

CNN-Based Multi-Mode Layered Transmission for Robust Speech Semantic Communication

Qi Gao, Jiaqi Li, Seung-Hoon Hwang
Department of Electronics and Electrical Engineering
Dongguk University
Seoul, Korea

gaoqi@dgu.ac.kr, kikilee@dgu.ac.kr, shwang@dongguk.edu

Abstract—Deep learning-based speech semantic communication aims to learn compact speech representations for robust transmission and reconstruction over noisy wireless channels. Existing waveform-level systems usually encode each speech segment into a single latent representation, which limits transmission flexibility under varying channel conditions or transmission budgets. This paper proposes MMLT-SC, a CNN-based multi-mode layered transmission framework that decomposes the learned speech representation into primary and secondary layers and jointly trains full, partial, and primary-only reconstruction modes. The primary layer supports coarse reconstruction with fewer transmitted channel-encoder output channels, while the secondary layer provides refinement when additional channels are available. Experiments under AWGN and Rayleigh fading channels show that MMLT-SC improves full-mode reconstruction over a single-stream baseline, reducing the reconstruction MSE by up to 27.9% under Rayleigh fading channels. Furthermore, compared with a reconstruction-only layered reference, MMLT-SC improves the average SDR from 1.16 dB to 13.50 dB at 50% transmission. These demonstrate that the proposed multi-mode supervision enables reliable reconstruction across multiple transmission budgets without sacrificing full-transmission performance.

Keywords—Semantic communication, speech transmission, CNN, layered representation, multi-mode reconstruction, waveform reconstruction, wireless channel

I. INTRODUCTION

Semantic communication has attracted increasing attention as a promising paradigm for efficient and robust wireless transmission in future intelligent networks [1], [2]. Unlike conventional communication systems that mainly focus on bit-level or symbol-level reliability, semantic communication aims to transmit information that is most relevant to the communication task. This paradigm is particularly suitable for speech transmission because speech signals contain information with unequal perceptual and functional importance. For example, coarse acoustic and phonetic cues are closely related to intelligibility, whereas detailed spectral characteristics mainly affect perceptual quality and naturalness. Such unequal importance motivates the development of speech semantic communication that can prioritize essential information under limited transmitted representation budgets while degrading less critical components.

Deep learning-based speech semantic communication has demonstrated the feasibility of learning speech-related representations directly from raw waveforms for transmission over noisy wireless channels. DeepSC-S is a representative convolutional neural network (CNN)-based speech semantic communication system that jointly optimizes semantic and channel coding modules to reconstruct speech signals at the receiver and has shown improved robustness under low signal-to-noise ratio (SNR) conditions [3]. However, DeepSC-S encodes each speech segment into a single latent

representation, which is transmitted as an indivisible feature stream. Consequently, the learned representation does not explicitly separate information required for basic intelligibility from information that mainly enhances perceptual quality. This design limits transmission scalability because the receiver can recover the speech only when the complete latent representation is available. In practical wireless environments, where channel quality and transmission resources may vary over time, such a single-stream representation provides limited flexibility for adaptive transmission under different communication budgets.

Layered semantic transmission has been investigated to improve transmission flexibility under varying channel conditions [4]. For waveform-level speech semantic communication, a layered transceiver can partition the learned representation into multiple branches, providing a structural basis for scalable transmission. However, introducing a layered architecture alone does not guarantee effective operation under reduced-transmission budgets. When training is performed solely to optimize full reconstruction using all branches, the receiver tends to rely on their joint representation. Consequently, reconstruction quality may degrade significantly when only a subset of the transmitted representation is available. Therefore, the key issue is not to design a layered representation, but to learn layer-wise features that remain independently useful and support reliable speech reconstruction under different transmission budgets.

