sensors

MDPI

Article

Rate–Distortion–Perception Optimized Neural Speech Transmission System for High-Fidelity Semantic Communications †

Shengshi Yao [1], Zixuan Xiao [1] and Kai Niu [1,2,*]

[1] Key Laboratory of Universal Wireless Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen 518066, China
* Correspondence: niukai@bupt.edu.cn
† This paper is an extended version of our paper published in ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.

Abstract: We consider the problem of learned speech transmission. Existing methods have exploited joint source–channel coding (JSCC) to encode speech directly to transmitted symbols to improve the robustness over noisy channels. However, the fundamental limit of these methods is the failure of identification of content diversity across speech frames, leading to inefficient transmission. In this paper, we propose a novel neural speech transmission framework named *NST*. It can be optimized for superior rate–distortion–perception (RDP) performance toward the goal of high-fidelity semantic communication. Particularly, a learned entropy model assesses latent speech features to quantify the semantic content complexity, which facilitates the adaptive transmission rate allocation. NST enables a seamless integration of the source content with channel state information through variable-length joint source–channel coding, which maximizes the coding gain. Furthermore, we present a streaming variant of NST, which adopts causal coding based on sliding windows. Experimental results verify that NST outperforms existing speech transmission methods including separation-based and JSCC solutions in terms of RDP performance. Streaming NST achieves low-latency transmission with a slight quality degradation, which is tailored for real-time speech communication.

Keywords: speech transmission; joint source–channel coding; semantic communications

Citation: Yao, S.; Xiao, Z.; Niu, K. Rate–Distortion–Perception Optimized Neural Speech Transmission System for High-Fidelity Semantic Communications. *Sensors* **2024**, *24*, 3169. https://doi.org/10.3390/s24103169

Academic Editor: Pablo Angueira

Received: 19 April 2024
Revised: 14 May 2024
Accepted: 15 May 2024
Published: 16 May 2024

1. Introduction

The vast demand of streaming audio and video communication poses significant challenges to wireless communication systems, underscoring the need to elevate both the quality and efficiency of speech transmission. Current wireless communication systems suffers from the cliff effect where the signal reconstruction quality breaks down if the channel quality falls below the level anticipated by the channel code. Learning-based speech transmission methods [1–3] are emerging as promising solutions to improve the end-to-end transmission performance in the context of semantic communication [4–7]. They mostly leverage the idea of joint source–channel coding (JSCC) to produce transmitted symbols directly from raw speech signals with neural networks, which is featured with graceful degradation with respect to channel quality [1,2,8]. However, these approaches fail to identify the content diversity among signals, leading to inefficient transmission. Streaming inference is also a fundamental aspect in real-time communication (RTC) scenarios. Although transmission errors can be compensated by retransmission such as hybrid automatic repeat requests, these lead to a loss of efficiency and transmission delay.

To address the above-mentioned issues, we make the first attempt to design a high-fidelity neural speech transmission framework (*NST*) for better end-to-end transmission performance. Motivated by learned data compression techniques [9,10], NST establishes a learned entropy model on latent speech features and then realizes semantic-guided

variable-length joint source–channel coding, thus achieving better coding gain. Specifically, a critical set of hyperprior variables is established upon the latent features, which estimate the entropy of speech features by variational modeling. Under the guidance, speech latent features are dynamically encoded to variable-length symbol sequences via a joint source–channel encoder. Based on our previous work [11], we further investigate the real-time speech transmission within latency-sensitive contexts, such as online conferencing and voice calls. In particular, we develop a streaming variant of NST tailored for low-latency transmission. All the operators of the model are strictly causal ones, which attend to the past speech signals only, to satisfy the real-time property. In addition, we design a sliding-window based inference mechanism in joint source–channel coding, which balances the performance of speech reconstruction and the overall delay.

We evaluate the performance by conducting simulations over wireless channels. The results demonstrate that the proposed NST model is source and channel-adaptive. In comparison to advanced speech coding combined with error correction coding, and the existing JSCC solution, the proposed NST achieves a superior rate–distortion–perception tradeoff. This translates to a high-fidelity speech reconstruction performance while incurring lower bandwidth costs. Notably, the streaming NST makes a slight compromise in speech quality to meet the low-latency requirement.

Notational Conventions: Throughout this paper, bold letters (e.g., $\boldsymbol{x}$) denote vectors and the scalars, and lowercase ones denote scales. Bold uppercase letters (e.g., $\mathcal{V}$) represent a collection. $\log(\cdot)$ is the logarithm to base 2. p_x denotes a probability density function (pdf) with respect to the continuous-valued random variable x. $\mathcal{U}(a-m, a+m)$ denotes a uniform distribution centered on a with width $2m$. $\mathbb{R}$ and $\mathbb{C}$ denote the real number set and the complex number set, respectively. $\mathbb{E}[\cdot]$ denotes the statistical expectation operation.

