文稿编号：IMT-2030-Semantic_2024xxx

会议名称	IMT-2030(6G)语义通信任务组第 4 次会议	会议地点	上海
提交单位	北京邮电大学	会议时间	2024 年 5 月 16 日

基于非线性变换编码的深度语音语义传输（DSST）技术提案

作者：甘善辉、肖子轩、姚圣时、戴金晟、王思贤、牛凯、张平

1. 背景

无线信道语音传输方法通常可以分为两个步骤：信源编码和信道编码[1]。信源编码对信号波形执行线性变换，从而消除冗余和解相关。信道编码则针对不完善的无线信道提供纠错。然而，传统的分离编码方案的最优性能仅在无限信源和信道编码块长度的渐近极限下成立，这在实际通信场景中是不可能的。此外，传统的分离编码方案在面对信道容量与通信速率不匹配时，会出现性能的急剧下降，即所谓的“悬崖效应”。

目前，针对语音信源的语义通信技术大多从联合信源信道编码的角度出发，通过联合考虑信源的语义特征以及信道的统计特性，对编码传输方法进行优化设计，实现端到端的性能提升。传统的联合信源信道编码（Joint Source-Channel Coding, JSCC）方法基于统计概率，而忽略信源语义方面。随着深度学习的发展，越来越多的研究人员致力于利用神经网络提取语义特征并实现 JSCC 以实现高效的语义传输[2-4]。这些方法相较于传统通信方法获得了有效的性能提升，并在低信噪比下展现了强大的鲁棒性。但是这种方法不能根据信源的语义特征的重要性不同实现变长速率的编码与传输，导致在传输过程中出现带宽的不必要浪费。最近的研究 DeepSC-S[5]采用深度联合信源信道编码的方法来准确传输语音波形，显示了联合编码方案的潜力。其直接对脉冲编码调制（pulse code modulation, PCM）信号进行编码，而没有有效消除样本之间的相关性，这会导致带宽的巨大浪费。

本提案介绍了一种基于非线性变换编码的深度语音语义传输（Deep Speech Semantic Transmission, DSST）技术方案。其结合经典传输方案和深度学习的优点，实现高效的端到端语音语义传输。本方案首先引入非线性变换来提取波形语义特征[6-9]，随后利用 Transformer 架构[10]将语义特征直接编码到信道输入符号。在此过程中，通过可学习的熵模型在语义潜在空间中估计语义特征的重要性，并将该值馈送到编码器中，以更合理地指导编码速率地分配，从而获得更高的编码增益。DSST 的设计框架被构建为一个优化问题，其目标是在保证听觉感知质量的同时，最小化端到端的传输速率失真（RD）性能。本方案在任何给定的传输速率下都能实现尽可能高的听觉感知性能。此外，本方案引入了信道信噪比（SNR）自适应机制，使得单个模型可以应用于各种信道状态。实验结果验证了 DSST 方案在客观和主观指标上明显优于当前工程语音传输系统。本提案给出的 DSST 方案，能够促进语音语义传输在未来移动通信网络中的应用，且具有信道带宽消耗小和语音重建质量高的特性。

会议名称	IMT-2030(6G)语义通信任务组第 4 次会议	会议地点	上海
提交单位	北京邮电大学	会议时间	2024 年 5 月 16 日

2. 基本框架

2.1 DSST 的基本框架

本提案提出的 DSST 框架利用了非线性变换编码的方法，将语音信源分帧加窗后的各帧音频序列映射到语义潜在空间，并将语义特征发送到联合信源信道编码器中，以生成信道输入序列，并在接收端进行相应的联合信源信道解码与音频重建。