To address this problem, we propose a CNN-based multi-mode layered transmission framework for waveform-level speech semantic communication (MMLT-SC). MMLT-SC decomposes the learned speech representation into a primary layer and a secondary layer, where the primary layer is encouraged to preserve information useful for coarse reconstruction and the secondary layer provides additional refinement when more transmission resources are available. Different from a reconstruction-only layered transceiver, MMLT-SC explicitly optimizes three transmission modes: primary-only, partial, and full transmission. In the primary-only mode, only the primary branch is transmitted. In the partial mode, the full primary branch and a subset of the secondary-branch channel-encoder output channels are transmitted. In the full mode, both branches are transmitted and jointly decoded. This design enables a controllable tradeoff between reconstruction quality and transmitted-feature budget under varying channel and resource conditions.

The main contributions of this paper are summarized as follows:

- We propose a CNN-based layered speech semantic transceiver that decomposes the learned waveform-level representation into primary and secondary branches, improving full-mode reconstruction over the single-stream baseline.

- We design a multi-mode reconstruction objective that jointly optimizes primary-only, partial, and full transmission modes, enabling reliable reconstruction under different transmitted-feature budgets.
- We conduct experiments and reduced-mode ablation studies under additive white Gaussian noise (AWGN) and Rayleigh fading channels, showing that multi-mode supervision is necessary for effective reduced-budget transmission.

II. SYSTEM MODEL AND PROPOSED METHOD

The overall architecture of the proposed MMLT-SC framework is illustrated in Fig. 1. The semantic encoder, channel encoders, channel decoders, and semantic decoder follow the CNN-based waveform-level speech semantic communication structure of DeepSC-S [3]. To enable scalable transmission, a layer-splitting module is introduced after the semantic encoder to decompose the learned speech representation into a primary layer and a secondary layer. The resulting layered representation supports three transmission modes, i.e., primary-only, partial, and full transmission, corresponding to different transmitted channel-encoder output budgets. During training, these modes are jointly optimized through a multi-mode reconstruction objective, allowing the receiver to reconstruct speech under different transmitted-representation budgets.

A. Layered Speech Semantic Transceiver

The input waveform x is first mapped into a shared latent representation by the semantic encoder:

$$z = S_\alpha(x), \quad (1)$$

where $S_\alpha(\cdot)$ denotes the semantic encoder with trainable parameters α , and z is the learned speech representation.

Instead of directly transmitting the learned latent representation as a single stream, MMLT-SC employs a trainable CNN-based layer-splitting module to generate two branch-specific latent representations:

$$(z_p, z_s) = G_\beta(z), \quad (2)$$

where $G_\beta(\cdot)$ is the layer-splitting network with the trainable parameters β . In this module, two parallel 1×1 convolutional heads project the shared representation z into a primary representation z_p and a secondary representation z_s . The resulting primary representation z_p and secondary representation z_s have the same spatial size and 64 feature channels each. Under the multi-mode reconstruction objective, the primary layer is encouraged to support coarse reconstruction when transmitted alone, while the secondary layer provides additional refinement when more channel-encoder output channels are transmitted.

The two branches are independently encoded for channel transmission:

$$s_p = C_p(z_p), s_s = C_s(z_s), \quad (3)$$

where $C_p(\cdot)$ and $C_s(\cdot)$ denote the channel encoders for the primary and secondary layers, respectively. The transmission budget and transmission ratio reported in this paper are measured by the number of transmitted channel-encoder output channels.

The wireless channel is modeled as:

$$y_i = h_i \cdot s_i + n_i \quad i \in \{p, s\}, \quad (4)$$

where h_i denotes the channel coefficient and n_i denotes additive white Gaussian noise. For the AWGN channel, $h_i = 1$. For the Rayleigh fading channel, h_i is modeled as a random fading coefficient, and the noise variance is determined by the SNR. Receiver-side channel equalization with perfect channel state information is assumed for Rayleigh fading. The equalized received signal is denoted by $\bar{y}_i$, with $\bar{y}_i = y_i$ for AWGN channels.

At the receiver, the primary and secondary representations are recovered by the corresponding channel decoders:

$$\hat{z}_p = D_p(\bar{y}_p), \hat{z}_s = D_s(\bar{y}_s), \quad (5)$$

where $D_p(\cdot)$ and $D_s(\cdot)$ denote the channel decoders for the primary and secondary layers, respectively.

