirement aimed at preserving the quality of speech transmission, with $R_{\text{th},k}$ indicating the threshold rate for the private message for the kth user.

IV. Problem Solution

A. Framework Overview

As shown in Fig. 3, we propose a novel framework for music performance live streaming, aiming at optimizing the transmission of audio content, including music and speech, in live streaming scenarios over wireless channels. This framework integrates semantic encoding techniques for music data with conventional encoding approaches for speech data, thereby enhancing the efficiency of channel utilization and ensuring the integrity of the transmitted content.

Specifically, the framework receives two distinct streams of data, i.e., music stream and speech stream. These streams are then processed through dedicated pathways: the semantic encoder for music, which is designed to capture and encode the essential semantic features of the musical content, and the speech encoder for speech, focusing on preserving the intelligibility and clarity of verbal communication. This bifurcated initial processing stage is crucial for preparing the disparate types of data for efficient transmission.

Subsequently, the encoded outputs are transmitted by rate splitting, which designates music as a common message for all users and speech as a private message targeted to specific users based on their preferences. This is predicated on the differentiation of data based on its relative importance and utility to the end-user, thereby optimizing the allocation of channel resources.

At the receiver side, the data stream is handled by two parallel decoding pathways. The semantic decoder is specifically tasked with reconstructing the music data by focusing on its semantic elements, ensuring that the essential qualities of the music are accurately reproduced. In parallel, the speech decoder is dedicated to the restoration of the speech data to its original form, emphasizing clarity and comprehensibility.

The final stage merges the decoded music and speech, which recombines the outputs from the semantic and speech decoders into two separate streams, one for music and another for speech. These streams are then presented to the listener, completing the transmission process.

Overall, the framework highlights the synergy between semantic processing and traditional encoding/decoding methodologies, significantly enhancing the live streaming experience for music and speech over wireless channels. It underscores the necessity for data-type specific processing—semantic encoding for music and conventional encoding for speech—and the strategic employment of rate splitting to maximize the effective use of limited channel resources. This approach not only optimizes bandwidth usage but also ensures the high fidelity of audio content delivered to end-users in live streaming applications.

B. Semantic Encoder and Decoder

The encoder of the lightweight encodec model is characterized by a stream-based architecture utilizing 1-D convolutions, tasked with converting the input audio signal into a latent representation. Segments of the audio signal are sampled to generate the sequence $\boldsymbol{y}$, which is then fed into the encoder as a 1-D vector. Initially, there is a 1-D convolutional layer with either 1 or 2 channels (depending on whether the audio is mono or stereo) and a convolutional kernel size of 7, followed by four convolutional blocks. Each convolutional block comprises a residual unit and a downsampling layer, where the residual unit consists of two convolutional layers with a kernel size of 3 and a skip connection. The downsampling layer is a convolutional layer with a stride of 2 and a kernel size of 4. With each downsampling operation, the number of channels is doubled to maintain the width of the feature maps. Following the convolutional blocks is a simple Transformer network employed for sequence modeling of the latent representation, capable of capturing long-term dependencies in the audio signal crucial for compression quality. The output of the encoder is a latent representation utilized for subsequent quantization and decoding processes.

The decoder segment, likewise stream-based, employs a 1-D transposed convolutional network structure and is responsible for reconstructing the encoder's output latent representation into a time-domain signal. The decoder receives the compressed latent representation from the encoder as input, utilizing transposed convolutional layers. The stride of the transposed convolutional layers matches that of the encoder but in reverse order. This allows the decoder to progressively recover high-resolution audio signals from low-resolution latent representations. The decoder outputs the final mono or stereo audio signal $\hat{\boldsymbol{y}}$, with separate processing for left and right channels in the case of stereo audio. The design of the decoder enables it to effectively recover high-quality audio signals from the encoder's compressed representations.

