Article

Title

Firstname Lastname [1], Firstname Lastname [2] and Firstname Lastname [2,*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)

Abstract

Driven by the evolution of next-generation wireless networks and ubiquitous sensing, the paradigm of communication is shifting from bit-level transmission to enabling intelligent, multimodal interactions within computational social systems. However, existing semantic communication frameworks predominantly prioritize textual accuracy while neglecting paralinguistic affective cues, which are paramount for human-centric interactions in the Social Internet of Things (SIoT). To address this limitation, we propose a Heterogeneous Affective Speech Semantic Communication System (HASCom), which facilitates robust and empathetic data transmission. Specifically, we design a heterogeneous dual-stream transmission architecture that decouples linguistic content from affective prosody. Reliable digital coding is employed for discrete semantic information to ensure intelligibility, whereas deep analog transmission is utilized for continuous emotional features to preserve high-fidelity details against quantization errors. Furthermore, we develop a prior-guided diffusion reconstruction module at the receiver. This module leverages the decoded semantic vectors as structural priors to guide the generative process, synthesizing high-quality speech waveforms conditioned on the recovered affective cues. Extensive experiments demonstrate that HASCom significantly outperforms state-of-the-art baselines, achieving superior performance in terms of both semantic intelligibility and emotional consistency, even under low-bandwidth and noisy channel conditions.

Keywords: semantic communication; computational social systems; affective computing; heterogeneous transmission; diffusion models; Social Internet of Things (SIoT)

1. Introduction

With the rapid evolution of next-generation wireless networks (e.g., 6G) and the proliferation of ubiquitous sensing devices, the paradigm of communication is shifting from merely transmitting bits to enabling intelligent, multimodal interactions within computational social systems [1,2]. The explosive growth of intelligent agents and high-fidelity sensors imposes unprecedented demands on bandwidth and latency, challenging traditional syntactic-level communication frameworks that focus solely on bit-level accuracy [3,4]. To address these challenges, Semantic Communication (SemCom) has emerged as a revolutionary paradigm, promising to extract and transmit only the meaning of data rather than raw signals [5–7]. Furthermore, recent advancements in Generative Artificial Intelligence (GAI) are fundamentally reshaping this landscape, empowering receivers to synthesize high-fidelity data from compressed semantic cues [8,9]. This evolution is critical for realizing the vision of the Social Internet of Things (SIoT) and Industry 5.0, where

Received:
Revised:
Accepted:
Published:

Copyright: © 2026 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

https://doi.org/10.3390/1010000

human-centric, emotionally aware, and robust interactions become paramount for bridging the gap between physical and digital worlds [10].

Among the diverse modalities facilitating social interactions, speech is paramount, serving as a dual-channel medium that conveys both explicit linguistic content and implicit affective prosody [11–13]. However, prevailing speech semantic communication frameworks predominantly prioritize the accuracy of textual content (i.e., *what* is said), often inadvertently stripping away the paralinguistic nuances (i.e., *how* it is said) that are vital for empathetic interaction [14,15]. Neuroscientific studies indicate that valid speech comprehension requires the dynamic integration of these heterogeneous cues [16], suggesting that content-only transmission is insufficient for high-fidelity social systems. Consequently, there is an urgent imperative to develop a mechanism that can simultaneously preserve both the semantic precision and the affective depth of speech signals, rather than treating them in isolation.

Despite this imperative, existing speech semantic communication systems are primarily categorized into speech-to-speech (S2S) and speech-to-text (S2T) paradigms. In the S2S framework, Weng and Qin proposed DeepSC-S, which utilizes Deep Joint Source-Channel Coding (JSCC) to map speech signals into semantic vectors for end-to-end transmission [17]. Similarly, DSST extends this approach to nonlinear wireless channels to improve signal robustness [18]. To further enhance efficiency, Han et al. developed a semantic preserved system for highly efficient speech transmission [14]. In contrast, S2T systems, such as DeepSC-ST [19], focus on extracting text-related features to minimize bandwidth usage. Recently, the EESC-S framework attempted to integrate emotion recognition to enhance reconstruction quality [20]. However, these approaches typically employ simple concatenation strategies to combine semantic and emotional features. They ignore the distinct characteristics of discrete linguistic information and continuous affective attributes, leading to inefficient bandwidth utilization and insufficient protection of emotional features against channel noise.

For speech reconstruction at the receiver, advanced generation technologies are essential to meet the high-fidelity interaction requirements of embodied artificial intelligence and the Internet of Vehicles [21]. Conventional methods often rely on Mean Squared Error (MSE) loss for decoding, which tends to produce over-smoothed spectrograms and reduces the perceptual quality of the synthesized speech [15]. To address this limitation, diffusion probabilistic models have been introduced to synthesize high-fidelity waveforms by iteratively refining the signal from noise [22–24]. Despite their potential, applying diffusion models in semantic communication remains challenging. The semantic features received from wireless channels inevitably contain noise and distortion. Existing methods fail to effectively condition the generative process on these noisy heterogeneous features, resulting in a degradation of both semantic accuracy and emotional naturalness in the reconstructed speech.

In summary, existing speech semantic communication frameworks exhibit critical limitations that impede high-fidelity applications. Semantic-centric systems typically prioritize linguistic content extraction, inadvertently discarding paralinguistic cues essential for human-centric interaction. Concurrently, while some emotion-aware models attempt to incorporate affective features, they often employ a homogeneous transmission strategy that fails to account for the distinct data characteristics of discrete text and continuous emotion. This mismatch leads to either semantic ambiguity due to channel noise or insufficient expressiveness due to quantization errors. Consequently, there is a lack of solutions that simultaneously achieve precise semantic understanding and robust emotional preservation, which is a key requirement for emerging embodied artificial intelligence systems [21].

To address this multifaceted challenge, we propose theHASCom. Unlike conventional monolithic architectures, HASCom introduces a heterogeneous dual-stream transmission paradigm. It strategically decouples the information flow, utilizing reliable digital coding for discrete semantic content and deep analog transmission for continuous affective features. At the receiver, a conditional diffusion model leverages the high-quality semantic prior to guide the generation of acoustic details, thereby bridging the gap between semantic precision and affective fidelity.

The main contributions of this study are as follows:

- We propose a heterogeneous semantic-affective transmission architecture. This mechanism transmits discrete linguistic features via conventional bit-level reliable transmission to ensure semantic accuracy, while mapping continuous emotion and timbre features directly into analog channel symbols via JSCC. This design avoids quantization loss for paralinguistic information while guaranteeing the correctness of the textual content.
- We develop a prior-guided diffusion reconstruction module. This module utilizes the accurate semantic features decoded from the digital channel to establish a reliable structural prior. This prior then guides the diffusion model to synthesize high-fidelity speech conditioned on the recovered analog affective cues, effectively restoring signal details even under noisy channel conditions.
- We formulate an end-to-end optimization objective and conduct extensive experiments. The results demonstrate that the proposed HASCom system significantly outperforms state-of-the-art baselines, achieving superior performance in terms of both speech quality and emotional consistency in bandwidth-constrained and noisy environments.

The remainder of this paper is organized as follows. Section 2 describes the system model and the formulation of the heterogeneous semantic-affective communication framework. Section 3 details the HASCom architecture, including the dual-stream decoupling transmission and the prior-guided diffusion reconstruction module. Experimental results, performance analysis, and comparative studies are presented in Section 4. Finally, Section 5 concludes the paper and discusses future research directions.

2. System Model

In this section, we propose the Heterogeneous Affective Speech Semantic Communication System (HASCom), which establishes a novel heterogeneous dual-stream transmission architecture. By strategically bifurcating the information flow into a *discrete semantic stream* and a continuous affective stream, the system reconciles the fundamental trade-off between semantic precision and affective fidelity. Unlike monolithic neural transmission schemes that treat all features uniformly, HASCom ensures that the reconstructed speech maintains rigorous semantic intelligibility via robust digital coding, while simultaneously exhibiting rich emotional fidelity preserved through deep analog transmission. As illustrated in Fig. 1, HASCom comprises two core stages: the Heterogeneous Semantic-Affective Transmission module and the Prior-Guided Affective Diffusion Reconstruction module. The transmission module employs a hybrid protocol, utilizing a reliable digital transmitter to convey quantized linguistic content (Z_c) for rigid semantic preservation, while employing an Analog JSCC Encoder to map affective embeddings (Z_e, Z_s) directly into continuous channel symbols, thereby preventing the quantization loss of granular prosodic nuances. At the receiver, the reconstruction module orchestrates a Skeleton-and-Flesh generative paradigm, where the Structural Prior Network anchors the semantic skeleton (μ) and the Conditional Diffusion model dynamically paints the emotional flesh based on the recovered analog affective cues ($\hat{Z}_a$), achieving high-fidelity speech reconstruction resilient to channel impairments.