2. Methodology

2.1. Architecture

The NST system architecture is illustrated in Figure 1. Assuming a sequence of T speech frames $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_T\}$, the analysis transform module $g_a\left(\cdot; \boldsymbol{\phi}_g\right)$, which consists of convolutional neural networks (CNNs) with temporal downsampling, transforms them into a semantic latent feature sequence $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots \boldsymbol{y}_T\}$. Then, the latent features $\boldsymbol{y}$ are fed into both a hyperprior encoder $h_a(\cdot; \boldsymbol{\phi}_h)$ and a variable-length JSCC encoder $f_e\left(\cdot; \boldsymbol{\phi}_f\right)$. On one hand, in order to conveniently quantify the amount of information for speech features, each element of $\boldsymbol{y}$ is variationally modeled by a simple Gaussian, whose parameters are encapsulated by the hyperprior variable $\boldsymbol{z}$. The means and variances of the Gaussians are encoded by $h_a(\cdot; \boldsymbol{\phi}_h)$ and $h_s(\cdot; \boldsymbol{\theta}_h)$ to capture the dependencies of $\boldsymbol{y}$. On the other hand, $f_e\left(\cdot; \boldsymbol{\phi}_f\right)$ encodes $\boldsymbol{y}$ into channel-input sequence $\boldsymbol{s} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots \boldsymbol{s}_T\}$, where $\boldsymbol{s}_i \in \mathbb{C}^{k_i}$ is a k_i-dimensional complex vector to transmit y_i. We consider a wireless channel denoted by $W(\cdot; \boldsymbol{\nu})$, where $\boldsymbol{\nu}$ denotes the channel parameters. Thus, the receiver obtains the sequence $\hat{\boldsymbol{s}} = W(\boldsymbol{s}; \boldsymbol{\nu})$ with the transition probability $p_{\hat{\boldsymbol{s}}|\boldsymbol{s}}(\hat{\boldsymbol{s}}|\boldsymbol{s})$. As illustrated in Fig. 1, with a mirrored design, the JSCC decoder $f_d\left(\cdot; \boldsymbol{\theta}_f\right)$ reconstructs latent representation $\hat{\boldsymbol{y}}$, and semantic synthesis transform $g_s\left(\cdot; \boldsymbol{\theta}_g\right)$ recovers speech waveform $\hat{\boldsymbol{x}}$. Hence, the total link of *NST* is formulated by

$$\boldsymbol{x} \xrightarrow{g_a\left(\cdot;\boldsymbol{\phi}_g\right)} \boldsymbol{y} \xrightarrow{f_e\left(\cdot;\boldsymbol{\phi}_f\right)} \boldsymbol{s} \xrightarrow{W(\cdot;\boldsymbol{\nu})} \hat{\boldsymbol{s}} \xrightarrow{f_d\left(\cdot;\boldsymbol{\theta}_f\right)} \hat{\boldsymbol{y}} \xrightarrow{g_s\left(\cdot;\boldsymbol{\theta}_g\right)} \hat{\boldsymbol{x}}, \tag{1}$$

with the latent prior $\boldsymbol{y} \xrightarrow{h_a(\cdot;\boldsymbol{\phi}_h)} \boldsymbol{z} \xrightarrow{h_s(\cdot\boldsymbol{\theta}_h)} \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$ and $(\boldsymbol{\theta}, \boldsymbol{\phi}) = (\boldsymbol{\phi}_g, \boldsymbol{\phi}_h, \boldsymbol{\phi}_f, \boldsymbol{\theta}_g, \boldsymbol{\theta}_h, \boldsymbol{\theta}_f)$ encapsulating the learnable parameters of each function above. Moreover, the hyperprior $\boldsymbol{z}$ can be viewed as side information, which is optionally sent via a digital link to the receiver to refine the latent feature $\boldsymbol{y}$.

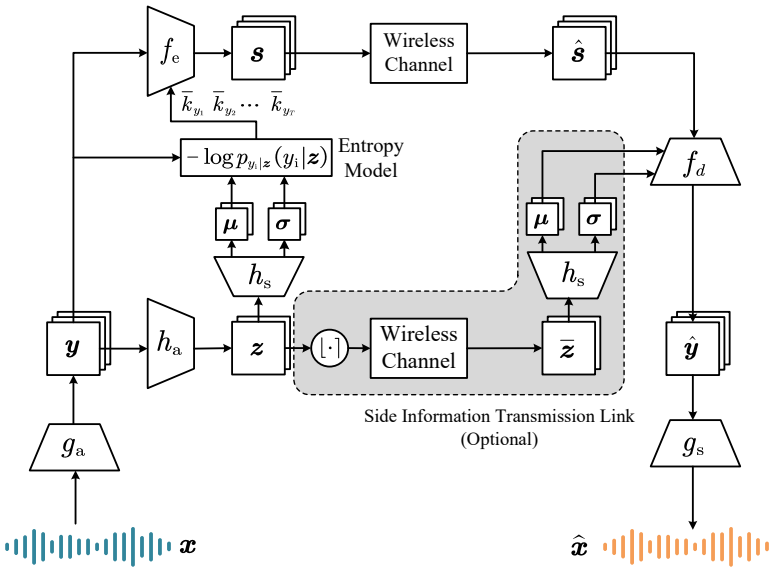


Figure 1. The architecture of the Neural Speech Transmission system (NST).

2.2. Dynamic Variable-Length Joint Source–Channel Coding

As defined previously, each y_i is variationally modeled as a Gaussian with mean μ_i and variance σ_i^2, whose density function is factorized as

$$p(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{\theta}_h,\boldsymbol{\psi}_h)=\prod_i \underbrace{\left(\mathcal{N}\left(\mu_i,\sigma_i^2\right)*\mathcal{U}\left(-\frac{1}{2},+\frac{1}{2}\right)\right)}_{p_{y_i|z}}(y_i), \tag{2}$$

with $(\boldsymbol{\mu},\boldsymbol{\sigma})=h_s(\boldsymbol{z})$, where $*$ is a convolutional operation. Dithered quantization is adopted [12], such that we can derive a non-negative entropy estimation of $-\log p_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z})$ by directly using the proxy $\tilde{y}_i=y_i+o, o\in\mathcal{U}\left(-\frac{1}{2},+\frac{1}{2}\right)$. The estimated entropy is directly linked to the channel bandwidth cost in the JSCC encoder for transmission. Intuitively, if $\boldsymbol{y}_i$ is tagged with high entropy, it will be allocated with more bandwidth and vice versa.