C. Precoder Optimization

After having the encoded information by the semantic encoder, we utilize rate splitting for multiuser VR streaming systems. Due to the fractional property of the multiratio terms $\gamma_{c,k}$ and $\gamma_{p,k}$, $R_{c,k}$ and $R_{p,k}$ are still nonconcave functions of $\mathbf{w}_k$. To tackle such nonconvexity, quadratic transform [40, Theorem 1] and convex approximation approach are applied. We propose to seek a linear surrogate function for the convex quadratic terms $R_{c,k}$ and $R_{p,k}$ by employing the first-order Taylor expansion of $|\mathbf{h}_k^H\mathbf{w}_{p,k}|^2$ at the current point $\mathbf{w}_{p,k}^{(m)}$ which is a global lower bound

$$\begin{aligned}&|\mathbf{h}_k^H\mathbf{w}_{p,k}|^2=(\mathbf{h}_k^H\mathbf{w}_{p,k})^H(\mathbf{h}_k^H\mathbf{w}_{p,k})\\&\geq 2\Re\left\{(\mathbf{w}_{p,k}^{(m)})^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_{p,k}\right\}-((\mathbf{w}_{p,k}^{(m)})^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_{p,k}^{(m)}).\end{aligned}\tag{14}$$

Therefore, plugging the result of (14) into (12), a surrogate function for u_k is constructed as

$$\begin{aligned}u_k&\geq \tilde{u_k}^{(m)}\triangleq\mu_{c,k}R_{c,k}^{(m)}+\mu_{p,k}R_{p,k}^{(m)}\\&=\mu_{c,k}\log_2(1+\gamma_{c,k}^{(m)})+\mu_{p,k}\log_2(1+\gamma_{p,k}^{(m)})\\&=\mu_{c,k}\log_2\Bigg(1+\mathbf{z}_c^*|\mathbf{h}_k^H\mathbf{w}_c|-\mathbf{z}_c^{*2}\sigma^2\\&\quad-\mathbf{z}_c^{*2}\operatorname{Re}\left(\sum_{j=1}^{K}\mathbf{Tr}((\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_{p,j}^{(m)})^H(2\mathbf{w}_{p,j}-\mathbf{w}_{p,j}^{(m)}))\right)\Bigg)\\&\quad+\mu_{p,k}\log_2\Bigg(1+\mathbf{z}_p^*|\mathbf{h}_k^H\mathbf{w}_p|-\mathbf{z}_p^{*2}\sigma^2\\&\quad-\mathbf{z}_p^{*2}\operatorname{Re}\left(\sum_{j=1,j\neq k}^{K}\mathbf{Tr}((\mathbf{h}_k\mathbf{h}_k^H\mathbf{w}_{p,j}^{(m)})^H(2\mathbf{w}_{p,j}-\mathbf{w}_{p,j}^{(m)}))\right)\Bigg)\end{aligned}\tag{15}$$

where $\mathbf{z}_{c,m}^*$ and $\mathbf{z}_{p,m}^*$ are the auxiliary variables, which can be, respectively, updated by $\mathbf{z}_{c,m}^*=(|\mathbf{h}_k^H\mathbf{w}_c^{(m)}|/\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{p,j}^{(m)}|^2+\sigma^2)$ and $\mathbf{z}_{p,m}^*=(|\mathbf{h}_k^H\mathbf{w}_{p,k}^{(m)}|/\sum_{j=1,j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_{p,j}^{(m)}|^2+\sigma^2)$. The surrogate function $\tilde{u_k}^{(m)}(\mathbf{W})$ is globally concave with respect to $\mathbf{W}$ which permits a premium solution for (13a). The optimization problems can be transformed into

$$\max_{\{\mathbf{w}_k\}_{k=1}^K}\ \sum_{k=1}^{K}\tilde{u_k}^{(m)}\tag{16a}$$

$$\text{s.t.}\quad (13b),(13c),(13d).\tag{16b}$$

The proposed algorithm is shown in Algorithm 1.

Algorithm 1: Proposed Iterative Algorithm for Handling (13).