https://doi.org/10.3390/1010000

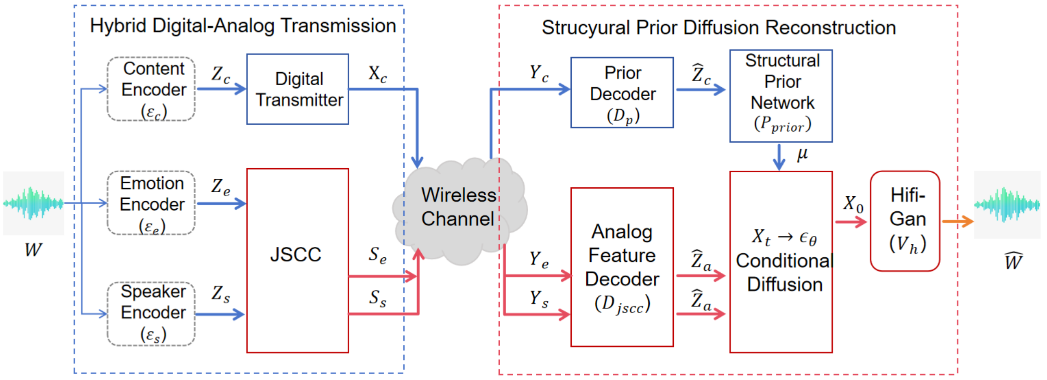
Hybrid Digital-Analog Transmission
Content Encoder (ε_c)
Z_c
Digital Transmitter
X_c
Emotion Encoder (ε_e)
Z_e
JSCC
S_e
Speaker Encoder (ε_s)
Z_s
S_s
W
Wireless Channel
Strucyural Prior Diffusion Reconstruction
Y_c
Prior Decoder (D_p)
$\widehat{Z}_c$
Structural Prior Network (P_{prior})
μ
Y_e
Y_s
Analog Feature Decoder (D_{jscc})
$\widehat{Z}_a$
$\widehat{Z}_a$
$X_t \to \epsilon_\theta$ Conditional Diffusion
X_0
Hifi-Gan (V_h)
$\widehat{W}$

Figure 1. Overall architecture of the proposed Heterogeneous Affective Speech Semantic Communication System (HASCom). The framework decouples speech into a discrete semantic stream and a continuous affective stream for heterogeneous transmission. At the receiver, the structural prior network establishes semantic alignment, while the recovered affective features guide the diffusion model to synthesize high-fidelity speech with precise intelligibility and robust emotional expressiveness.

2.1. *Hybrid Digital-Analog Transmission* 133

Let $W \in \mathbb{R}^L$ denote the raw input speech waveform of length L. The system first decomposes W into disentangled latent representations corresponding to linguistic content, emotional state, and speaker identity. The multi-modal feature extraction is formulated as: 134 135 136

$$Z_c = \mathcal{E}_c(W; \omega_c), \quad Z_e = \mathcal{E}_e(W; \omega_e), \quad Z_s = \mathcal{E}_s(W; \omega_s), \tag{1}$$ 137

where $\mathcal{E}_c$, $\mathcal{E}_e$, and $\mathcal{E}_s$ denote the content, emotion, and speaker encoders parameterized by ω_c, ω_e, and ω_s, respectively. Here, $Z_c \in \mathbb{R}^{D_c}$ represents the discrete phoneme-level linguistic features, while $Z_e \in \mathbb{R}^{D_e}$ and $Z_s \in \mathbb{R}^{D_s}$ represent the continuous embedding vectors for emotion and speaker timbre. 138 139 140 141

2.1.1. Discrete Semantic Link 142

To eliminate semantic ambiguity caused by channel noise, the linguistic features Z_c are transmitted via a reliable digital link. The features are first quantized and source-encoded into a binary bit sequence, followed by channel coding and digital modulation. This process generates the discrete digital transmit symbols X_c: 143 144 145 146

$$X_c = \mathcal{M}(\mathcal{C}(\mathcal{Q}(Z_c))) \in \mathbb{C}^{N_c}, \tag{2}$$ 147

where $\mathcal{Q}(\cdot)$ denotes the quantization operation, $\mathcal{C}(\cdot)$ represents the channel coding, $\mathcal{M}(\cdot)$ is the digital modulation function, and N_c is the frame length of the digital content symbols. 148 149

2.1.2. Continuous Affective Link 150

For paralinguistic information, which exhibits high redundancy and requires graceful degradation, we employ analog transmission. To capture the full spectrum of expressive attributes, we concatenate the emotion vector Z_e and the speaker vector Z_s to form a unified expressive style latent Z_a. Subsequently, this latent vector is mapped directly to continuous-valued analog channel symbols S_a via a deep neural network-based Joint Source-Channel Coding (JSCC) encoder $\mathcal{E}_{jscc}$: 151 152 153 154 155 156

$$S_a = \mathcal{E}_{jscc}(Z_a; \omega_j) \in \mathbb{C}^{N_a}, \tag{3}$$ 157

where ω_j denotes the parameters of the JSCC encoder, and N_a represents the dimension of the analog affective symbols. This analog mapping preserves the infinite resolution of emotional intensity that digital quantization might otherwise destroy.

The encoded symbols X_c and S_a are transmitted over the wireless channel. To explicitly model the distinct physical propagation characteristics and processing logic for the digital and analog streams, we define the received signals Y_c and Y_a separately:

$$\begin{aligned} Y_c &= h_c X_c + n_c, \\ Y_a &= h_a S_a + n_a, \end{aligned} \tag{4}$$

where h_c, h_a represent the channel fading coefficients for the content and affective links respectively, and $n_c, n_a \sim \mathcal{CN}(0, \sigma^2 I)$ denote the additive white Gaussian noise (AWGN). The receiver employs a heterogeneous decoding strategy to recover the latent features from these two distinct signal streams.

2.1.3. Robust Semantic Restoration

For the discrete stream Y_c, the reception process aims to recover the rigid linguistic skeleton of the speech. Drawing inspiration from the generative diffusion logic of Prior decoder, we formulate the semantic restoration not merely as hard bit-decoding, but as a robust structural decoding process. We define a Prior Decoder, denoted as $\mathcal{D}_p$, to map the received noisy digital symbols back to the discrete linguistic latent space:

$$\hat{Z}_c = \mathcal{D}_p(Y_c; \psi_c), \tag{5}$$

where ψ_c represents the parameters of the decoder, which conceptually encapsulates demodulation and error-correction mechanisms.

Physically implemented via a robust digital receiver chain, this process focuses on the accurate retrieval of discrete symbols. Mathematically abstracting this as $\mathcal{D}_p$ emphasizes its function in stripping away channel noise and restoring the integrity of the linguistic content from Y_c. This decoder ensures that the output $\hat{Z}_c$ faithfully preserves the original phoneme-level semantic information, providing a noise-free and high-fidelity linguistic foundation for the subsequent structural alignment and generation steps.

2.1.4. Affective Feature Reconstruction via Neural Inversion

Conversely, the continuous stream Y_a carries the paralinguistic style information, which requires a decoding capability that is resilient to channel variations. We employ a specialized deep neural feature decoder, denoted as $\mathcal{D}_{jscc}$, to handle this analog signal. Functioning as a non-linear inverse operator to the JSCC encoder, $\mathcal{D}_{jscc}$ extracts the expressive latent vector from the noisy channel symbols:

$$\hat{Z}_a = \mathcal{D}_{jscc}(Y_a; \psi_j), \tag{6}$$

where ψ_j represents the trainable parameters of the decoder. This network is trained to implicitly perform joint channel equalization and source denoising, effectively filtering out channel noise n_a while preserving the fine-grained values of the emotion and speaker embeddings. The recovered $\hat{Z}_a$ is subsequently sliced to retrieve the explicit emotion estimate $\hat{Z}_e$ and speaker estimate $\hat{Z}_s$, which will serve as the style conditions for the subsequent diffusion generation.

https://doi.org/10.3390/1010000

2.1.5. Structural Prior Generation

The reconstruction begins by establishing a semantic time-frequency skeleton. We utilize a Structural Prior Network, denoted as $\mathcal{P}_{prior}$, which incorporates a text encoder and a Monotonic Alignment Search (MAS) mechanism. Taking the robustly decoded linguistic features $\hat{Z}_c$ as input, it predicts the mean μ of the terminal Gaussian distribution for the diffusion process:

$$\mu = \mathcal{P}_{prior}(\hat{Z}_c; \psi_p) \in \mathbb{R}^{M \times T}, \tag{7}$$

where M is the number of Mel-frequency bands and T is the temporal duration. Here, μ provides the structural alignment of phonetic content, serving as the semantic anchor for the generation process.

$$\mu = \mathcal{P}_{prior}(\hat{Z}_c; \psi_p) \in \mathbb{R}^{M \times T}, \tag{8}$$

where M is the number of Mel-frequency bands and T is the temporal duration. Here, μ provides the structural alignment of phonetic content, serving as the semantic anchor for the generation process.

2.1.6. Affective-Guided Reverse Diffusion

Subsequently, the conditional diffusion generator, denoted as $\mathcal{G}_d$, leverages the recovered expressive style latent $\hat{Z}_a$ (encompassing both emotion and speaker attributes) as the primary conditioning input to modulate the reverse denoising trajectory. Unlike unconditional generation, here the network effectively navigates the latent manifold by conditioning on the heterogeneous priors. The prediction of the noise component ϵ_θ at each diffusion time step t is mathematically formulated as:

$$\epsilon_\theta = \mathcal{G}_d(\mathcal{X}_t, \mu, t, \hat{Z}_a; \psi_d) \in \mathbb{R}^{M \times F}, \tag{9}$$

where ψ_d represents the learnable parameters of the neural generator, and $\mathcal{X}_t$ denotes the noisy Mel-spectrogram state at time t.

In this formulation, the mechanism of explicit condition injection plays a pivotal role. The recovered style vector $\hat{Z}_a$ is projected and injected into the denoising network (e.g., via adaptive layer normalization or cross-attention mechanisms). This ensures that the generated speech $\mathcal{X}_0$ strictly retains the intricate prosody, intonation, and timbre transmitted via the *continuous affective link*, while simultaneously adhering to the rigid linguistic alignment constraints defined by the semantic structural prior μ. The prior μ acts as the semantic skeleton while $\hat{Z}_a$ provides the emotional flesh, achieving a synergistic reconstruction of the target speech.