In practice, the total bandwidth cost $K_{\boldsymbol{y}}$ for transmitting $\boldsymbol{y}$ is formulated by

$$K_{\boldsymbol{y}}=\sum_{i=1}^{T}\bar{k}_{y_i}=\sum_{i=1}^{T}Q(k_{y_i})=\sum_{i=1}^{T}Q(-\eta_y\log p_{\boldsymbol{y}_i|\boldsymbol{z}}(\boldsymbol{y}_i|\boldsymbol{z})), \tag{3}$$

where η_y controls the scaling between the estimated entropy and the number of transmitted symbols, and Q denotes a 2^n-level scalar quantization with the quantized value set as $\mathcal{V}=\{v_1,v_2,\cdots,v_{2^n}\}$. Hence, n bits are transmitted as side information to inform the receiver in which $\bar{k}_{y_i}\in\mathcal{V}$ is selected for transmitting $\boldsymbol{y}_i$.

We adopt a pair consisting of a Transformer-like [13] JSCC encoder and decoder as f_e and f_d, as plotted in Figure 2. Guided by the entropy model $-\log p_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z})$, a set of learnable rate token embeddings with the same dimension with y_i are developed, each of which corresponds to a value in $\mathcal{V}$. To adapt to various channel environments, we assume a channel state information feedback to inform the sender of the instant signal-to-noise ratio (SNR). Similarly, a set of learnable SNR tokens are developed. T frames of speech features are gathered and fused with respective rate tokens and an SNR token, and they are finally fed into the Transformer block with N_e Transformer layers. A bunch of fully connected (FC) layers with output dimensions of $v_q, q=1,2,\cdots,2^n$ are employed to map the embeddings into s_i with given dimensions. A toy visualization of the rate allocation result is displayed in Figure 3. It can be observed that more bandwidth is allocated to frames with prominent

contents and less is allocated to ones in silence. The overall bandwidth is adjusted by tuning the hyperparameter η_y.

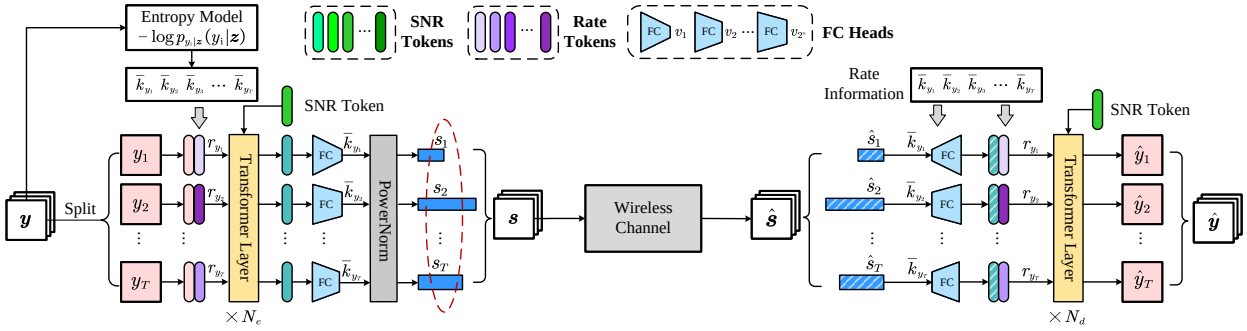
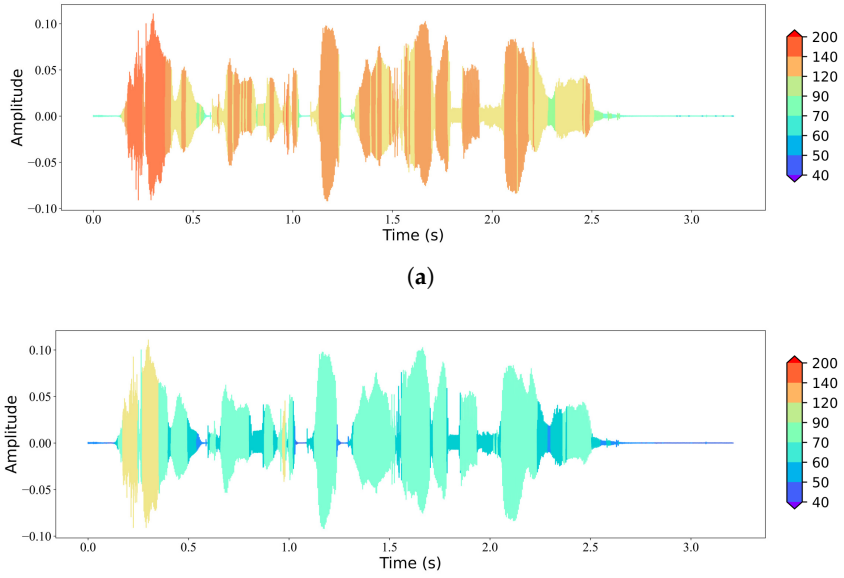


Figure 2. The pipeline of the variable-length joint source-channel coding (JSCC) via f_e and f_d. FC denotes fully connected layers.



(**a**)

(**b**)

Figure 3. Visualization of rate allocation along temporal domain. (**a**) Bandwidth K = 10 kHz with η_y = 0.16. (**b**) Bandwidth K = 4 kHz with η_y = 0.105.