Require: $\mathbf{h}, K, M, \sigma^2$.
Ensure: $\mathbf{W}^*$.
1: Initialize $\mathbf{h}, K, M, \sigma^2, \mathbf{W}^{(0)}$.
2: **while** not converged **do**
3: Update $\mathbf{z}_{c,m}^*\leftarrow\frac{|\mathbf{h}_k^H\mathbf{w}_c^{(m)}|}{\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{p,j}^{(m)}|^2+\sigma^2}$.
4: Update $\mathbf{z}_{p,m}^*\leftarrow\frac{|\mathbf{h}_k^H\mathbf{w}_{p,k}^{(m)}|}{\sum_{j=1,j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_{p,j}^{(m)}|^2+\sigma^2}$.
5: $\mathbf{W}^{(m+1)}\leftarrow$ Update $\mathbf{W}^{(m)}$ by solving problem (17).
6: $m\leftarrow m+1$.
7: **end while**
8: **return** $\mathbf{W}^*=\mathbf{W}^{(\mathbf{m+1})}$.

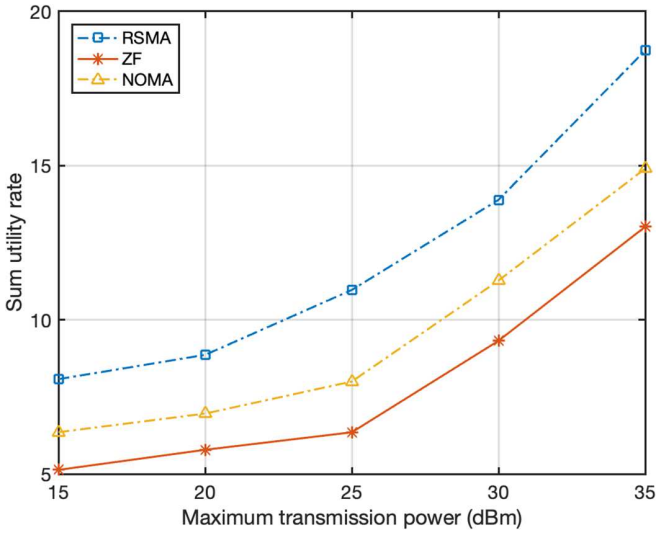

Fig. 4. Sum utility rate versus maximum transmission power $P_{\max}$ (dBm) with $K = 4, M = 8$ on RSMA, ZF, and NOMA

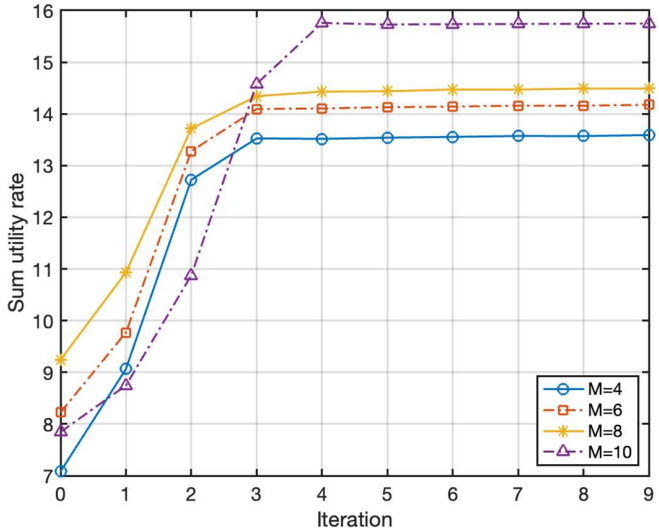

Fig. 5. Sum utility rate versus iteration with $K = 4$, M = 4, M = 6, M = 8, M = 10.

V. Numerical Results

A. Experiment Setup

1) Dataset: The model is trained on 24-kHz monophonic data across various domains, including speech, noisy speech, and music, whereas the fullband stereo model is exclusively trained on 48-kHz music. Speech segments from DNS challenge 4 and FSD50K are utilized for speech training. For music training and evaluation, the Jamendo dataset is employed, supplemented by an in-house proprietary music dataset for further evaluation. Dataset splits are established as follows: 90% of clean segments from DNS challenge 4 are allocated for training, with 5% each for validation and testing. A similar approach is applied to FSD50K, utilizing the development set for training and dividing the evaluation set for validation and testing purposes.