Finally, the denoised Mel-spectrogram $\mathcal{X}_0$, obtained at the end of the reverse diffusion process, is transformed into the time-domain waveform to complete the end-to-end communication loop. We employ a pretrained HiFi-GAN vocoder, denoted as $\mathcal{V}_h$, to synthesize the high-fidelity waveform $\hat{W}$ from the spectrogram:

$$\hat{W} = \mathcal{V}_h(\mathcal{X}_0) \in \mathbb{R}^L. \tag{10}$$

This non-autoregressive vocoding stage ensures that the final output preserves the perceptual quality and phase continuity of the original speech signal.

2.2. End-to-End Optimization Objective

The optimization objective of the Heterogeneous Affective Speech Semantic Communication System (HASCom) is to minimize the discrepancy between the source speech waveform W and the reconstructed waveform $\hat{W}$ over the stochastic channel conditions. Let Φ denote the end-to-end composite mapping function parameterized by the set of

https://doi.org/10.3390/1010000

learnable parameters θ, such that the reconstruction process is defined as $\hat{W} = \Phi(W, H; \theta)$, where H represents the channel state. Consequently, the system training is formulated as an optimization problem to minimize the expected loss $\mathcal{L}$ with respect to θ:

$$\min_{\theta} \mathbb{E}[\mathcal{L}(W, \hat{W})]. \tag{11}$$

To effectively guide the model convergence, we formulate the system design as an end-to-end optimization problem. Our global objective is to minimize a composite loss function $\mathcal{L}$, which serves as a tractable surrogate for the perceptual distance between the original and reconstructed speech. This objective is formulated as a weighted sum of four distinct components, each governing a specific physical dimension of the speech signal:

$$\mathcal{L} = \mathcal{L}_d + \lambda_a \mathcal{L}_a + \lambda_s (\mathcal{L}_s + \mathcal{L}_t), \tag{12}$$

where λ_a and λ_s are hyperparameters balancing the trade-off between generative quality, transmission robustness, and structural alignment.

The components of the objective function are designed to constrain the system across multiple modalities. The diffusion denoising loss $\mathcal{L}_d$ governs the generative link, ensuring that the diffusion model synthesizes high-fidelity acoustic details conditioned on the heterogeneous features. Simultaneously, the analog consistency loss $\mathcal{L}_a$ ensures the robustness of the affective link by minimizing the distortion of the continuous emotion and speaker embeddings under channel noise. To maintain linguistic accuracy, the structural alignment loss $\mathcal{L}_s$ enforces the correctness of the semantic skeleton by penalizing misalignment between the linguistic prior and the ground truth spectrogram, while the temporal duration loss $\mathcal{L}_t$ constrains the rhythm of the speech, ensuring that the predicted phoneme durations match the target prosody. By jointly minimizing $\mathcal{L}$, the system effectively harmonizes the rigid semantic structure with the fluid affective flow, achieving robust and expressive speech communication.

3. Heterogeneous Affective Speech Semantic Communication System

In this section, we provide a comprehensive description of the proposed Heterogeneous Affective Speech Semantic Communication System (HASCom). As depicted in Fig. 1, HASCom establishes a novel Skeleton-and-Flesh communication paradigm. Unlike conventional frameworks that treat all speech features uniformly or rely solely on digital coding, HASCom strategically disentangles speech into a rigid semantic skeleton and fluid affective flesh, transmitting them via heterogeneous digital and analog streams, respectively. This architecture ensures that the reconstructed speech maintains rigorous semantic intelligibility while exhibiting rich emotional fidelity. In the following subsections, we detail the core components: the heterogeneous semantic-affective transmission module, the prior-guided affective diffusion reconstruction module, and the corresponding two-stage optimization strategy.

3.1. Heterogeneous Semantic-Affective Feature Encoding

To ensure the precise delivery of linguistic content while preserving the rich nuances of emotional expression, we propose a Heterogeneous Semantic-Affective Feature Encoding mechanism at the transmitter side. Unlike conventional approaches that fuse all features into a unified embedding for homogeneous transmission , this mechanism strategically bifurcates the encoding process based on the physical nature of the features, generating a robust digital stream for semantics and a high-resolution analog stream for affect.

To implement the heterogeneous transmission paradigm, the system decomposes the raw speech waveform u into disentangled latent representations. We design specialized extraction pathways that map the high-dimensional temporal signal into compact semantic and affective embeddings, utilizing state-of-the-art self-supervised encoders as the backbone feature extractors.

3.1.1. Semantic Content Extraction

To ensure the rigorous preservation of linguistic content independent of speaker identity and emotional prosody, we employ the Wav2Vec 2.0 framework [25] as the backbone content encoder. This module functions as a linguistic filter, processing the raw waveform u to extract a sequence of frame-level discrete representations. Unlike conventional spectral features, these deep latent variables capture the contextual phonetic structure of the speech while being robust to channel variations. Mathematically, the semantic extraction process maps the time-domain signal to a downsampled linguistic feature space via the encoder function $\mathcal{E}_c$, expressed as:

$$Z_c = \mathcal{E}_c(u; \omega_c) \in \mathbb{R}^{T_c \times D_c}, \tag{13}$$

where ω_c denotes the pre-trained parameters, T_c represents the effective temporal resolution aligned with phonemic boundaries, and D_c is the feature dimension. These representations Z_c constitute the "semantic skeleton," providing the rigid structural foundation for the subsequent digital transmission.

3.1.2. Unified Affective Extraction

Conversely, to capture the paralinguistic flesh which inherently entangles time-invariant speaker timbre and time-variant prosodic intensity, we propose a unified affective extraction strategy based on the WavLM architecture [26]. Instead of explicitly separating emotion and speaker identity which may lead to feature accumulation errors, we utilize this comprehensive model to distill a global style descriptor.

The extraction mechanism aggregates the disentangled paralinguistic features into a unified continuous embedding suitable for analog transmission. To preserve the complete information of both emotional intensity and speaker identity without premature mixing or compression, we construct the unified affective latent Z_a via a concatenation operation:

$$Z_a = [Z_e \oplus Z_s] \in \mathbb{R}^{D_e + D_s}, \tag{14}$$

where $\oplus$ denotes the vector concatenation operator, and $Z_e \in \mathbb{R}^{D_e}$ and $Z_s \in \mathbb{R}^{D_s}$ are the extracted emotion and speaker embeddings, respectively. This composite vector Z_a is subsequently fed into the JSCC encoder, ensuring that the continuous channel symbols encapsulate both the global prosody and timbre with high fidelity.

3.2. Channel Transmission Model

The HASCom framework operates over a hybrid wireless environment where the discrete semantic skeleton and the continuous affective flesh are transmitted via physically distinct logic. We model the wireless propagation for both streams assuming a Additive White Gaussian Noise (AWGN).

3.2.1. Digital Semantic Transmission

To guarantee the rigorous recoverability of the linguistic skeleton, the quantized semantic features are transmitted via a classic digital link protected by Low-Density Parity-Check (LDPC) codes. This traditional coding scheme provides a hard reliability guarantee, correcting bit-level errors up to the Shannon limit. The received digital signal Y_c is modeled as:

$$Y_c = h_c S_c + n_c, \tag{15}$$

where $S_c \in \mathbb{C}^{N_c}$ denotes the QAM-modulated symbols derived from the LDPC-coded bits, $h_c \sim \mathcal{CN}(0,1)$ represents the complex Rayleigh fading coefficient, and $n_c \sim \mathcal{CN}(0,\sigma_c^2 I)$ is the Gaussian noise vector.

3.2.2. Deep Analog Affective Transmission (JSCC)

Unlike the discrete semantic stream, the affective features Z_a (extracted by the large-scale WavLM backbone) represent continuous-valued deep representations where the Euclidean distance directly correlates with perceptual similarity. Traditional digital coding would introduce irreversible quantization noise and suffer from the cliff effect when channel conditions deteriorate. Therefore, we employ a Deep JSCC strategy.

The JSCC encoder $\mathcal{E}_{jscc}$ functions as a neural projection network, mapping the high-dimensional affective latent Z_a directly into a sequence of complex-valued channel input symbols. To ensure strict adherence to the hardware transmission power constraints and to maximize energy efficiency, we impose a global power normalization layer immediately following the neural mapping. The transmitted analog symbols S_a are obtained by:

$$\tilde{S}_a = \mathcal{E}_{jscc}(Z_a; \omega_j), \quad S_a = \sqrt{P}\frac{\tilde{S}_a}{\sqrt{\mathbb{E}[\|\tilde{S}_a\|^2]}}, \tag{16}$$

where P is the average transmission power constraint, and the normalization ensures $\mathbb{E}[\|S_a\|^2] = P$. This operation aligns the variance of the deep features with the physical channel capacity, preventing signal saturation or attenuation.

The signal propagates through the analog fading channel, arriving at the receiver as:

$$Y_a = h_a S_a + n_a, \tag{17}$$

where h_a and n_a denote the fading and noise components for the analog link. In this paradigm, the neural decoder $\mathcal{D}_{jscc}$ at the receiver side acts as a non-linear Minimum Mean Square Error (MMSE) estimator. By jointly optimizing the encoder and decoder, the system learns to map the most critical affective syntax to the high-SNR subspaces of the channel, achieving graceful degradation where channel noise results in subtle emotional shifts rather than catastrophic decoding failures.

3.3. Diffusion Model

3.3.1. Theoretical Foundation

We adapt the score-based generative framework to facilitate robust speech reconstruction. In contrast to standard unconditional diffusion models that initiate the reverse process from a standard isotropic Gaussian distribution $\mathcal{N}(0, I)$, our approach leverages the linguistic information recovered from the digital link to construct a structural prior. Specifically, we utilize the semantic skeleton to define the mean μ of the terminal distribution. This strategy provides an informative starting point for the generative process, allowing the model to focus on refining fine-grained acoustic details and emotional prosody rather than synthesizing linguistic structure from pure noise.