2.3. Streaming NST for Real-Time Communication

In this subsection, we propose a streaming NST model variant to facilitate real-time speech communication.

Firstly, all the convolutional operators in g_a and g_s are substituted by causal ones. The transposed convolution operations in g_s only pad on the past steps to meet the causal property. Secondly, the joint source–channel coding of speech latent features is modified from that in Figure 3 to reduce the latency. Traditionally, Transformer uses multi-head attention that jointly learns diverse relationships between queries and keys, which are the speech features $\boldsymbol{y}$ in this paper, from different representation subspaces with j-th head computing

$$\boldsymbol{Q}_j = \boldsymbol{W}_j^Q \boldsymbol{y}, \boldsymbol{K}_j = \boldsymbol{W}_j^K \boldsymbol{y}, \boldsymbol{V}_j = \boldsymbol{W}_j^V \boldsymbol{y}. \tag{4}$$

To meet the real-time requirement, a causal masked attention method is proposed together with a sliding-window inference mechanism. As shown in Figure 4, the JSCC encoder f_e has a limited contextual window with W frames, which hops along the temporal domain with a stride of N frames. In particular, we follow Transformer-XL [14] and create

a segment-level recurrence of output by intermediate layers. In this paper, we define the attention span of each layer as $2N$ frames, which ends with the last frame of the current N target frames. The self-attention is computed after a causal mask $\boldsymbol{M}$ is applied, whose elements satisfy

$$\boldsymbol{M}_{t,\tau} = \begin{cases} 1, & t - 2N < \tau \leq t \\ -\infty, & \text{others} \end{cases}. \tag{5}$$

Then, the output of the j-th head self-attention $\boldsymbol{a}_j$ is formulated by

$$\boldsymbol{a}_j = \text{Softmax}\left(\frac{\boldsymbol{Q}_j \boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d_h}} \odot \boldsymbol{M}\right)\boldsymbol{V}_j, \tag{6}$$

where d_h is the dimension of each head. Thus, the length of contextual window W grows linearly with respect to the number of Transformer layers as well as the window stride, which can be written as $W = N(N_e + 1)$ frames.

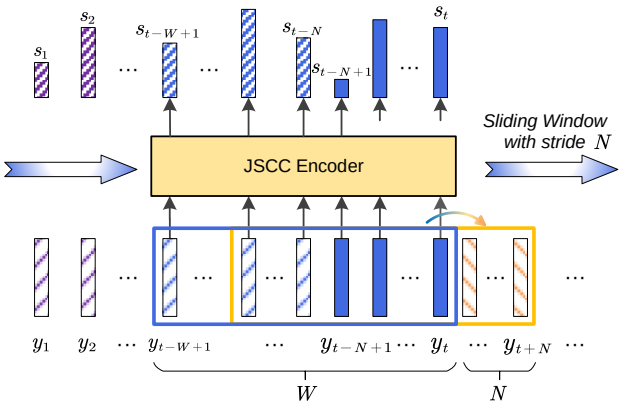


Figure 4. Streaming joint source–channel encoding for real-time inference. It encodes latent features of N frames into transmitted symbols in each inference, e.g., $\boldsymbol{y}_{t-N+1}, \cdots, \boldsymbol{y}_t$ for the blue window in the figure and then the ones in orange in the next inference.

2.4. Optimization Goal

The analysis transform together with the joint source–channel encoder creates a parametric density $q_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}}$ to approximate the true posterior distribution $p_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}}$. The optimization goal is to minimize the Kullback–Leibler (KL) divergence between the above two components. After reformulation, it minimizes its upper bound, i.e.,

$$\min \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \mathop{\mathbb{E}}_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}\sim q_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}}} D_{\mathrm{KL}}\left[q_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}} \| p_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}}\right] \leq \min \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \mathop{\mathbb{E}}_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}\sim q_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}}} \Big[\underbrace{-\log p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})}_{\text{side info. coding rate}} \underbrace{-\log p_{\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}}}(\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}})}_{\text{bandwidth}} \underbrace{-\mathbb{E}_{\boldsymbol{y}\sim p_{\boldsymbol{y}|\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}}} \log p_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})}_{\text{distortion}} \Big] + \text{const.} \tag{7}$$

The first term of (7) represents the cost of encoding the side information assuming $p_{\tilde{\boldsymbol{z}}}$ as the entropy model, where $\tilde{z}_i = z_i + o$ is the proxy quantization of z_i. Since there is no prior information about $\boldsymbol{z}$, $p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})$ is modeled as a non-parametric fully factorized density [9] $p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) = \prod_i \left(p_{z_i|\boldsymbol{\psi}^{(i)}}(z_i|\boldsymbol{\psi}^{(i)}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2})\right)(z_i)$. The second term represents the bandwidth cost of encoding $\hat{\boldsymbol{s}}$. In practice, the intermediate variable $\boldsymbol{y}$ is utilized by $p_{\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}}} = W(p_{\boldsymbol{s}|\tilde{\boldsymbol{z}}}|\boldsymbol{h}) = W(f_e(p_{\boldsymbol{y}|\tilde{\boldsymbol{z}}})|\boldsymbol{h})$. The third term denotes the weighted distortion of the reconstructed speech waveform. $d(\cdot,\cdot)$ indicates the objective signal distortion. To enrich the distortion term in alignment with human perceptual quality, a differentiable $F(\cdot)$ is employed as a perceptual feature extractor, and the distance between perceptual features $d_p(\cdot,\cdot)$ is minimized to improve the listening quality.