2) Parameter Setting: We train all models for ten epochs, with one epoch consisting of more than 2000 updates using the Adam optimizer with a batch size of 32 examples, each of 1-s duration. The learning rate is set to 2×10^{-4}, with $\beta_1 = 0.5$ and $\beta_2 = 0.9$. All models are trained on four GeForce RTX 4080 GPUs. We utilize the balancer introduced before, with weights $\lambda_t = 0.5$, $\lambda_f = 1$, and $\lambda_w = 1$ for the 24-kHz models, and $\lambda_t = 0.5$, $\lambda_f = 1$, and $\lambda_w = 2$ for the 48-kHz models.

B. Results

We start with the results for lightweight encodec model with a bandwidth of 6 kbps, ensuring little loss of the compression performance. The results of communication optimizing sum utility rate and semantic lightweight encodec model are shown as follows.

In Fig. 4, the performance comparison clearly demonstrates that RSMA consistently outperforms the baseline methods ZF and NOMA in terms of sum utility rate across various levels of maximum transmission power. RSMA achieves significant rate gains due to its flexible power and resource allocation strategies, effectively leveraging transmission power to enhance overall transmission rates. This flexibility also extends to the design of semantic information extraction schemes, reducing computational complexity while maintaining robustness and adaptability under varying transmission conditions. Additionally, RSMA's integrated encoding and decoding strategies improve spectral efficiency without increasing system complexity, making it a promising technology for high-efficiency semantic communication in future 6G networks.

Fig. 5 illustrates that the proposed algorithm achieves convergence within a few iterations. Specifically, the algorithm generally begins to converge by the fifth iteration and fully converges within ten iterations under the given parameters. Additionally, the figure shows that increasing the number of transmitting antennas M leads to an enhancement in communication performance, as evidenced by the higher sum utility rates achieved with larger values of M. This indicates that the algorithm not only converges efficiently but also scales well with the number of antennas, offering improved performance and making it suitable for systems requiring rapid convergence and high efficiency.

Fig. 6 demonstrates the sum utility rate achieved by the proposed algorithm under varying transmission power budgets. As observed, increasing the maximum transmission power $P_{\max}$ results in a corresponding improvement in the sum utility rate for all configurations of transmitting antennas M. Specifically, higher transmission power consistently enhances communication performance, with the sum utility rate rising steadily as $P_{\max}$ increases from 15 to 35 dBm. Additionally, the figure indicates that configurations with a greater number of transmitting antennas (i.e., $M = 6$, $M = 8$, and $M = 10$) consistently outperform those with fewer antennas (i.e., $M = 4$), highlighting the benefits of utilizing more antennas in achieving higher utility rates.

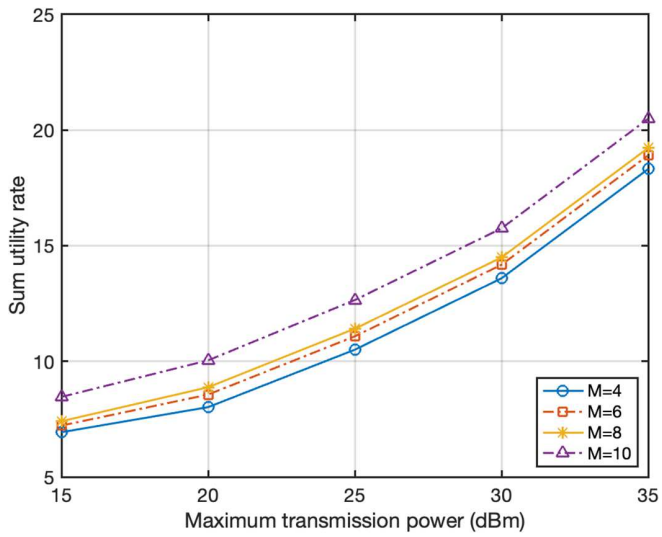

Fig. 6. Sum utility rate versus maximum transmission power $P_{\max}$ (dBm) with $K = 4$, $M = 4$, $M = 6$, $M = 8$, $M = 10$.