Formally, we model the forward degradation process not as a destruction into random noise, but as a mean-reverting Ornstein-Uhlenbeck (OU) stochastic differential equation. Given the ground-truth Mel-spectrogram $\mathcal{X}_0$, the state evolution at time t is governed by:

$$d\mathcal{X}_t = \frac{1}{2}\beta_t(\mu - \mathcal{X}_t)dt + \sqrt{\beta_t}d\mathbf{w}_t, \quad t \in [0, T], \tag{18}$$

where β_t represents the non-negative noise schedule and $\mathbf{w}_t$ denotes the standard Wiener process. Crucially, μ is the acoustic structural prior, which is deterministically predicted

https://doi.org/10.3390/1010000

from the decoded semantic features $\hat{Z}_c$ via the Structural Prior Network. Under this formulation, as t approaches T, the distribution of the noisy state $\mathcal{X}_T$ converges to $\mathcal{N}(\mu, I)$. This implies that the terminal distribution of the forward process is anchored by the linguistic content rather than a standard Gaussian distribution.

Consequently, the generation of high-fidelity speech corresponds to the reverse-time solution of the SDE. To ensure the reconstruction possesses the correct emotional intensity and speaker timbre, we formulate the reverse dynamics as a probability flow Ordinary Differential Equation (ODE) conditioned on the analog affective cues. The deterministic trajectory evolving from the structural prior $\mathcal{X}_T$ back to the data distribution $\mathcal{X}_0$ is defined as:

$$d\mathcal{X}_t = \frac{1}{2}\beta_t\left[(\mu - \mathcal{X}_t) - \epsilon_\theta(\mathcal{X}_t, \mu, t, \hat{Z}_a)\right]dt, \tag{19}$$

where $\epsilon_\theta(\mathcal{X}_t, \mu, t, \hat{Z}_a)$ approximates the score function $\nabla_{\mathcal{X}_t} \log p_t(\mathcal{X}_t|\hat{Z}_a)$.

In this theoretical framework, the term $(\mu - \mathcal{X}_t)$ functions as a conservative force field that anchors the generative trajectory to the semantic skeleton, ensuring linguistic intelligibility. Simultaneously, the neural score estimator ϵ_θ, conditioned on the recovered affective features $\hat{Z}_a$, introduces the necessary gradient field to synthesize the fine-grained prosodic and timbral details. This formulation mathematically realizes the proposed Skeleton-and-Flesh paradigm, where the structural prior provides the geometric boundary and the diffusion model fills in the affective texture.

3.3.2. Structural Prior Network

The initial phase of the generative reconstruction focuses on establishing the deterministic semantic anchor that defines the terminal state of the proposed short-cut diffusion trajectory. To transform the non-isochronous linguistic features $\hat{Z}_c$ into the time-synchronized acoustic skeleton μ, we formulate the alignment process as a latent variable optimization problem solved via Monotonic Alignment Search (MAS).

The Structural Prior Network $\mathcal{P}_{prior}$ first projects the digital semantic inputs into a sequence of unaligned distribution parameters $\tilde{\mu}$. The MAS algorithm then identifies the optimal monotonic path A^* that maximizes the log-likelihood of the ground-truth spectrogram $\mathcal{X}_0$ being generated from this prior distribution. Mathematically, this process aligns the discrete phoneme frames to the continuous acoustic time axis, yielding the aligned mean:

$$\mu = \text{Align}(\tilde{\mu}, A^*). \tag{20}$$

We explicitly train this aligned mean to serve as the geometric center of the diffusion terminal distribution, postulating that $\mathcal{X}_T \sim \mathcal{N}(\mu, I)$. Consequently, the primary optimization objective is to minimize the Euclidean distance between this generated skeleton and the ground truth, formulated as the Structural Alignment Loss:

$$\mathcal{L}_s = \|\mu - \mathcal{X}_0\|_2^2. \tag{21}$$

Simultaneously, to ensure the system can autonomously determine the temporal structure during inference without ground-truth guidance, we jointly optimize an internal duration predictor to approximate the optimal token durations. The target durations d_{mas} are derived by aggregating the alignment path A^* along the time axis. We impose the Temporal Duration Loss on the predicted durations d_{pred} in the logarithmic domain to robustly handle the dynamic range of speech rhythm:

$$\mathcal{L}_t = \left\|\log(d_{pred}) - \log(d_{mas})\right\|_2^2. \tag{22}$$

By minimizing this objective, the network learns to construct a stable, time-aligned acoustic skeleton μ solely from the digital linguistic stream, providing the necessary deterministic boundary condition for the subsequent affective diffusion process.

3.3.3. Affective-Guided Diffusion Optimization

To reconstruct semantically faithful and emotionally expressive speech, we propose an emotion-conditioned diffusion architecture that explicitly embeds the recovered affective features as dominant guidance signals in the reverse generative process. The core of this module is the neural score estimator ϵ_θ, implemented as an improved U-Net architecture tailored to synergize the rigid semantic skeleton with the fluid affective flesh.

To strictly enforce the linguistic alignment constraints, we adopt a structural concatenation strategy at the network input. The aligned semantic skeleton μ, derived from the prior network, is concatenated with the noisy intermediate variable $\mathcal{X}_t$ along the channel dimension. This combined tensor forms the diffusion input, anchoring the generative search space and allowing the network to focus its capacity on refining acoustic textures rather than inferring geometries from scratch. Simultaneously, to embed the continuous affective flesh, we propose a multi-level conditioning mechanism. The recovered analog features $\hat{Z}_a$ are projected through a Multi-Layer Perceptron (MLP) to generate time-aware style embeddings. These embeddings are then injected into each residual block of the U-Net, dynamically modulating the internal feature maps. This design ensures that the speaker timbre and emotional intensity are deeply integrated into the generation process.

With the neural estimator parameterized, the generative reconstruction corresponds to solving the reverse-time probability flow Ordinary Differential Equation (ODE). Starting from the terminal distribution centered at the semantic skeleton, the trajectory evolves deterministically towards the clean data manifold, governed by the network's gradient estimation:

$$d\mathcal{X}_t = \frac{1}{2}\beta_t\left[(\mu - \mathcal{X}_t) - \epsilon_\theta(\mathcal{X}_t, \mu, t, \hat{Z}_a)\right]dt, \tag{23}$$

where β_t represents the noise schedule. In this dynamic equation, the drift term $(\mu - \mathcal{X}_t)$ acts as a conservative structural force maintaining linguistic intelligibility, while the neural term $\epsilon_\theta(\cdot)$ provides the "affective gradient" that effectively renders the fine-grained prosodic details onto the semantic skeleton.

The optimization objective is to train the network parameters θ to accurately approximate this score function. Consistent with the theoretical formulation of the mean-reverting stochastic differential equation, we formulate the diffusion denoising loss $\mathcal{L}_d$ as the weighted expected mean squared error between the predicted noise term and the actual Gaussian noise ξ injected during the forward process:

$$\mathcal{L}_d = \mathbb{E}_{t,\mathcal{X}_0,\xi}\left[\lambda_t \left\|\epsilon_\theta(\mathcal{X}_t, \mu, t, \hat{Z}_a) - \xi\right\|_2^2\right], \tag{24}$$

where t is sampled uniformly from the time horizon $[0, T]$, $\mathcal{X}_t$ is sampled from the forward transition kernel, and λ_t denotes the weighting coefficient. By minimizing $\mathcal{L}_d$, the model learns to utilize the analog affective cues $\hat{Z}_a$ to reconstruct the high-fidelity acoustic details missing from the digital skeleton.

3.4. Training Strategy

3.5. Analog Affective Transmission Optimization

The training process of HASCom is divided into two decoupled stages to ensure stability. The first stage focuses exclusively on establishing a robust analog channel for the affective features. The detailed optimization procedure is outlined in Algorithm 1.

Algorithm 1 Stage I: Training Procedure for the Analog Affective Link

```
Input: Frozen Modules: Emotion Encoder $\mathcal{E}_e$, Speaker Encoder $\mathcal{E}_s$;
    Trainable Modules: JSCC Encoder $\mathcal{E}_{jscc}$, JSCC Decoder $\mathcal{D}_{jscc}$ (collectively $\theta_{jscc}$);
    Hyperparameters: Learning rate $\eta$, batch size $B$, transmit power $P$.
Output: Optimized Parameters $\theta_{jscc}$.
Initialize trainable parameters $\theta_{jscc}$ randomly.
repeat
    for each batch $W \sim \mathcal{D}$ (batch size $B$) do
        Extract emotion vectors: $Z_e = \mathcal{E}_e(W)$
        Extract speaker vectors: $Z_s = \mathcal{E}_s(W)$
        Concatenate latent features: $Z_a = [Z_e \oplus Z_s]$
        Map to channel symbols: $\tilde{S}_a = \mathcal{E}_{jscc}(Z_a)$
        Apply Power Normalization:
        $S_a = \sqrt{P} \cdot \frac{\tilde{S}_a}{\sqrt{\mathbb{E}\|\tilde{S}_a\|^2 + \epsilon}}$
        Sample channel state $h_a \sim \mathcal{CN}(0,1)$ and noise $n_a \sim \mathcal{CN}(0,\sigma^2)$
        Transmission: $Y_a = h_a \cdot S_a + n_a$
        Neural Inversion: $\hat{Z}_a = \mathcal{D}_{jscc}(Y_a)$
        Compute Reconstruction Loss:
        $\mathcal{L}_{trans} = \frac{1}{B} \sum_{i=1}^{B} \left\| Z_a^{(i)} - \hat{Z}_a^{(i)} \right\|_2^2$
        Update gradients: $\theta_{jscc} \leftarrow \theta_{jscc} - \eta \nabla_\theta \mathcal{L}_{trans}$
    end for
until convergence or max epochs reached
return $\theta_{jscc}$
```

3.5.1. Generative Reconstruction Optimization

Following the convergence of the analog transmission link, the second stage focuses on the joint optimization of the structural prior and the affective diffusion model. In this phase, the JSCC module serves as a frozen, noise-robust feature extractor. The training procedure is detailed in Algorithm 2.