In summary, the RDP function is formulated as

$$L_{\mathrm{RDP}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[-\eta_y \log p_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z}) - \eta_z \log p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) + \lambda_D d(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_P d_p(F_{\boldsymbol{x}}, F_{\hat{\boldsymbol{x}}})], \tag{8}$$

where the Lagrange multipliers λ_D, λ_P control the tradeoff among the total transmission rate, the distortion and the perceptual quality. The scaling factor η_y is adjusted for RDP tradeoff, while η_z is determined according to the channel capacity of the optional tranmission link.

3. Results

In this section, we provide numerical results in terms of objective quality metrics and subjective scores to evaluate the quality of speech transmission.

3.1. Experimental Setup

The mono speech signals are sampled at 16 kHz from the TIMIT dataset [15]. Compared to our conference paper [11], to adapt to RTC scenarios, a shorter frame length is considered in this paper. Each speech frame has $L = 128$ samples with an overlap of eight samples. The analysis transform module g_a and synthesis transform module g_s consist of stacks of 1D convolutional layers with a residual connection. The number of channels of the convolutional kernel of the output/input layer for g_a/g_s is configured with $C_g = 4$, while the one for h_a/h_s is set as $C_h = 2$. In the variable-length JSCC coder f_e and f_d, we use $N_e = 3$ Transformer layers with eight-head self-attention. The quantized channel bandwidth cost value set is defined as $\mathcal{V} = \{10, 40, 90, 120, 200, 250, 300, 400\}$. Each speech frame $\boldsymbol{x}_i \in \mathbb{R}^{1\times L}$ is transformed into latent feature $\boldsymbol{y}_i \in \mathbb{R}^{C_g \times \frac{L}{4}}$ with a downsampling factor of four. It is then flattened into an embedding vector with a dimension of $\frac{C_g L}{4} = 128$, which is identical to the dimension of the Transformer in JSCC coders.

In (8), the object signal distortion d is evaluated by the mean square error in the time domain. In terms of perceptual optimization, we minimize the difference of Mel-frequency cepstral coefficients (MFCCs) [16], which is a hand-crafted speech perceptual feature. Specifically, a mean square loss function d_p for MFCCs is employed, where F denotes the function of the MFCC extractor.

We compare our NST model with traditional separation-based transmission schemes. Specifically, we employ the widely used speech codec AMR-WB [17] and Opus [18] for source coding and convolutional codes, 5G LDPC [19] for channel coding, and follow the principle of adaptive modulation coding (AMC) [20]. Moreover, we also compare our NST model with another JSCC model DeepSC-S [1] for speech transmission, which is a non-streaming model with CNN modules. We modify its model to support low bandwidth transmission with 12 kHz and 32 kHz, separately.

3.2. Evaluation Metrics

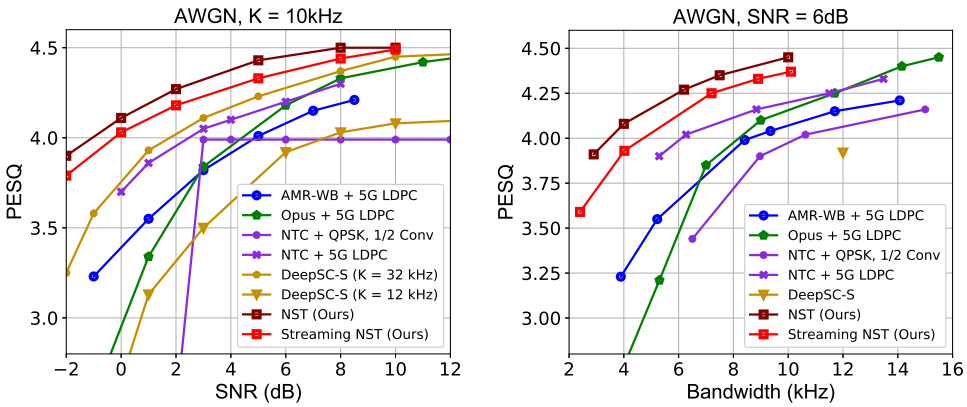
In terms of objective metrics for perceptual quality, we report the perceptual evaluation of speech quality (PESQ) [21] scores, which range from 1.0 to 4.5. Furthermore, we implement a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) subjective test [22] for human preference evaluation. As a widely used approach in the subject quality assessment method, the MUSHRA test allows users to compare multiple variants of reconstructed audio and provides the relative score between 0 and 100. We randomly select 10 speech segments from the test set.

3.3. Results Analysis

Figure 5 reports the PESQ performance over additive white Gaussian noise (AWGN) channels. In Figure 5a, with a fixed channel bandwidth cost of $K = 10$ kHz, we find that the proposed NST brings a performance gain for all SNRs by incorporating source and channel information into JSCC, especially in a low SNR region. In addition to traditional speech source coding methods, we also compare with a nonlinear neural speech compressor which employs the similar entropy model as NST to entropy encode the latent speech features. This scheme is marked in the figure as "NTC + QPSK, 1/2 Conv" when using convolutional codes with a rate of 1/2 and QPSK modulation as "NTC + 5G LDPC" when using LDPC codes. NST demonstrates graceful performance degradation with the decrease of SNR,

while the performance of other separation-based methods probably breaks down (cliff effect) when using a single-channel coding rate and modulation level, e.g., "NTC + QPSK, 1/2 Conv". For the 5G LDPC, we plot the envelope of several curves, corresponding to different coding rates and modulation levels. Compared with another JSCC method DeepSC-S, our model achieves better perceptual quality by introducing an explicit perceptual loss function with much less bandwidth cost. In addition, we notice a slight quality drop under the streaming inference setting, but it remains better than other methods. NST adapts well to various channel conditions by means of SNR token fusion in f_e and f_d using a single model.