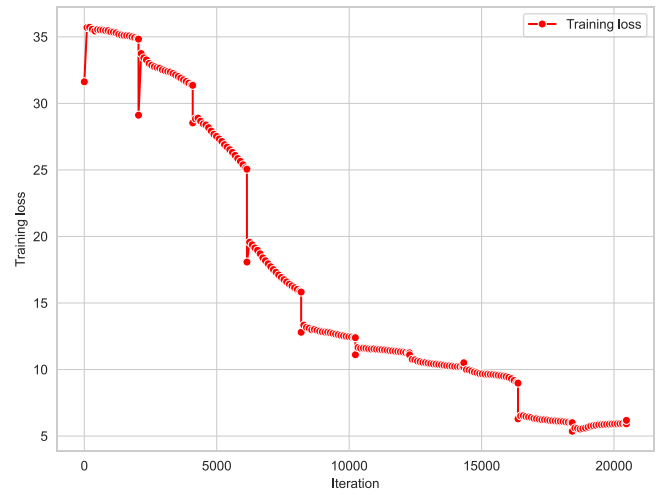

Fig. 7. Training process of L_G versus iteration on training set.

Fig. 7 depicts the loss function L_G throughout the training procedure utilizing multiple NVIDIA GeForce RTX 4080 GPUs. Following approximately 20 000 iterations, the loss function L_G ultimately achieves convergence, indicating the effectiveness of the training strategy. To expedite the convergence process, we incorporate a warm-up initialization at the commencement of each epoch, which serves to stabilize the initial training iterations and facilitate a smoother descent toward the optimal solution. This approach ensures that the model is well conditioned for further optimization, ultimately leading to improved performance.

Figs. 8 and 9 display the average mean squared error (MSE) and mean absolute error (MAE) losses across mel-frequency and time sequences derived from 100 h of data sourced from the FSD50k dataset at varying bandwidth sampling rates. Opting for a smaller bandwidth notably enhances the compression efficacy of the model, albeit at the expense of diminished quality.

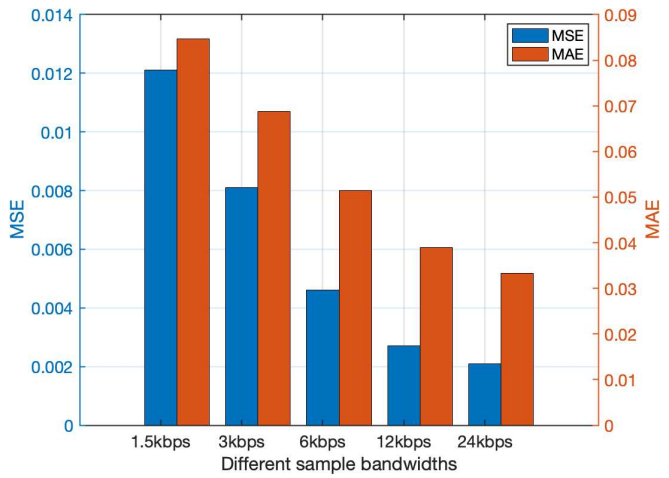

Fig. 8. MSE and MAE of mel-frequency on different sampling bandwidths.

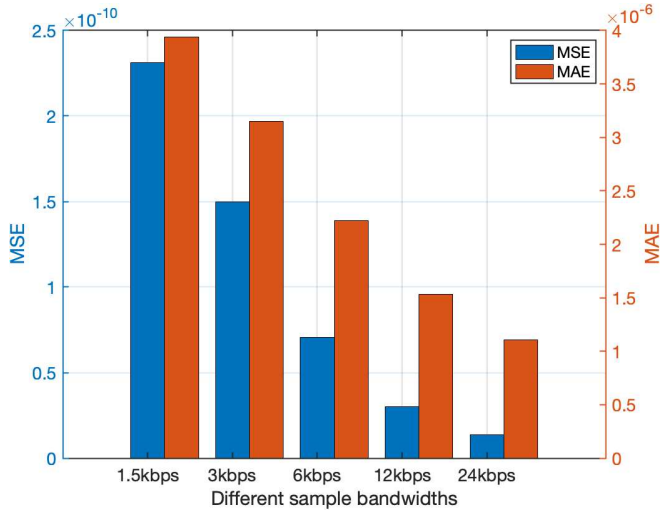

Fig. 9. MSE and MAE of time sequence on different sampling bandwidths.