图 1 为 DSST 的基本结构示意图。

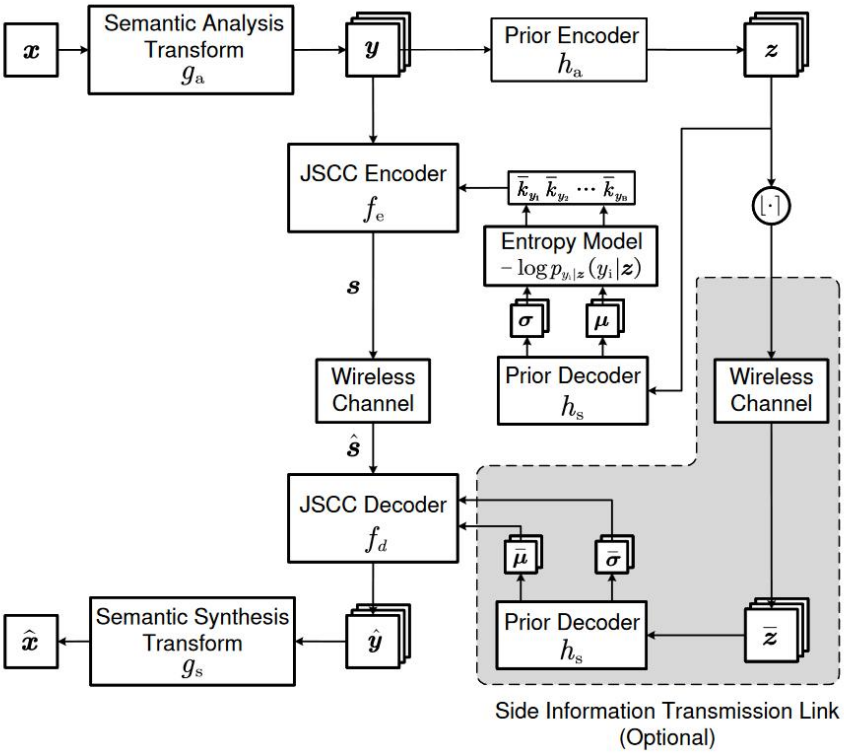

图 1 DSST 的基本结构示意图

总体来看，图 1 的上半部分为 DSST 的发射机，中间为信道，下半部分是 DSST 的接收机。信源为语音信源，其直接送入发射机中。在发射机中，一方面，利用语义解析变换提取信源语义特征作为语义特征空间中的语义潜在表示。随后 DJSCC 编码器在这个潜在空间上进行操作；另一方面，引入超先验解析变换，对潜在的语义变量进一步解相关，作为边信息，并通过超先验综合变换估计每帧语音的语义特征信息量，用于指导联合信源信道编码的输出维度，起到传输速率自适应的效果。接收机可以在边信息的指导下进行联合信源信道译码，再通过语义综合变换重构语音帧，也可以无需边信息，直接进行联合信源信道译码之后再重建信源。

DSST 系统的总链路表述如下

会议名称	IMT-2030(6G)语义通信任务组第 4 次会议	会议地点	上海
提交单位	北京邮电大学	会议时间	2024 年 5 月 16 日

$$\boldsymbol{x} \xrightarrow{g_a(\cdot)} \boldsymbol{y} \xrightarrow{f_e(\cdot)} \boldsymbol{s} \xrightarrow{W(\cdot|\boldsymbol{h})} \hat{\boldsymbol{s}} \xrightarrow{f_d(\cdot)} \hat{\boldsymbol{y}} \xrightarrow{g_s(\cdot)} \hat{\boldsymbol{x}} \tag{1}$$

潜在超先验链路表述如下

$$\boldsymbol{y} \xrightarrow{h_a(\cdot)} \boldsymbol{z} \xrightarrow{h_s(\cdot)} \{\boldsymbol{\mu}, \boldsymbol{\sigma}\} \tag{2}$$

在本提案中考虑 AWGN 信道与 COST2100 无线衰落信道[11]

$$\hat{\boldsymbol{s}} = W(\boldsymbol{s}|\boldsymbol{h}) = \boldsymbol{h} \odot \boldsymbol{s} + \boldsymbol{n} \tag{3}$$

其中，$\boldsymbol{h}$ 是 CSI 向量，$\odot$ 是哈达马积，$\boldsymbol{n}$ 定义为噪声向量，其独立采样于高斯分布 $\boldsymbol{n} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_n^2 \mathbf{I})$。接收机的结构采用镜像架构。