3.6. End-to-End Training Protocol

To orchestrate the heterogeneous components into a cohesive system, we implement a sequential optimization protocol that mimics the physical causality of the communication link. This ensures that the generative receiver is effectively trained to handle both the complexity of speech synthesis and the stochasticity of channel distortions.

Initially, the system is trained in an ideal, noise-free environment to establish a robust generative baseline. In this phase, the semantic quantization and Deep JSCC modules are bypassed, feeding the ground-truth linguistic features Z_c and affective features Z_a directly into the reconstruction module. The structural prior network and the conditional diffusion model are jointly optimized to learn the fundamental mapping from heterogeneous latent representations to high-fidelity Mel-spectrograms. This pre-training strategy ensures that the generator captures the intrinsic correlation between the linguistic structure and the paralinguistic details without the interference of channel noise, preventing the model from collapsing due to early-stage signal distortion.

Building upon this stable generative foundation, we subsequently introduce the physical channel simulation for fine-tuning. The analog JSCC encoder is activated, injecting random channel noise n_a and fading h_a into the affective stream to simulate real-world transmission conditions. The pre-trained diffusion model is then adapted to these distorted features $\hat{Z}_a$. During this stage, the model shifts its focus from basic synthesis to robust compensation, learning to hallucinate missing details from the noisy analog cues and effectively bridging the gap between the clean semantic prior and the corrupted affective condition.

Algorithm 2 Stage II: Prior-Guided Affective Diffusion Optimization

```
Input: Frozen Modules: JSCC Encoder/Decoder $\theta_{jscc}$, Semantic Encoder $\mathcal{E}_c$;
    Trainable Modules: Structural Prior Network $\mathcal{P}_{prior}$ (including Duration Predictor),
    Diffusion U-Net $\epsilon_\theta$;
    Data: Speech dataset $\mathcal{D}$, Ground-truth Mel-spectrogram $\mathcal{X}_0$;
    Hyperparameters: Loss weights $\lambda_s, \lambda_t$, noise schedule $\beta_t$.
Output: Optimized Generative Parameters $\theta_{gen} = \{\mathcal{P}_{prior}, \epsilon_\theta\}$.
Initialize $\theta_{gen}$ randomly.
repeat
    for each batch $(u, \mathcal{X}_0) \sim \mathcal{D}$ do
        Digital Stream: Extract & Quantize semantic features:
            $\hat{Z}_c \leftarrow \text{Quantize}(\mathcal{E}_c(u))$
        Analog Stream: Transmit affective features via frozen JSCC:
            $Z_a = \text{Agg}(\mathcal{E}_a(u))$
            $\hat{Z}_a = \mathcal{D}_{jscc}(\text{Channel}(\mathcal{E}_{jscc}(Z_a)))$
        Predict unaligned distributions: $\tilde{\mu} = \mathcal{P}_{prior}(\hat{Z}_c)$
        Monotonic Alignment Search (MAS):
            $A^* = \text{MAS}(\tilde{\mu}, \mathcal{X}_0)$
        Get aligned semantic skeleton & durations:
            $\mu = \text{Align}(\tilde{\mu}, A^*), \quad d_{mas} = \text{Sum}(A^*)$
        Compute Structural Losses:
            $\mathcal{L}_s = \|\mu - \mathcal{X}_0\|^2, \quad \mathcal{L}_t = \|\log d_{pred} - \log d_{mas}\|^2$
        Sample time $t \sim [0, T]$ and noise $\xi \sim \mathcal{N}(0, I)$
        Sample noisy state $\mathcal{X}_t$ via OU forward process anchored at $\mu$:
            $\mathcal{X}_t = \rho_t \mathcal{X}_0 + (1 - \rho_t)\mu + \delta_t \xi$
        Predict noise residual with Affective FiLM conditioning:
            $\hat{\xi} = \epsilon_\theta(\mathcal{X}_t, \mu, t, \hat{Z}_a)$
        Compute Diffusion Loss:
            $\mathcal{L}_d = \|\hat{\xi} - \xi\|^2$
        Aggregate Total Loss:
            $\mathcal{L}_{total} = \mathcal{L}_d + \lambda_s(\mathcal{L}_s + \mathcal{L}_t)$
        Update gradients: $\theta_{gen} \leftarrow \theta_{gen} - \eta \nabla_\theta \mathcal{L}_{total}$
    end for
until convergence
return $\theta_{gen}$
```

In the reconstruction phase, the flow bifurcates and then converges. The structural prior network processes the semantic tokens to generate the acoustic skeleton μ, while the diffusion model utilizes the affective condition $\hat{Z}_a$ to refine this skeleton. Although the internal loss of the diffusion model is calculated on the noise residual, the global optimization objective is mathematically equivalent to minimizing the reconstruction error between the generated spectrogram and the original ground truth. To ensure the highest fidelity, our ultimate training goal is to minimize the divergence between the ground-truth spectrogram S (derived from W) and the reconstructed spectrogram $\hat{S}$ (equivalent to $\mathcal{X}_0$):

$$\min \mathbb{E}\left[\|S - \hat{S}\|_2^2\right]. \tag{25}$$

By minimizing this objective, the system ensures that the final output converges to the original spectral distribution, preserving both the linguistic content and the rich paralinguistic details.

4. Simulation

In this section, we conduct comprehensive experiments to evaluate the performance of the proposed HASCom framework. To demonstrate the superiority of the heterogeneous Skeleton-and-Flesh architecture, we compare HASCom against traditional digital communication schemes and state-of-the-art semantic-only communication systems. The evaluation focuses on three key aspects: semantic intelligibility under noisy conditions, affective fidelity preservation, and the necessity of the dual-stream heterogeneous design.

4.1. Experimental Setup

4.1.1. Datasets and Preprocessing

To train and evaluate the proposed framework, we utilize the Emotional Speech Dataset (ESD)[27]. This high-quality multi-speaker dataset is specifically designed for speech synthesis and emotion conversion tasks. The ESD dataset comprises recordings from 10 native English speakers and 10 native Chinese speakers, covering five distinct emotion categories: neutral, happy, angry, sad, and surprise. Each speaker contributes 350 parallel utterances per emotion category, providing sufficient data diversity in terms of both linguistic content and emotional variance.

For data preprocessing, all audio recordings are resampled to 16 kHz to ensure consistency across the system. We extract 80-dimensional Mel-spectrograms as acoustic features using the Short-Time Fourier Transform (STFT). The STFT configuration includes a 1024-point FFT size, a 256-point hop length, and a 1024-point Hann window. These Mel-spectrograms serve as the ground truth targets for the generative diffusion decoder. Furthermore, we employ the Montreal Forced Aligner (MFA) to extract phoneme-level alignments from the raw audio, which serve as the input for the semantic coding stream.

The proposed HASCom framework was implemented using the PyTorch library and trained on a single NVIDIA RTX 4090 GPU with 24 GB of video memory. The network parameters were optimized using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and an epsilon value of 10^{-9}. A constant learning rate of 1×10^{-4} was employed throughout the training process to ensure stable convergence of the joint source-channel coding module. The batch size was set to 32 to accommodate the memory constraints while maintaining accurate gradient estimation.

For the conditional diffusion decoder, the total number of diffusion steps T_{train} was set to 1000 during the training phase to capture the fine-grained acoustic details of the Mel-spectrograms. During the inference phase, a fast sampling strategy with $T_{infer} = 50$ steps was adopted to reduce computational latency. A temperature parameter $\tau = 1.3$ was applied to the stochastic noise generator to enhance the expressiveness of the synthesized speech. To ensure the robustness of the system against varying channel conditions, the model was trained under a dynamic SNR environment, where the signal-to-noise ratio was uniformly sampled from $[0, 20]$ dB. The detailed hyperparameter settings are summarized in Table 1.

Table 1. Hyperparameter settings of the proposed framework.

Parameter	**Value**	**Parameter**	**Value**
Total Epochs	200	Optimizer	Adam
Learning Rate	1×10^{-4}	Batch Size	32
Training Steps (T_{train})	1000	Inference Steps (T_{infer})	50
Min. Noise Level (β_{min})	0.05	Max. Noise Level (β_{max})	20.0
Temperature (τ)	1.3	Length Scale	0.91
λ_{rec}	1.0	λ_{diff}	1.0
Channel Type	AWGN	Training SNR	$U(0, 20)$ dB
FFT Size	1024	Hop Length	256

4.1.2. Baselines

To accurately evaluate the performance of HASCom, we compare it against three representative semantic communication baselines. These baselines represent the standard end-to-end learning paradigm and two typical cascaded architectures for handling noise:

DeepSC-SR (Deep Semantic Communication for Speech Reconstruction): This baseline represents the standard state-of-the-art JSCC-based speech semantic communication system (e.g., DeepSC-S). It employs a monolithic encoder-decoder architecture to map speech features directly to continuous analog channel symbols. It focuses on minimizing the end-to-end reconstruction loss without explicit disentanglement of semantic and affective features, serving as the fundamental anchor for neural speech transmission performance.