Figure 5b compares the rate–distortion–perception performance using different methods for the 6 dB AWGN channel. Since the NST model learns an adaptive rate allocation mechanism, we traverse the η_y from 0.1 to 0.3 and finetune the model with a fixed λ_D and λ_P. It can be observed that a remarkable bandwidth saving can be accomplished for NST by integrating source semantic information as well as channel information.



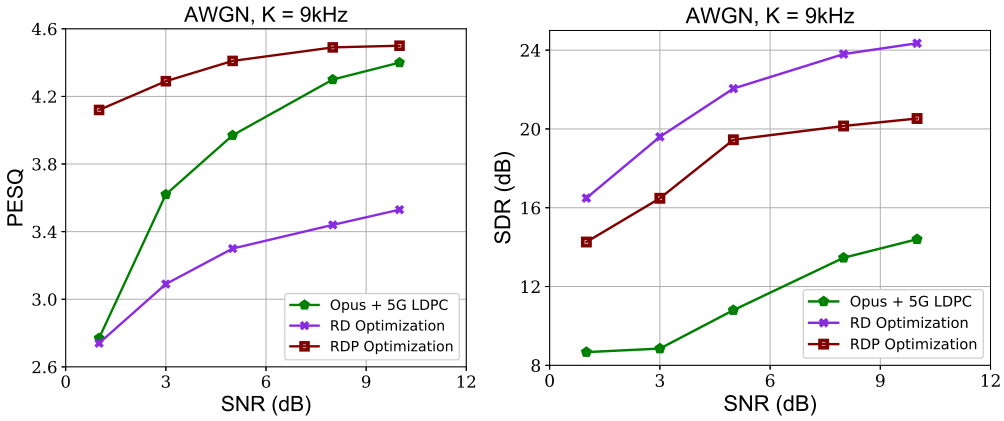
(**a**) (**b**)

Figure 5. Perceptual evaluation of speech quality (PESQ) performance over additive white Gaussian noise (AWGN) channel. (**a**) PESQ scores versus signal-to-noise ratio (SNR). The bandwidth of all methods K is 10 kHz, except those of DeepSC-S are 12 kHz and 32 kHz (yellow lines). (**b**) PESQ scores versus channel bandwidth cost when SNR = 6 dB.

To delineate the distortion–perception tradeoff, we conduct an ablation study examining the impact of perceptual optimization. We evaluate the signal–to–distortion ratio (SDR) performances to assess the traditional signal distortion. The results in Figure 6 demonstrate that the proposed NST using an RDP optimization objective function (8) outperforms its counterpart solely optimized toward reduced signal distortion in terms of perceptual quality. NST with rate–distortion (RD) optimization (omitting perceptual loss in (8)) exhibits inferior perceptual quality despite there being less objective signal distortion. Performance using the traditional speech coding method is also included in the figure, which also underscores the significance of perceptual optimization in addition to minimizing the objective signal distortion.

Figure 7 displays the effect of SNR fusion in our SNR-adaptive joint source–channel coding. It can be observed that the PESQ-SNR curve of the proposed NST trained under multiple SNRs with SNR token fusion closely approximates the envelope of the curves obtained from models trained using single SNR values.

We additionally carry out experiments on the widely used COST2100 fading channel [23] to verify the robustness of the NST model. Figure 8 shows the results. With a feedback of average SNR and the SNR token fusion, our model adapts to the channel states well, while performances of DeepSC-S are evaluated on models trained at multiple SNRs. With lower bandwidth in Figure 8, NST also shows better transmission efficiency compared to traditional methods.



(**a**) (**b**)

Figure 6. Distortion–perception tradeoff using different optimization objectives with 9 kHz channel bandwidth cost over AWGN channel. (**a**) PESQ for assessing perceptual quality. (**b**) Signal–to–distortion ratio (SDR) for assessing signal distortion.

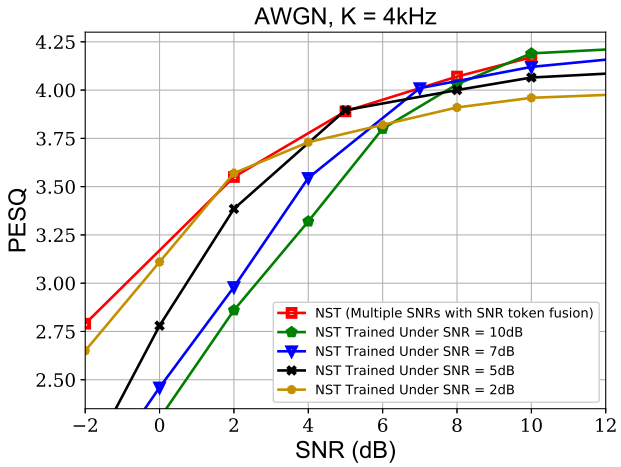
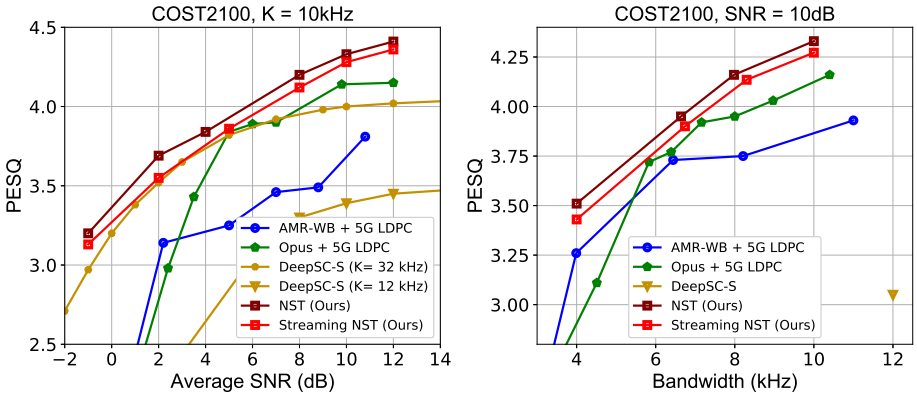