B. Multi-Mode Layered Transmission

MMLT-SC supports three transmission modes corresponding to different transmission budgets: full, partial, and primary-only transmission. These modes are realized by controlling the number of channel-encoder output channels that are passed through the wireless channel while maintaining the same network architecture. Consequently, the receiver can reconstruct speech using either the complete layered representation or only a subset of transmitted features.

In the full mode, both the primary and secondary channel-encoder outputs are transmitted and jointly used for reconstruction:

$$\hat{x}_{full} = R_Y([\hat{z}_p, \hat{z}_s]), \quad (6)$$

where $R_Y(\cdot)$ denotes the semantic decoder and $[\cdot]$ denotes channel-wise concatenation.

In the partial-transmission mode, the complete primary branch and only a subset of the secondary-branch channel-encoder outputs are transmitted. Let $T_r(\cdot)$ denote a channel-selection operator that keeps a ratio r of the secondary-branch channel-encoder outputs. In our implementation, $T_r(\cdot)$ retains the first r fraction ratio of the secondary channels. Using the same fixed subset during both training and inference allows the retained secondary channels to be directly optimized by the partial reconstruction loss, encouraging them to support reduced-budget reconstruction. The received secondary transmission is expressed as

$$y_{s,r} = h_s T_r(s_s) + n_{s,r}, \quad (7)$$

where $n_{s,r}$ denotes the additive noise corresponding to the reduced secondary-channel subset. After channel equalization, the received secondary subset is zero-padded to its original secondary-channel dimension by $P_r(\cdot)$, and then decoded as:

$$\hat{z}_{s,r} = D_s(P_r(\bar{y}_{s,r})). \quad (8)$$

where $P_r(\cdot)$ zero-pads the omitted positions to recover the original secondary-channel dimension. The partial-mode reconstruction is then obtained by:

$$\hat{x}_{par} = R_Y([\hat{z}_p, \hat{z}_{s,r}]). \quad (9)$$

In this work, $r=0.5$, corresponding to transmission of the complete primary branch and half of the secondary-branch channel-encoder outputs.

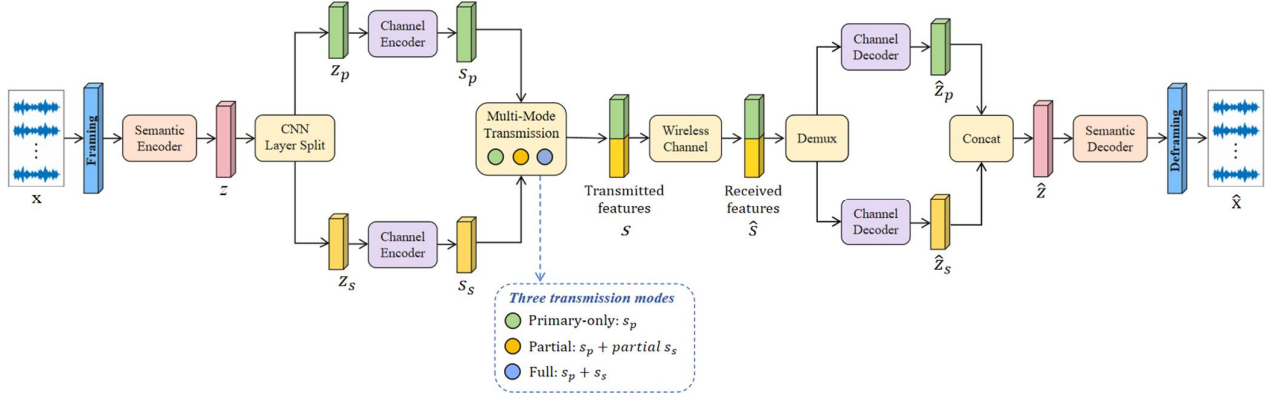


Fig. 1. Architecture of the proposed MMLT-SC framework.