SE-DeepSC (Speech Enhancement + DeepSC): This baseline represents a Denoise-then-Transmit cascaded strategy designed to mitigate environmental noise. A dedicated Speech Enhancement (SE) module is deployed at the transmitter to filter out background noise from the raw waveform before the clean speech is fed into the DeepSC encoder. This baseline evaluates whether pre-processing the source signal can effectively improve the robustness of semantic transmission.

DeepSC-S-SE (DeepSC + Speech Enhancement): This baseline represents a Transmit-then-Denoise cascaded strategy. The speech is first transmitted and reconstructed via the standard DeepSC system, and subsequently, a Speech Enhancement (SE) module is applied at the receiver end as a post-processor. This setup aims to suppress channel-induced noise and artifacts in the synthesized speech, testing the efficacy of post-hoc restoration compared to our intrinsic generative denoising approach.

4.1.3. Channel Models

We simulate the physical layer transmission over Additive White Gaussian Noise (AWGN). The Signal-to-Noise Ratio (SNR) ranges from 0 dB to 18 dB, covering conditions from extreme interference to high-quality links.

4.2. Performance Evaluation Metrics

To evaluate the proposed HASCom framework, we utilize three key metrics to assess physical layer reliability, semantic consistency, and perceptual quality:

Feature Reconstruction MSE: We use the Mean Squared Error (MSE) to evaluate the reconstruction quality of the transmitted features. This metric calculates the squared difference between the original feature vectors generated by the encoder and the noisy features recovered by the decoder. A lower MSE value means that the system can effectively recover the semantic and emotional features even under noisy channel conditions.

Similarity Score: This metric measures the semantic and emotional consistency between the original and the reconstructed speech. It is computed based on the cosine similarity between the source emotion embeddings and the received ones. A higher similarity score indicates that the meaning and emotional intent of the speaker are well preserved during the communication process.

Mean Opinion Score (MOS): MOS is a subjective metric used to evaluate how natural the speech sounds to human ears. In our experiments, human listeners were invited to rate the generated audio samples on a scale from 1 (Bad) to 5 (Excellent). A higher MOS score indicates that the reconstructed speech sounds clearer and closer to natural human recording.

https://doi.org/10.3390/1010000

4.3. Results and Analysis

4.3.1. Training Convergence Analysis of Analog Transmission

To validate the learnability and optimization stability of the proposed Deep Analog Transmission module, we monitor the training loss curve of the JSCC encoder-decoder network. The convergence behavior is illustrated in Fig. 2, which depicts the Mean Squared Error (MSE) loss of the affective feature reconstruction over 100 training epochs.

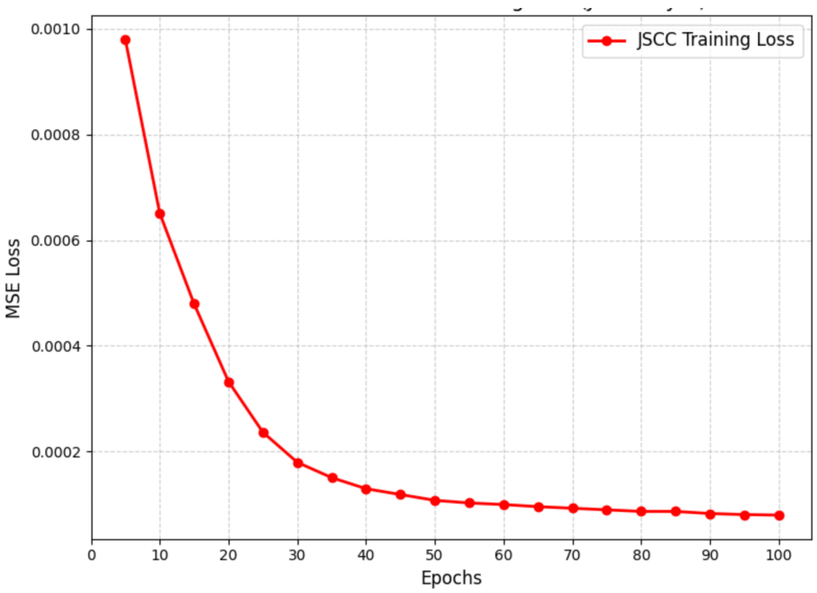

Figure 2. The training convergence curve of the Deep Analog JSCC module.

The curve demonstrates exceptional optimization efficiency for the heterogeneous affective link. As shown in Fig. 2, the training loss exhibits a sharp descent within the first 20 epochs, dropping from approximately 1.0×10^{-3} to 3.3×10^{-4}. This rapid initial convergence indicates that the feature adapter and JSCC encoder can efficiently learn to map the high-dimensional emotion and speaker embeddings into a compact, robust analog constellation without suffering from gradient vanishing.

Subsequently, the training enters a stable fine-tuning phase. From epoch 30 to 100, the error decays monotonically and smoothly, eventually converging to a negligible noise floor of approximately 7.9×10^{-5}. The absence of significant oscillations or spikes throughout the process confirms the numerical stability of the proposed architecture. This result theoretically guarantees that the analog affective link can provide accurate and consistent conditioning signals for the downstream generative tasks, serving as a reliable foundation for the entire HASCom system.

Following this initial phase, the model enters a stable refinement stage. From epoch 30 to 200, the loss decreases monotonically and smoothly, eventually stabilizing at a low residual error of approximately 0.96. The absence of significant oscillations or divergence confirms that the heterogeneous conditioning mechanism—combining discrete linguistic priors with continuous affective embeddings—provides robust and consistent guidance for the denoising process. This stable convergence ensures that the HASCom receiver can reliably synthesize high-fidelity speech with rich emotional details, validating the feasibility of the proposed generative semantic communication paradigm.

4.3.2. Semantic Consistency Analysis

Next, we assess the system's ability to maintain semantic consistency using the Similarity metric. Fig. 3 compares the performance of the proposed HASCom framework against three semantic communication baselines: DeepSC-SR, SE-DeepSC, and DeepSC-S-SE.

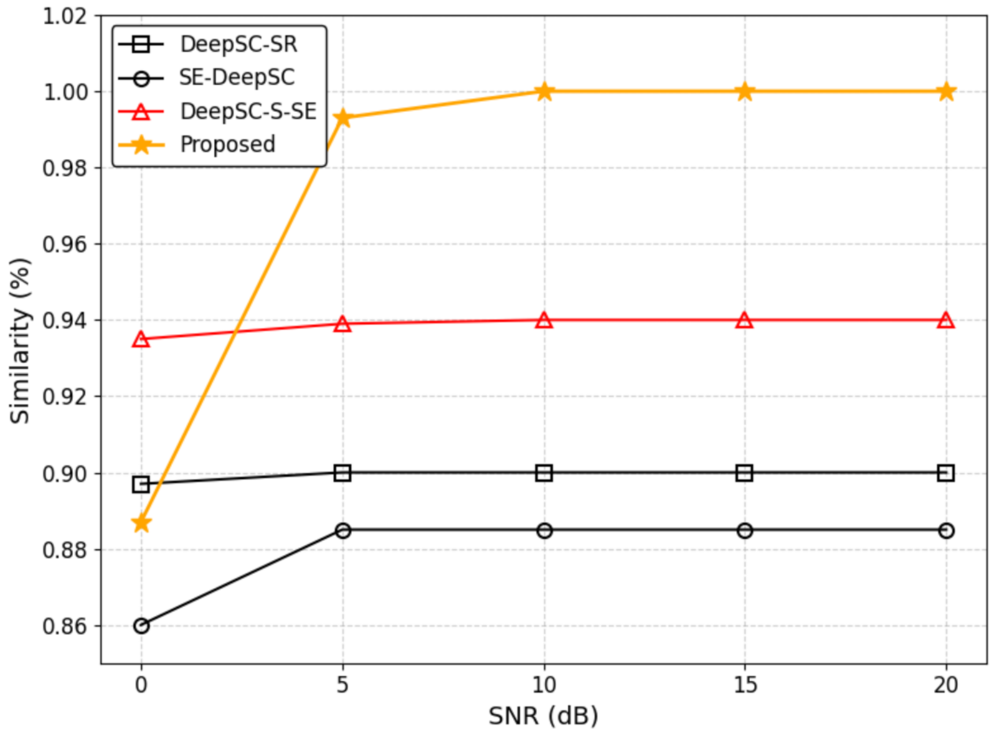

Figure 3. Performance comparison of multiple models in terms of Similarity under different AWGN noise levels

The comparison reveals a fundamental divergence in performance characteristics between the proposed heterogeneous architecture and the baseline models.

As shown in Fig. 3, the baseline methods suffer from a performance ceiling, saturating between 88% and 94%. This limitation arises because they rely on end-to-end analog transmission for the entire speech representation, where channel noise inevitably degrades the discrete linguistic information. In contrast, the proposed HASCom framework demonstrates a distinct threshold behavior driven by its digital linguistic skeleton. In the low SNR regime (0 dB), the similarity score dips to approximately 89%, slightly lower than the DeepSC-S-SE baseline. This is attributed to the residual bit errors in the digital stream before the channel coding fully converges. However, a sharp performance leap is observed as the SNR improves. Once the SNR exceeds 5 dB—surpassing the decoding threshold of the digital channel—the similarity score rapidly ascends to nearly 100%, significantly outperforming all baselines. The superiority of the proposed HASCom framework is directly attributed to its heterogeneous transmission strategy, where the linguistic text is transmitted via a traditional digital channel while emotion and timbre are conveyed through the JSCC stream.This confirms that our skeleton-and-flesh strategy successfully breaks the analog bottleneck, ensuring error-free semantic transmission in moderate-to-good channel conditions while maintaining competitive robustness in harsh environments.

4.3.3. Perceptual Quality Evaluation

Finally, we evaluate the subjective perceptual quality of the synthesized speech using the Mean Opinion Score (MOS). Fig. 4 presents the MOS results from human listening tests, comparing the proposed HASCom against the DeepSC-S and SC-HEST baselines.