Specifically, at a bandwidth of 1.5 kbps, the model achieves a compression ratio of 5%, yet at the cost of noticeable audio information loss, resulting in perceptible differences between the decoded and original audio. Conversely, at a bandwidth of 6 kbps, the compression ratio reaches 40%, rendering the compressed audio nearly indistinguishable from the original in human auditory assessments. A bandwidth of 12 kbps facilitates lossless transmission but fails to effectively reduce file size. These findings underscore the merits of semantic communication, which delivers robust compression capabilities with minimal loss of music or speech information.

Fig. 10 comprehensively showcases the robust performance of the semantic communication model across three distinct datasets, each comprising 50 h of audio data. Specifically, the MSE and MAE metrics for DNS challenge 4, FSD50K, and Jamendo are presented. Among these, DNS challenge 4 exhibits a MSE and MAE of 0.08, indicating a consistent yet slightly higher error rate. The FSD50K dataset, on the other hand, achieves a lower MSE of 0.07 and MAE of 0.06, suggesting

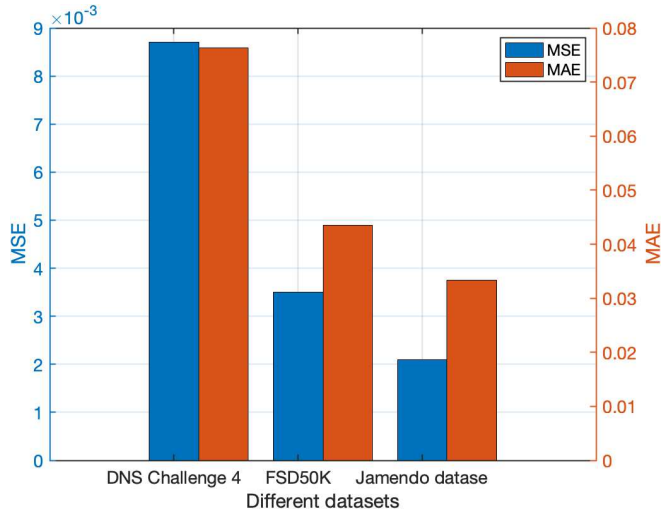

Fig. 10. MSE and MAE of mel-frequency on different datasets.

improved accuracy. Notably, the Jamendo dataset stands out with the lowest MSE of 0.05 and MAE of 0.04, demonstrating the superior performance of the semantic communication model on this particular dataset. These findings underscore the generalizability and adaptability of the proposed model across diverse audio datasets.

In summary, to transmit public information (music), a bandwidth of 12 kbps is employed for compression to uphold the quality standards of the public content. For private information (speech), a bandwidth of 6 kbps is utilized to optimize compression efficiency while preserving fundamental speech quality. Users with lower quality demands for speech information may opt for compression at 3 kbps or even 1.5-kbps bandwidth.

VI. Conclusion

In conclusion, this study has successfully developed and introduced a cutting-edge framework that significantly enhances the VR music live streaming experience by integrating semantic communication with rate splitting. This innovative approach efficiently transmits music and speech components, utilizing a semantic encoder to distinguish between common and private messages for users, based on their unique preferences. Key achievements of our research include the creation of a framework that uniquely combines semantic communication and rate splitting for VR applications, the effective use of semantic communication to process and transmit streamlined semantic information, and the strategic implementation of rate splitting to optimize the delivery of music and speech. These contributions collectively address critical challenges in streaming high-quality VR content and pave the way for more immersive and enjoyable live streaming experiences.