In the primary-only mode, only the primary branch is transmitted, and the secondary input to the semantic decoder is set to zero:

$$\hat{x}_{pri} = R_Y([\hat{z}_p, 0]). \quad (10)$$

This mode transmits only 64 primary channel-encoder output channels. Thus, the full, partial, and primary-only modes correspond to 128, 96, and 64 transmitted channel-encoder output channels, respectively.

C. Multi-Mode Reconstruction Objective

A layered architecture alone does not guarantee effective reconstruction under reduced transmitted-channel budgets. When the model is trained only with the full-mode reconstruction objective, the primary and secondary representations may be optimized mainly for joint decoding, which can lead to performance degradation when only part of the transmitted representation is available.

The reconstruction losses for the three modes are defined as

$$L_{pri} = \frac{1}{L} \|x - \hat{x}_{pri}\|_2^2, \quad (11)$$

$$L_{par} = \frac{1}{L} \|x - \hat{x}_{par}\|_2^2, \quad (12)$$

$$L_{full} = \frac{1}{L} \|x - \hat{x}_{full}\|_2^2. \quad (13)$$

These losses are computed at the waveform level between the original speech segment and the reconstructed output of each transmission mode. The total objective is given by

$$L = w_{pri}L_{pri} + w_{par}L_{par} + w_{full}L_{full}, \quad (14)$$

where w_{pri} , w_{par} , and w_{full} are the weights of the primary-only, partial, and full reconstruction losses, respectively. In our experiments, $w_{pri} = 0.05$, $w_{par} = 0.05$, and $w_{full} = 1.0$. The full reconstruction loss is assigned the largest weight to preserve high-quality reconstruction when all layers are transmitted, while the primary-only and partial losses encourage the model to maintain functional reconstruction under reduced transmission budgets. Through joint optimization of the three reconstruction objectives, the primary representation is encouraged to support coarse reconstruction when transmitted alone, while the secondary representation provides additional refinement as more channel-encoder output channels are transmitted.

D. Evaluation Metrics

To evaluate waveform-level reconstruction quality, we adopt four metrics: mean squared error (MSE), signal-to-distortion ratio (SDR), scale-invariant signal-to-distortion ratio (SI-SDR), and perceptual evaluation of speech quality (PESQ). Let $x \in \mathbb{R}^L$ and $\hat{x} \in \mathbb{R}^L$ denote the original and reconstructed speech waveform segments, respectively.

The SDR is defined as

$$\text{SDR} = 10 \log_{10} \frac{\|x\|_2^2}{\|x - \hat{x}\|_2^2}. \quad (15)$$

A higher SDR indicates better waveform-level reconstruction.

For SI-SDR, let $x_0 = x - \text{mean}(x)$ and $\hat{x}_0 = \hat{x} - \text{mean}(\hat{x})$ denote the zero-mean signals. The scale-aligned target and residual distortion are defined as

$$x_{\text{target}} = \frac{\langle \hat{x}_0, x_0 \rangle}{\|\hat{x}_0\|_2^2} x_0, \quad (16)$$

$$e_{\text{noise}} = \hat{x}_0 - x_{\text{target}}. \quad (17)$$

Then SI-SDR is given by

$$\text{SI-SDR} = 10 \log_{10} \frac{\|x_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2}. \quad (18)$$

PESQ is used to evaluate perceptual speech quality [5]. Since all speech signals are sampled at 8 kHz, PESQ is computed in the narrow-band mode.

III. EXPERIMENTAL SETUP AND RESULTS

A. Experimental Setup

The proposed MMLT-SC framework is evaluated using the Edinburgh DataShare speech dataset, which contains more than 10,000 training waveform files and 800 test waveform files [6]. All speech signals are resampled to 8 kHz and segmented into 16384-sample frames. Following the preprocessing procedure adopted by DeepSC-S [3], each waveform frame is reshaped into a $(128 \times 128 \times 1)$ input tensor for CNN-based processing. This operation only reorganizes waveform samples and does not introduce handcrafted acoustic features. The model adopts a 64+64 layered representation, where the primary and secondary branches each contain 64 channel-encoder output channels. The primary-only, partial, and full modes transmit 64, 96, and 128 channels, respectively; the partial mode transmits the complete primary branch and the first half of the secondary branch.