接下来对图 1 表示的 DSST 框架每个模块的作用进行详细介绍：

- 语义解析变换模块 g_a：解析变换 g_a 的作用是利用非线性变换，将信源 $\boldsymbol{x}$ 映射到语义特征空间中，并通过相关性的去除，解析获得隐式特征 $\boldsymbol{y}$；
- 语义综合变换模块 g_s：综合变换 g_s 的作用是利用对隐变量的估计 $\hat{\boldsymbol{y}}$，对信源数据进行重建（综合），输出重建后的信源 $\hat{\boldsymbol{x}}$；
- 超先验解析变换模块 h_a：超先验解析变换 h_a 用于对隐变量 $\boldsymbol{y}$ 进一步地进行解相关，输出隐变量 $\boldsymbol{z}$，其中 $\boldsymbol{z}$ 也被称作超先验；
- 超先验综合变换模块 h_s：超先验综合变换 h_s 以具有随机性的隐变量 $\tilde{\boldsymbol{z}}$（或量化后的隐变量）为条件，为隐式特征 $\boldsymbol{y}$ 输出相应的均值序列和方差序列，以建立混合高斯分布，并在 $y_i - \frac{1}{2}$ 到 $y_i + \frac{1}{2}$ 的范围内计算概率值。
- 熵模型：经过超先验解析变换模块 h_a 与超先验综合变换模块 h_s 的处理后，熵模型利用 $-\log P(y_i|\tilde{\boldsymbol{z}})$ 计算每个语音帧的语义特征信息量，作为联合信源信道编码的指导，可以对传输速率进行控制；
- 可变速率 DJSCC 模块：DJSCC 编码器 f_e 以隐变量 $\boldsymbol{y}$ 为输入，熵模型计算出的 $-\log P(y_i|\tilde{\boldsymbol{z}})$ 为指导，为每个语音帧进行不同输出维度的降维变换，输出变长的序列 $\boldsymbol{s}$；DJSCC 解码器 f_d 以变长传输序列 $\hat{\boldsymbol{s}}$ 为输入，为每个语音帧进行重建，并拼接组合成隐变量的估计值 $\hat{\boldsymbol{y}}$；

需要说明的是，经过量化后的超先验 $\overline{\boldsymbol{z}}$，既可以传输到接收端，作为编解码器共享的边信息，辅助进行信源的重建，也可以不将其传输到接收端，仅在得知每个语音帧的信道符号长度的条件下，恢复隐式特征和信源信息。若传输 $\overline{\boldsymbol{z}}$ 时，接收端可以联合地利用接收

的符号$\hat{\boldsymbol{s}}$，与$\overline{\boldsymbol{z}}$经过超先验综合变换$h_s$后得到的均值$\overline{\boldsymbol{\mu}}$和方差$\overline{\boldsymbol{\sigma}}$作为联合信源信道解码器$f_d$的输入，以获得更好的重建性能。相应地，不传输$\overline{\boldsymbol{z}}$可以一定程度的节省信道带宽成本，但解码质量会有所下降。

2.2 变分模型和信源信道编码

在人类的直觉中，无声时的语音信号携带很少的信息。因此，应为具有零幅度的信号分配较少的信道带宽。如果为所有语音帧的语义特征分配相同的编码率，则无法最大化编码效率。为了解决这个问题，DSST 使用了一个熵模型来准确估计$\boldsymbol{y}$的分布，来评估不同语义特征的重要性。此外，DSST 还使 Deep JSCC 编码器能够感知信源的语义特征，并根据语义特征的重要性合理分配编码速率。

DSST 的熵模型如图 1 的中间部分所示。按照[8]的工作，语义特征$\boldsymbol{y}$被变分建模为具有均值$\boldsymbol{\mu}$和标准差$\boldsymbol{\sigma}$的多元高斯分布。因此，$\boldsymbol{y}$的真实后验分布可以通过完全分解的概率密度模型来建模：

$$p_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_i \left(\mathcal{N}(y_i|\mu_i,\sigma_i) * \mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)\right)(y_i)$$
$$with\ (\boldsymbol{\mu},\boldsymbol{\sigma}) = h_s(\boldsymbol{z}) \tag{4}$$

其中，$*$是卷积运算，$\mathcal{U}$表示均匀分布，其用于将先验分布与边缘分布相匹配，使得估计的熵$-\log p_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z})$为非负值。在优化过程中，$h_s$创建的概率分布$q_{\boldsymbol{y}|\boldsymbol{z}}$将逐渐逼近真实分布$p_{\boldsymbol{y}|\boldsymbol{z}}$，因此熵模型可以准确地估计$\boldsymbol{y}$的分布。