Looking ahead, there are numerous exciting directions for future research. One potential area is to further explore the integration of advanced semantic analysis techniques to enable even more personalized and context-aware VR music live streaming experiences. Additionally, we plan to investigate the application of our framework to other VR scenarios, such as immersive concerts, virtual events, and educational simulations. Moreover, with the increasing popularity of 5G and beyond networks, we believe that our approach can play a crucial role in enabling real-time, high-quality VR content streaming, opening up new possibilities for more immersive and enjoyable live streaming experiences. We are excited about the prospects of our research and look forward to exploring these future directions.

References

[1] S. Gunkel, M. Prins, H. Stokking, and O. Niamut, "WebVR meets WebRTC: Towards 360-degree social VR experiences," in *Proc. IEEE Virtual Reality (VR)*, 2017, pp. 457–458.

[2] D. Ochi, Y. Kunita, A. Kameda, A. Kojima, and S. Iwaki, "Live streaming system for omnidirectional video," in *Proc. IEEE Virtual Reality (VR)*, 2015, pp. 349–350, doi: 10.1109/VR.2015.7223439.

[3] C. Li, "The composition of VR system and the construction of VR teaching model in innovation and entrepreneurship education," in *Proc. 2nd Int. Conf. Inf. Sci. Educ. (ICISE-IE)*, 2021, pp. 1504–1507, doi: 10.1109/ICISE-IE53922.2021.00335.

[4] X. Ning, K. Tian, Y. Shen, Y. Liu, and H. Yang, "Optimization of VR video wireless transmission based on fountain code," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2023, pp. 1–5, doi: 10.1109/BMSB58369.2023.10211160.

[5] N. Rendevski et al., "PC VR vs standalone VR fully-immersive applications: History, technical aspects and performance," in *Proc. 57th Int. Sci. Conf. Inf., Commun. Energy Syst. Technol. (ICEST)*, 2022, pp. 1–4, doi: 10.1109/ICEST55168.2022.9828656.

[6] Y. Son, J. Yeom, S. Lim, D.-H. Kim, and K.-S. Choi, "Method and system for evaluating tracking performance OF VR/AR/MR devices," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Convergence (ICTC)*, 2022, pp. 2074–2076, doi: 10.1109/ICTC55196.2022.9952887.

[7] K. Niu et al., "A paradigm shift toward semantic communications," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 113–119, Nov. 2022, doi: 10.1109/MCOM.001.2200099.

[8] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surveys Tut.*, vol. 24, no. 4, pp. 2073–2126, Fourthquart. 2022, doi: 10.1109/COMST.2022.3191937.

[9] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 133, May 2018. [Online]. Available: http://dx.doi.org/10.1186/s13638-018-1104-7

[10] Y. Tong, D. Li, Z. Yang, Z. Xiong, N. Zhao, and Y. Li, "Outage analysis of rate splitting networks with an untrusted user," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2626–2631, Feb. 2023, doi: 10.1109/TVT.2022.3209794.

[11] M. Wu, Z. Gao, Y. Huang, Z. Xiao, D. W. K. Ng, and Z. Zhang, "Deep learning-based rate-splitting multiple access for reconfigurable intelligent surface-aided tera-hertz massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1431–1451, May 2023, doi: 10.1109/JSAC.2023.3240781.

[12] S. Aditya et al., "Rate splitting multiple access: Prototypes, experiments and standardization efforts," in *Proc. IEEE Conf. Stand. Commun. Netw. (CSCN)*, 2023, pp. 376–376, doi: 10.1109/CSCN60443.2023.10453147.

[13] G. Arora and A. Jaiswal, "Zero SIC based rate splitting multiple access technique," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2430–2434, Oct. 2022, doi: 10.1109/LCOMM.2022.3191737.

[14] B. Clerckx et al., "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1265–1308, May 2023, doi: 10.1109/JSAC.2023.3242718.

[15] H. T. Huong Giang, P. D. Thanh, H. Ko, and S. Pack, "Deep reinforcement learning-based power allocation for downlink RSMA system," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Convergence (ICTC)*, 2022, pp. 775–777, doi: 10.1109/ICTC55196.2022.9952717.