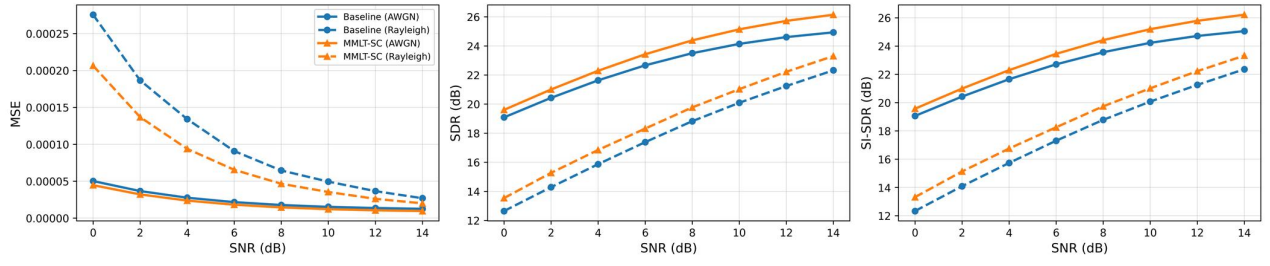


Fig. 2. Full-mode reconstruction performance comparison between baseline and MMLT-SC under AWGN and Rayleigh fading channels: (a) MSE, (b) SDR, and (c) SI-SDR.

MMLT-SC is trained using a two-stage strategy. First, a two-branch layered transceiver (Layered-Full-64+64) is trained using only the full reconstruction objective to obtain a stable layered representation. Then, MMLT-SC is initialized from this checkpoint and further fine-tuned with the proposed multi-mode reconstruction objective. This setting isolates the effect of multi-mode training on the same layered structure.

The tested SNR values are $\{0, 2, 4, 6, 8, 10, 12, 14\}$ dB. Both AWGN and Rayleigh fading channels are considered. All models are evaluated using MSE, SDR, SI-SDR, and PESQ as defined in Section II-D.

The evaluated models are summarized as follows:

- 1) Baseline (DeepSC-S): a single-stream speech semantic transceiver for waveform reconstruction.
- 2) Layered-Full-64+64: a reconstruction-only two-branch reference trained only for full-mode reconstruction.
- 3) MMLT-SC: the proposed multi-mode layered transmission framework trained with primary-only, partial, and full reconstruction objectives.

B. Overall Performance Comparison

Fig. 2 compares the reconstruction performance of the single-stream baseline and MMLT-SC under AWGN and Rayleigh fading channels. Across the entire SNR range, MMLT-SC consistently achieves lower MSE, higher SDR, and higher SI-SDR than the baseline, showing improved full-mode waveform reconstruction over the single-stream transceiver.

The performance improvement is particularly noticeable under Rayleigh fading channels, where the received signal is affected by both additive noise and fading-induced amplitude variation. For example, at 0 dB, the Rayleigh-channel MSE decreases from 2.76×10^{-4} to 2.07×10^{-4} , corresponding to a 25.0% reduction, while SDR and PESQ improve from 12.64 dB to 13.53 dB and from 2.295 to 2.365, respectively. At a moderate SNR of 8 dB, MMLT-SC reduces the Rayleigh-channel MSE by 27.9% compared with the baseline, while improving SDR and PESQ by 0.95 dB and 0.097, respectively. Even under favorable channel condition, i.e., 14 dB, MMLT-SC reduces the Rayleigh-channel MSE from 2.70×10^{-5} to 2.00×10^{-5} , corresponding to a 25.7% reduction. The SDR and PESQ are also improved by 0.98 dB and 0.087, respectively.

Under AWGN channels, MMLT-SC also consistently outperforms the baseline. For example, at 14 dB, the MSE is reduced from 1.27×10^{-5} to 9.61×10^{-6} , and SDR

increases from 24.93 dB to 26.14 dB. Overall, these results indicate that MMLT-SC improves full-mode waveform reconstruction over the single-stream baseline under both Gaussian and fading channels. The consistent MSE, SDR, and PESQ gains show that the proposed layered transceiver maintains reconstruction robustness across different SNR conditions.