Figure 7. Effect of SNR token fusion in joint source–channel coding.



(**a**) (**b**)

Figure 8. PESQ performance over COST2100 fading channel. (**a**) PESQ scores versus average SNR. (**b**) PESQ scores versus channel bandwidth cost.

The subjective user rating results in Figure 9 verify that the proposed NST recovers perceptually satisfying speech over 6 dB AWGN channels, even consuming much less bandwidth than separation-based speech coding methods. Compared to DeepSC-S,

which is only optimized for lower distortion, the RDP-optimized NST achieves semantics-guided dynamic rate allocation, thus much improving the end-to-end system gain. The perceptual quality of streaming NST exhibits no substantial degradation compared to the non-streaming one, which is of practical value in RTC scenarios.

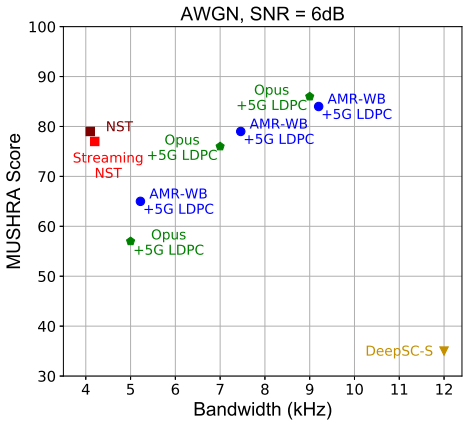


Figure 9. MUSHRA scores evluated under 6 dB AWGN channel. Audio samples are available at https://ximoo123.github.io/NSTSpeech (accessed on 1 March 2024).

3.4. Discussion on the Quality–Latency Tradeoff

In terms of streaming NST, we investigate the tradeoff between the perceptual quality and the transmission latency. As is defined previously, each frame of speech feature $\boldsymbol{y}_i$ accounts for 8 milliseconds (ms) of a 16 kHz signal.

Table 1 shows the tradeoff between speech quality and transmission delay, which consists of encoding and decoding time (runtime) and the latency. In the context of sliding-window-based inference, a longer stride will increase the latency as it needs to wait for the arrival of future frames to collect all features belonging to the same window.

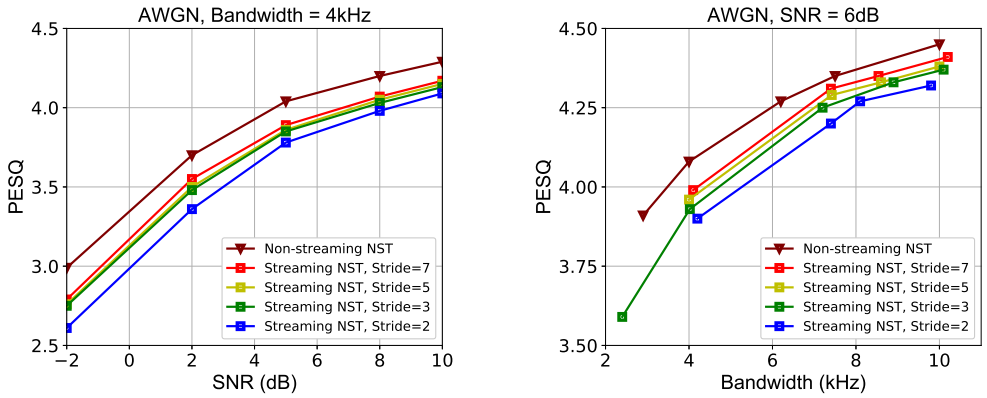
Table 1. Quality–delay tradeoff for the streaming NST model tested with SNR = 10 dB over the AWGN channel.

Stride	PESQ	Total Delay	Runtime	Maximum Latency
2	4.09	67.1 ms	51.1 ms	16 ms
3	4.13	83.2 ms	59.2 ms	24 ms
5	4.15	112.7 ms	72.7 ms	40 ms
7	4.17	140.1 ms	84.1 ms	56 ms

We also compare the PESQ performances versus stride frames across different SNRs and bandwidth cost. The results in Figure 10 verify that a longer window stride as well as the length of contextual windows in JSCC consistently presents a better coding gain across different transmission conditions at the cost of longer delay according to Table 1. Except this subsection, performances of streaming NST are reported with a stride of $N = 3$ and a total delay of less than 100 ms. It satisfies the real-time property and ensures a high-quality speech restoration simultaneously. The runtime is evaluated on an Intel(R) Core i9-12900K CPU (Intel Corporation, Santa Clara, CA, USA). Table 2 presents the model complexity comparison of both computational (measured by giga floating point operations per second, i.e., GFLOPs) and space complexity. Due to the employment of a tiny Transformer in the joint source–channel encoder, our model is comparably lightweight and computational efficient. Extra measures for accelerating inference may be taken to facilitate lightweight deployment in resource-limited devices.