[16] C. Zeng et al., "Task-oriented semantic communication over rate splitting enabled wireless control systems for URLLC services," *IEEE Trans. Commun.*, vol. 72, no. 2, pp. 722–739, Feb. 2024, doi: 10.1109/TCOMM.2023.3325901.

[17] R. Huang, V. W. Wong, and R. Schober, "Rate-splitting for intelligent reflecting surface-aided multiuser VR streaming," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1516–1535, May 2023.

[18] Z. Yang, M. Chen, W. Saad, W. Xu, and M. Shikh-Bahaei, "Sum-rate maximization of uplink rate splitting multiple access (RSMA) communication," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2596–2609, Jul. 2022.

[19] Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5988–6002, Sep. 2021.

[20] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu, and M. Shikh-Bahaei, "Energy efficient rate splitting multiple access (RSMA) with reconfigurable intelligent surface," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2020, pp. 1–6.

[21] W. Yang et al., "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tut.*, vol. 25, no. 1, pp. 213–250, Firstquart. 2023.

[22] P. Zhang, Y. Liu, Y. Song, and J. Zhang, "Advances and challenges in semantic communications: A systematic review," *Nat. Sci. Open*, vol. 3, no. 4, 2023, Art. no. 20230029. [Online]. Available: https://api.semanticscholar.org/CorpusID:266318115

[23] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019. Available: https://arxiv.org/abs/1810.04805

[25] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2328–2332, Sep. 2023.

[26] T. Ren and H. Wu, "Asymmetric semantic communication system based on diffusion model in IoT," in *Proc. IEEE 23rd Int. Conf. Commun. Technol. (ICCT)*, 2023, pp. 1–6.

[27] Q. Lan et al., "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, 2021.

[28] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," 2022. Available: https://arxiv.org/abs/2201.01389

[29] S. Iyer et al., "A survey on semantic communications for intelligent wireless networks," *Wireless Pers. Commun.*, vol. 129, no. 1, pp. 569–611, Nov. 2022. [Online]. Available: http://dx.doi.org/10.1007/s11277-022-10111-7

[30] Y. Liu, X. Wang, Z. Ning, M. Zhou, L. Guo, and B. Jedari, "A survey on semantic communications: Technologies, solutions, applications and challenges," *Digit. Commun. Netw.*, vol. 10, no. 3, pp. 528–545, 2023.

[31] K. Li, B. P. L. Lau, X. Yuan, W. Ni, M. Guizani, and C. Yuen, "Toward ubiquitous semantic metaverse: Challenges, approaches, and opportunities," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21855–21872, Dec. 2023.

[32] Y. Lin, C. Wu, J. Wu, L. Zhong, X. Chen, and Y. Ji, "Meta-networking: Beyond the Shannon limit with multi-faceted information," *IEEE Netw.*, vol. 37, no. 4, pp. 256–264, Jul./Aug. 2023.

[33] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.

[34] H. Rezaei, T. Sivalingam, and N. Rajatheva, "Automatic and flexible transmission of semantic map images using polar codes for end-to-end semantic-based communication systems," in *Proc. IEEE 34th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2023, pp. 1–6.

[35] M. Chen, M. Liu, W. Wang, H. Dou, and L. Wang, "Cross-modal semantic communications in 6G," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2023, pp. 1–6.

[36] Q. Wu, F. Liu, H. Xia, and T. Zhang, "Semantic transfer between different tasks in the semantic communication system," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 566–571.

[37] B. Tang, L. Huang, Q. Li, A. Pandharipande, and X. Ge, "Cooperative semantic communication with on-demand semantic forwarding," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 349–363, 2024.

[38] Y. Sheng, F. Li, L. Liang, and S. Jin, "A multi-task semantic communication system for natural language processing," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC-Fall)*, 2022, pp. 1–5.

[39] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," 2022. Available: https://arxiv.org/abs/2210.13438

[40] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, Mar. 2018.