C. Reference Comparison with Structural Layered Model

Table I reports the average reconstruction performance over all tested SNR values. The comparison among the baseline, Layered-Full-64+64, and MMLT-SC is used to analyze the roles of the layered structure and the proposed multi-mode reconstruction objective.

Compared with the single-stream baseline, Layered-Full-64+64 improves full-mode reconstruction under both channel conditions. Under AWGN, Layered-Full-64+64 reduces the average MSE from 2.45×10^{-5} to 2.17×10^{-5} and increases SDR from 22.62 dB to 23.21 dB. Under Rayleigh fading, it reduces the average MSE from 1.08×10^{-4} to 8.28×10^{-5} and improves SDR from 17.83 dB to 18.60 dB. These results indicate that the two-branch layered architecture contributes to full-mode waveform reconstruction.

MMLT-SC further achieves comparable or slightly better full-mode reconstruction than Layered-Full-64+64. Under AWGN, MMLT-SC obtains an average MSE of 2.07×10^{-5} , SDR of 23.46 dB, SI-SDR of 23.49 dB, and PESQ of 3.364. Under Rayleigh fading, it obtains an average MSE of 7.89×10^{-5} , SDR of 18.78 dB, SI-SDR of 18.72 dB, and PESQ of 2.896. This indicates that the proposed multi-mode objective does not degrade full-mode reconstruction performance. Therefore, Layered-Full-64+64 serves as a structural reference, while MMLT-SC further extends the layered transceiver toward controllable multi-budget transmission, as analyzed in Section III-D.

TABLE I. AVERAGE RECONSTRUCTION PERFORMANCE OVER ALL TESTED SNR VALUES.

Channel	Model	MSE $\times 10^{-5}$	SDR	SI-SDR	PESQ
AWGN	Baseline	2.45	22.62	22.68	3.305
	Layered-Full-64+64	2.17	23.21	23.26	3.342
	MMLT-SC	2.07	23.46	23.49	3.364
Rayleigh	Baseline	10.81	17.83	17.74	2.808
	Layered-Full-64+64	8.28	18.60	18.54	2.859
	MMLT-SC	7.89	18.78	18.72	2.896

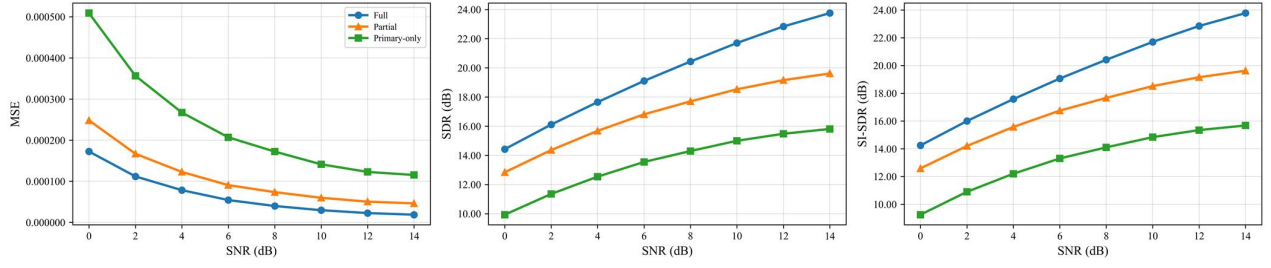


Fig. 3. Transmission mode analysis of MMLT-SC under Rayleigh fading channels: (a) MSE, (b) SDR, and (c) SI-SDR.