Consistent with the objective semantic consistency results, the proposed framework significantly outperforms the baselines in terms of speech naturalness and intelligibility.

https://doi.org/10.3390/1010000

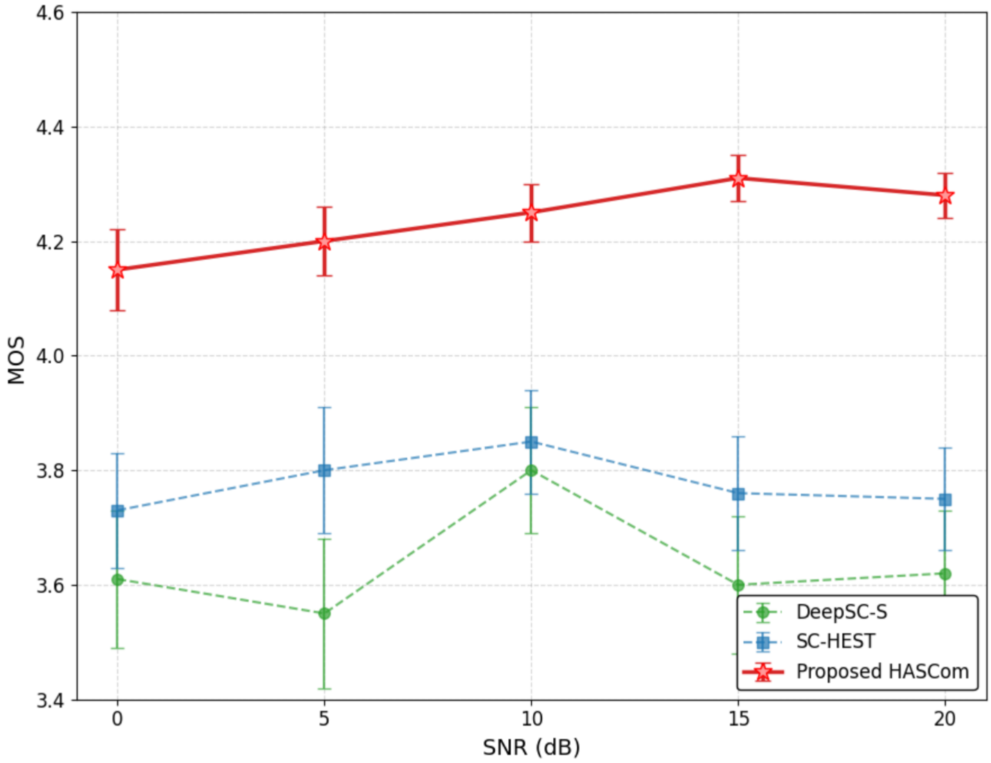

Figure 4. MOS under AWGN channel for different speech to speech transmission approaches.

As observed, the pure analog baselines (DeepSC-S and SC-HEST) fluctuate between a MOS of 3.6 and 3.8. These methods often suffer from residual channel noise, resulting in audible artifacts and robotic prosody, particularly in low SNR conditions.

In sharp contrast, HASCom achieves a remarkable MOS of 4.15 even at 0 dB. This superior performance is primarily driven by our heterogeneous transmission strategy: the digital linguistic stream guarantees accurate content reconstruction , ensuring high intelligibility, while the robust JSCC stream effectively preserves the paralinguistic features. Furthermore, as the SNR improves, the MOS of our method stabilizes around 4.3, approaching ground-truth quality with narrow confidence intervals. This indicates that the conditional diffusion decoder successfully utilizes the robust semantic cues to synthesize high-fidelity waveforms, effectively decoupling the perceptual quality from channel fluctuations.

4.3.4. Analysis of Mel-spectrogram Characteristics

To intuitively verify the generation capability of the proposed HASCom framework, we visualize the intermediate representations and the final synthesized output. Fig. 7 presents the progression of the speech generation process, comparing the Ground Truth , the Encoder Output, the Synthesized Mel-spectrogram, and the Alignment map.

As illustrated in the four mel-spectrograms demonstrate the progression of the diffusion model output at different stages. First, regarding the Linguistic Prior, the output of the text encoder in Fig. 6 exhibits a smoothed and averaged texture. This is expected behavior, as the encoder predicts the general linguistic distribution based on the text input, providing a foundational representation without fine-grained prosodic variations.

Simultaneously, the system demonstrates Robust Alignment as shown in Fig. 8. The clear, unbroken diagonal line in the alignment map indicates that the system maintains robust synchronization between text characters and acoustic frames. This precise phoneme

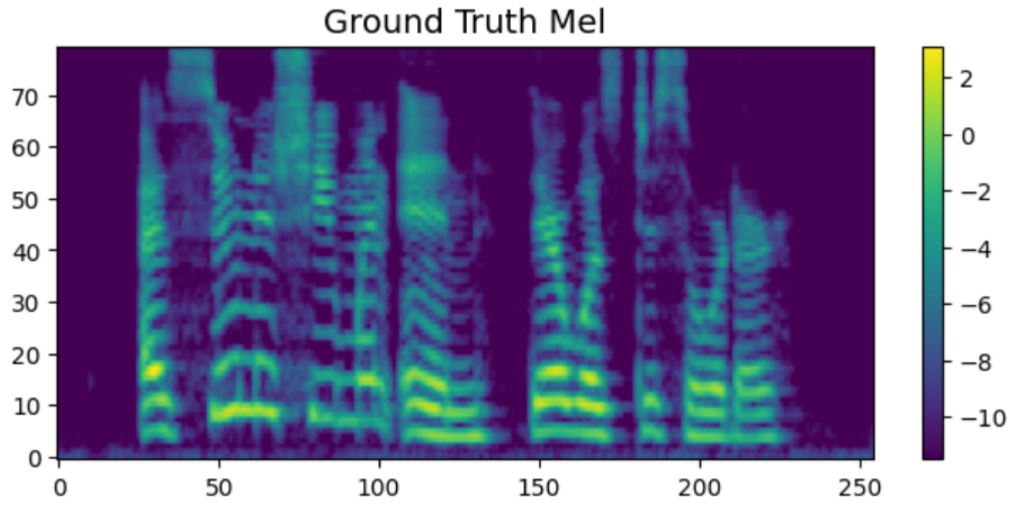

Figure 5. Ground Truth Mel-spectrogram: The original target spectrogram derived from the source speech.

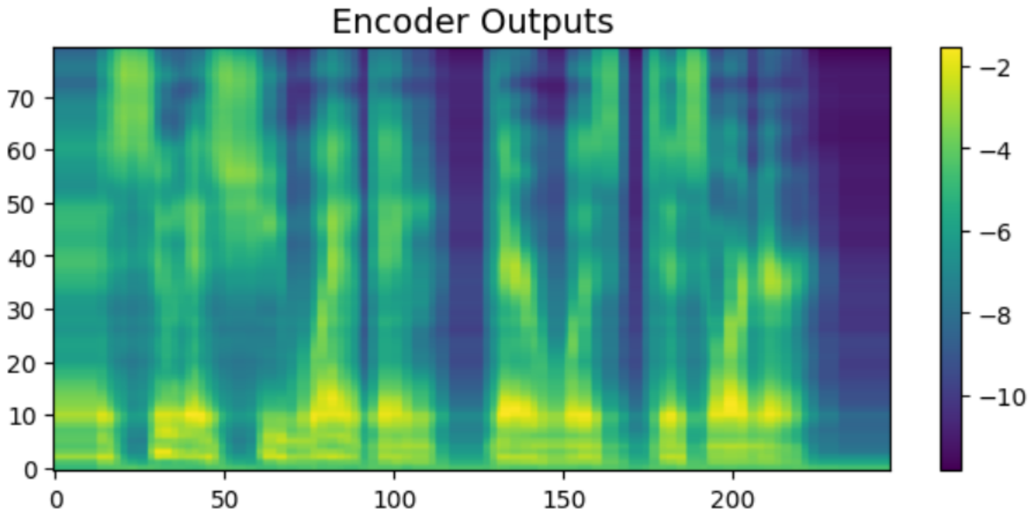

Figure 6. Encoder Outputs (Prior Distribution): The latent representation generated by the text encoder before diffusion.