Table 2. Model complexity comparison.

Model	GFLOPs	#Params (Unit: Million)
[3]	>31	>106
DeepSC-S [1]	7.60	0.24
NST (Ours)	9.87	2.49



(**a**) (**b**)

Figure 10. PESQ performances using different strides N over AWGN channels. (**a**) PESQ scores versus SNR. (**b**) PESQ scores versus bandwidth.

4. Conclusions

In this paper, we present the NST, which is a novel neural speech transmission framework. The model features dynamic rate allocation for variable-length JSCC, which is guided by the variational modeling of speech latent features. It presents good adaptability to varying channel conditions by channel information fusing in JSCC. A streaming variant of NST is also designed for RTC. Simulation results verify that the proposed method consumes much less bandwidth cost than classical methods when achieving similar perceptual performances. It highlights NST's potential in high-efficiency and high-fidelity speech transmission in the realm of semantic communication.

Author Contributions: Conceptualization, S.Y.; Methodology, S.Y. and Z.X.; Validation, S.Y. and Z.X.; Formal analysis, S.Y. and Z.X.; Writing—original draft preparation, S.Y. and Z.X.; Writing—review and editing, S.Y.; Visualization, S.Y and Z.X.; Supervision, K.N.; Project administration, K.N; Funding acquisition, S.Y. and K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China under Grant 92267301 and Grant 62071058 and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2023305.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Weng, Z.; Qin, Z. Semantic communication systems for speech transmission. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2434–2444. [CrossRef]
2. 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. [CrossRef]
3. Guo, J.; Zhang, Y.; Liu, C.; Xu, W.; Bie, Z. SNR-Adaptive Multi-Layer Semantic Communication for Speech. In Proceedings of the 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Toronto, ON, Canada, 5–8 September 2023; pp. 1–6. [CrossRef]
4. Qin, Z.; Tao, X.; Lu, J.; Tong, W.; Li, G.Y. Semantic communications: Principles and challenges. *arXiv* **2021**, arXiv:2201.01389.

5. Dai, J.; Zhang, P.; Niu, K.; Wang, S.; Si, Z.; Qin, X. Communication beyond transmitting bits: Semantics-guided source and channel coding. *IEEE Wirel. Commun.* **2023**, *30*, 170–177. [CrossRef]
6. Xu, J.; Tung, T.Y.; Ai, B.; Chen, W.; Sun, Y.; Gündüz, D.D. Deep joint source-channel coding for semantic communications. *IEEE Commun. Mag.* **2023**, *61*, 42–48. [CrossRef]
7. Lu, Z.; Li, R.; Lu, K.; Chen, X.; Hossain, E.; Zhao, Z.; Zhang, H. Semantics-empowered communications: A tutorial-cum-survey. *IEEE Commun. Surv. Tutor.* **2023**, *26*, 41–79. [CrossRef]
8. Bourtsoulatze, E.; Kurka, D.B.; Gündüz, D. Deep joint source-channel coding for wireless image transmission. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 567–579. [CrossRef]
9. Ballé, J.; Laparra, V.; Simoncelli, E.P. End-to-end optimized image compression. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
10. Ballé, J.; Minnen, D.; Singh, S.; Hwang, S.J.; Johnston, N. Variational image compression with a scale hyperprior. In Proceedings of the International Conference on Learning Representations, Vancouver, QC, Canada, 30 April–3 May 2018.
11. Xiao, Z.; Yao, S.; Dai, J.; Wang, S.; Niu, K.; Zhang, P. Wireless deep speech semantic transmission. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
12. Schuchman, L. Dither signals and their effect on quantization noise. *IEEE Trans. Commun. Technol.* **1964**, *12*, 162–165. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017): 31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
14. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.G.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2978–2988.
15. Garofolo, J.S. *Timit Acoustic Phonetic Continuous Speech Corpus*; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
16. Muda, L.; Begam, M.; Elamvazuthi, I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv* **2010**, arXiv:1003.4083.
17. Bessette, B.; Salami, R.; Lefebvre, R.; Jelinek, M.; Rotola-Pukkila, J.; Vainio, J.; Mikkola, H.; Jarvinen, K. The adaptive multirate wideband speech codec (AMR-WB). *IEEE Trans. Speech Audio Process.* **2002**, *10*, 620–636. [CrossRef]
18. Valin, J.M.; Vos, K.; Terriberry, T. Definition of the Opus Audio Codec, Technical Report. 2012. Available online: https://www.rfc-editor.org/rfc/pdfrfc/rfc6716.txt.pdf (accessed on 1 July 2022).
19. Ryan, W.; Lin, S. *Channel Codes: Classical and Modern*; Cambridge University Press: Cambridge, UK, 2009.
20. Peng, F.; Zhang, J.; Ryan, W.E. Adaptive modulation and coding for IEEE 802.11 n. In Proceedings of the 2007 IEEE Wireless Communications and Networking Conference, Hong Kong, 11–15 March 2007; pp. 656–661.
21. ITU-T. *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*; International Telecommunication Union: Geneva, Switzerland, 2001.
22. BS Series. *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*; International Telecommunication Union: Geneva, Switzerland, 2014.
23. Liu, L.; Oestges, C.; Poutanen, J.; Haneda, K.; Vainikainen, P.; Quitin, F.; Tufvesson, F.; De Doncker, P. The COST 2100 MIMO channel model. *IEEE Wirel. Commun.* **2012**, *19*, 92–99. [CrossRef]