D. Transmission Mode Analysis

Fig. 3 analyzes the reconstruction quality of MMLT-SC under three trained transmission modes: full, partial, and primary-only transmission, corresponding to 100%, 75%, and 50% transmitted channel-encoder output budgets, respectively. Under Rayleigh fading channels, MMLT-SC forms a stable quality hierarchy across these modes. Averaged over all tested SNR values, the full, partial, and primary-only modes achieve 19.51 dB, 16.84 dB, and 13.50 dB SDR, respectively, with corresponding MSE values of 6.61×10^{-5} , 1.07×10^{-4} , and 2.36×10^{-4} . These results indicate that MMLT-SC provides trained operating points that enable controllable tradeoffs between transmitted-channel budget and reconstruction quality.

To examine whether this behavior can be obtained by a layered architecture alone, Table II compares MMLT-SC with the reconstruction-only Layered-Full-64+64 reference. Since Layered-Full-64+64 is trained only with the full reconstruction objective and has no explicit primary/secondary roles, its reduced-mode results are reported as symmetric diagnostic averages. The Avg. 75%-TX diagnostic result averages two candidates using one complete branch and half of the other branch, while the Avg. 50%-TX diagnostic result averages the two single-branch candidates.

As shown in Table II, Layered-Full-64+64 and MMLT-SC achieve comparable full-mode performance, with SDR values of 19.32 dB and 19.51 dB, respectively. However, their reduced-mode behavior differs substantially. At the 75% transmission ratio, Layered-Full-64+64 obtains 10.54 dB SDR and 10.59 dB SI-SDR, whereas MMLT-SC achieves 16.84 dB SDR and 16.76 dB SI-SDR. At the 50% transmission ratio, Layered-Full-64+64 degrades to 1.16 dB SDR and -0.98 dB SI-SDR, while MMLT-SC still maintains 13.50 dB SDR and 13.20 dB SI-SDR. These results show that a layered architecture trained only for full reconstruction does not naturally provide reliable reconstruction when only part of the transmitted representation is available. In contrast, the proposed multi-mode reconstruction objective provides more controlled partial and primary-only reconstruction.

TABLE II. REDUCED-MODE DIAGNOSTIC COMPARISON UNDER RAYLEIGH FADING, AVERAGED OVER ALL TESTED SNR VALUES.

Model	Mode	Ratio	MSE	SDR	SI-SDR
Layered-Full-64+64	Full	100%	6.82×10^{-5}	19.32	19.29
	Avg. 75%-TX diagnostic	75%	3.78×10^{-4}	10.54	10.59

	Avg. 50%-TX diagnostic	50%	3.17×10^{-3}	1.16	-0.98
MMLT-SC	Full	100%	6.61×10^{-5}	19.51	19.46
	Partial	75%	1.07×10^{-4}	16.84	16.76
	Primary-only	50%	2.36×10^{-4}	13.50	13.20

IV. CONCLUSION

This paper proposed MMLT-SC, a CNN-based multi-mode layered transmission framework for waveform-level speech semantic communication. The proposed framework decomposes the learned speech representation into primary and secondary layers and jointly trains full, partial, and primary-only reconstruction modes. Experimental results under AWGN and Rayleigh fading channels show that MMLT-SC improves full-mode reconstruction over the single-stream baseline and provides more reliable reduced-mode behavior than a reconstruction-only layered reference. Future work will investigate channel-aware mode selection and evaluate the channel-use reduction and quality tradeoff in adaptive transmission scenarios.

REFERENCES

- [1] W. Yang *et al.*, "Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213-250, Firstquarter 2023, doi: 10.1109/COMST.2022.3223224.
- [2] Z. Lu *et al.*, "Semantics-Empowered Communications: A Tutorial-Cum-Survey," in *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 41-79, Firstquarter 2024, doi: 10.1109/COMST.2023.3333342.
- [3] Z. Weng and Z. Qin, "Semantic Communication Systems for Speech Transmission," in *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434-2444, Aug. 2021, doi: 10.1109/JSAC.2021.3087240.
- [4] F. Liu *et al.*, "Adaptive Semantic Speech Transmission for High-Speed Scenarios," in *IEEE Transactions on Wireless Communications*, vol. 25, pp. 3060-3074, 2026, doi: 10.1109/TWC.2025.3601158.
- [5] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.
- [6] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR), 2017. doi: 10.7488/ds/2117.