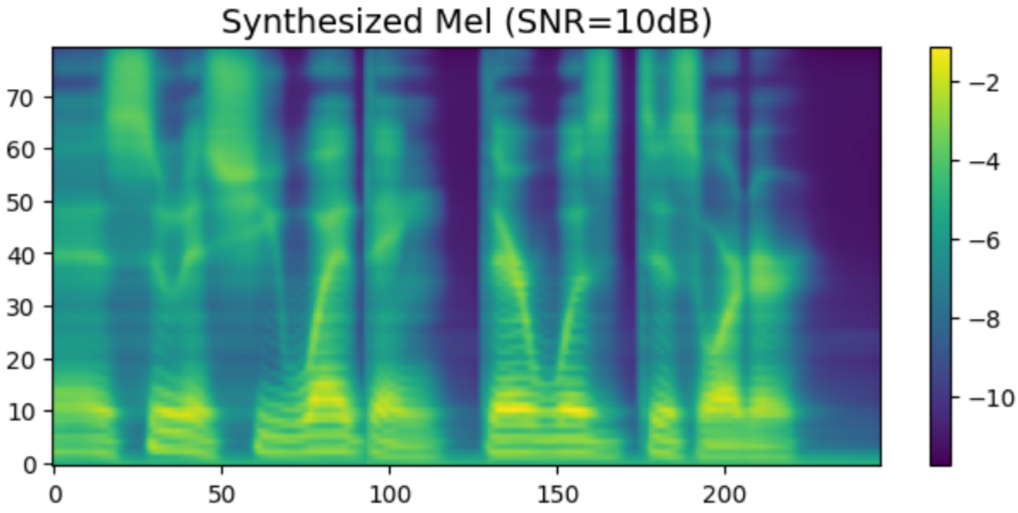

Figure 7. Synthesized Mel-spectrogram (Final Output): The reconstructed spectrogram generated by the diffusion model.

https://doi.org/10.3390/1010000

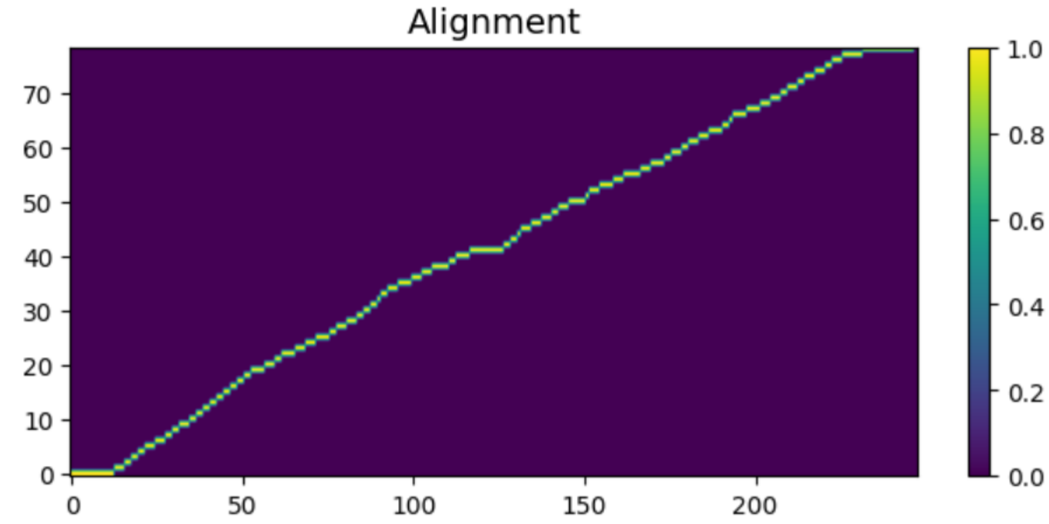

Figure 8. Monotonic Alignment Search: The alignment path between text characters and acoustic frames.

duration modeling ensures that the synthesized speech is coherent and stable, even when transmitted through a noisy channel.

Most importantly, the final Expressive Synthesis is visualized in Fig. 7. By incorporating the transmitted semantic and affective features, the conditional diffusion decoder significantly enriches the spectral representation. In contrast to the smoothed encoder output, the synthesized spectrogram recovers rich high-frequency details, distinct harmonic structures, and subtle pitch nuances. The result closely resembles the Ground Truth in Fig. 5, confirming that the proposed model is capable of generating highly expressive and natural-sounding speech with complex acoustic characteristics.

5. Conclusions

In this paper, we propose the Heterogeneous Affective Speech Semantic Communication System (HASCom), a novel framework that effectively reconciles the trade-off between semantic precision and affective fidelity. Unlike monolithic transmission schemes, HASCom strategically bifurcates the speech signal into a discrete semantic stream and a continuous affective stream, matching the transmission protocol to the physical nature of the features. Specifically, we construct a hybrid digital-analog transmission architecture to ensure rigorous linguistic alignment while preserving infinite-resolution emotional cues. Furthermore, we design a prior-guided conditional diffusion model at the receiver, which synergistically reconstructs high-fidelity speech by painting affective details onto a semantic structural skeleton. Experimental results demonstrate that the proposed method achieves superior performance in both objective and subjective evaluations compared to existing baselines. In terms of semantic preservation, the system achieves a semantic similarity score approaching 100% under stable channel conditions, significantly outperforming DeepSC-S. Regarding perceptual quality, HASCom attains a Mean Opinion Score (MOS) of approximately 4.3, verifying its ability to generate natural and emotionally expressive speech even in noisy environments. In future work, we aim to extend our heterogeneous framework to support real-time multi-speaker scenarios and investigate lightweight adaptation strategies for deployment on resource-constrained edge devices.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis,

X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Where GenAI has been used for purposes such as generating text, data, or graphics, or for study design, data collection, analysis, or interpretation of data, please add "During the preparation of this manuscript/study, the author(s) used [tool name, version information] for the purposes of [description of use]. The authors have reviewed and edited the output and take full responsibility for the content of this publication."

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflicts of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

https://doi.org/10.3390/1010000

References

1. Zhao, C.; Du, H.; Niyato, D.; Kang, J.; Xiong, Z.; Kim, D.I.; Shen, X.; Letaief, K.B. Generative AI for secure physical layer communications: A survey. *IEEE Trans. Cogn. Commun. Netw.* **2025**, *11*, 3–26.
2. Miao, J.; Wang, Z.; Wang, M.; Garg, S.; Hossain, M.S.; Rodrigues, J.J. Secure and efficient communication approaches for Industry 5.0 in edge computing. *Comput. Netw.* **2024**, *242*, 110244.
3. Dong, P.; Ge, J.; Wang, X.; Guo, S. Collaborative edge computing for social internet of things: Applications, solutions, and challenges. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 291–301.
4. He, X.; Jiang, Y.; Liu, Y.; Cui, H.; Pan, H.; Mao, Y. Transforming 6G mobile edge intelligence with large models. *IEEE Netw.* **2025**, Early Access, 1–1.
5. Niyato, D.; Kim, D.I.; Xiong, Z.; et al. Generative AI for semantic communication: Architecture, challenges, and outlook. *IEEE Wirel. Commun.* **2025**, *32*, 132–140.
6. Liu, Y.; Du, H.; Niyato, D.; Kang, J.; Xiong, Z.; Mao, S.; Zhang, P.; Shen, X. Cross-modal generative semantic communications for mobile AIGC: Joint semantic encoding and prompt engineering. *IEEE Trans. Mob. Comput.* **2024**, *23*, 14871–14888.
7. Qin, Z.; Tao, X.; Lu, J.; Tong, W.; Li, G.Y. Semantic communications: Principles and challenges. *IEEE Netw.* **2021**, *35*, 70–76.
8. Zhang, Q.; Saad, W.; Bennis, M.; Debbah, M. Generative AI for Physical Layer Communications: A Survey. *IEEE Trans. Cogn. Commun. Netw.* **2024**, Early Access.
9. Xie, H.; Ye, Z.; Li, G.Y.; Juang, B.H.F. Deep learning enabled semantic communication systems. *IEEE Trans. Signal Process.* **2021**, *69*, 2663–2675.
10. Al-Hawawreh, M.; Hossain, M.S. A human-centered quantum machine learning framework for attack detection in IoT-based healthcare Industry 5.0. *IEEE Internet Things J.* **2025**, Early Access.
11. Yeo, Y.; Kim, J.; Song, H.-Y. Enhanced semantic communication schemes for speech signals. *Electron. Lett.* **2024**, *60*, e13183.
12. Weng, Z.; Qin, Z. Semantic communication systems for speech transmission. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2434–2444.
13. Belin, P.; Zatorre, R.J.; Lafaye, P.; Lioret, P.; Pike, B.; Penhune, V. Voice-selective areas in human auditory cortex. *Nature* **2000**, *403*, 309–312.
14. Han, T.; Yang, Q.; Shi, Z.; He, S.; Zhang, Z. Semantic-preserved communication system for highly efficient speech transmission. *IEEE J. Sel. Areas Commun.* **2022**, *41*, 245–259.
15. Kumar, Y.; Koul, A.; Singh, C. A deep learning approaches in text-to-speech system: A systematic review and recent research perspective. *Multimed. Tools Appl.* **2023**, *82*, 15171–15197.
16. Obert, A.; Gunter, T.C.; Kotz, S.A. Neural integration of affective prosodic and semantic cues in non-literal forms of speech understanding. *bioRxiv* **2026**, Preprint.
17. Weng, Z.; Qin, Z.; Tao, X.; Pan, C.; Liu, G.; Li, G.Y. Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 6227–6240.
18. Xiao, Z.; Yao, S.; Dai, J.; Wang, S.; Niu, K.; Zhang, P. Wireless deep speech semantic transmission. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023; pp. 1–5.
19. Xie, H.; Ye, Z.; Li, G.Y.; Juang, B.H.F. A deep learning enabled semantic communication system for speech-to-text transmission. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021; pp. 6678–6682.
20. Tan, K.; Zhao, H.; Zhang, Y.; Cao, K.; Luo, P.; Zhang, Y.; Wei, J. EESC-S: An emotion-enhanced semantic communication framework for speech transmission. *IEEE Trans. Cogn. Commun. Netw.* **2025**, Early Access.
21. Chen, M.; Wang, C.; He, X.; Zhu, F.; Wang, L.; Vasilakos, A.V. Embodied artificial intelligence-enabled internet of vehicles: Challenges and solutions. *IEEE Veh. Technol. Mag.* **2025**, Early Access, 2–9.
22. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS 2020)*, 2020; pp. 6840–6851.
23. Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; Kudinov, M. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021; pp. 8599–8608.
24. Grassucci, E.; Marinoni, C.; Rodriguez, A.; Comminiello, D. Diffusion models for audio semantic communication. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024; pp. 13136–13140.
25. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NIPS 2020*, 2020; pp. 12449–12460.
26. Chen, S.; Wang, C.; Chen, Z.; et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518.
27. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Emotional voice conversion: Theory, databases and ESD. *Speech Commun.* **2022**, *137*, 1–18.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/1010000