每帧的熵值$-\log p_{y_i|\boldsymbol{z}}(y_i|\boldsymbol{z})$将被馈送到 JSCC 的编码器并相应地指导编码速率的分配。如果熵模型指示y_i为高熵，则编码器将分配更多资源来传输它。传输$\boldsymbol{y}$的总信道带宽成本K可以写为

$$K = \sum_{i=1}^{B} \overline{k}_{y_i} = \sum_{i=1}^{B} Q(k_{y_i}) = \sum_{i=1}^{B} Q(-\eta_y \log p_{y_i|\boldsymbol{z}}(y_i|\boldsymbol{z})) \tag{5}$$

其中，B是语音信号的帧数，η_y是用于平衡信道带宽成本与估计熵的超参数，$\overline{k}_{y_i}$是第i帧的带宽消耗，Q表示标量量化，其范围包括n个整数。

2.3 优化目标

DSST 系统的优化目标是在信道带宽成本和语音重建质量之间实现 RD 权衡。Balle 证明，将整个模型建模为变分自编码器[8]更为合理，这样最小化 KL 散度就相当于优化模型的 RD 性能。因此，推理模型的目标是创建变分概率密度$q_{\hat{\boldsymbol{s}},\hat{\boldsymbol{z}}|\boldsymbol{x}}$以近似真实的后验$p_{\hat{\boldsymbol{s}},\hat{\boldsymbol{z}}|\boldsymbol{x}}$。RD 函数可以近似为

会议名称	IMT-2030(6G)语义通信任务组第 4 次会议	会议地点	上海
提交单位	北京邮电大学	会议时间	2024 年 5 月 16 日

$$L_{RD} = \mathbb{E}_{\boldsymbol{x}\sim p_x} D_{KL}\left[q_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}} || p_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}}\right] \Leftrightarrow$$
$$\mathbb{E}_{\boldsymbol{x}\sim p_x}\mathbb{E}_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}\sim q_{\hat{s},\tilde{z}|x}}\left[\left(-\log p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) - \log p_{\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}}}(\hat{\boldsymbol{s}}|\hat{\boldsymbol{z}})\right) + \mathbb{E}_{\boldsymbol{y}\sim p_{y|\hat{s},\cdot\tilde{z}}}\left[\log p_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})\right]\right] \tag{6}$$

第一项为边缘分布 $q_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) = \mathbb{E}_{\boldsymbol{x}\sim p_x}\mathbb{E}_{\hat{\boldsymbol{s}}\sim q_{\hat{s}|x}}\left[q_{\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x}}(\hat{\boldsymbol{s}},\tilde{\boldsymbol{z}}|\boldsymbol{x})\right]$ 和超先验概率 $p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})$ 之间的交叉熵，其代表着假设 $p_{\tilde{\boldsymbol{z}}}$ 作为熵模型对边信息进行编码的成本。$p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})$ 被建模为非参数完全分解概率密度[8]

$$p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) = \prod_i \left(p_{z_i|\psi^{(i)}}(\psi^{(i)}) * \mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)\right) \tag{7}$$

其中 $\psi^{(i)}$ 封装了 $p_{z_i|\psi^{(i)}}$ 的所有参数。此外为了在模型训练阶段通过梯度下降进行优化[7]，$\boldsymbol{z}$ 添加了服从均匀分布的偏移量 $\boldsymbol{o}$，而不是标量量化 $\bar{\boldsymbol{z}} = \lfloor \boldsymbol{z} \rceil$，即 $\tilde{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{o}$ 且 $o_i \sim \mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)$。

第二项与第一项类似，代表编码 $\hat{\boldsymbol{s}}$ 的成本，表示语音信号的传输速率。在实践中，可以利用中间变量 $\boldsymbol{y}$ 来推导 $p_{\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}}}$。在传输过程中，$\boldsymbol{y}$ 不经过量化直接送入 JSCC 编码器和信道中。这个过程可以被描述为

$$p_{\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}}} = W(p_{\boldsymbol{s}|\tilde{\boldsymbol{z}}}|\boldsymbol{h}) = W(f_e(p_{\boldsymbol{y}|\tilde{\boldsymbol{z}}})|\boldsymbol{h}) \tag{8}$$

因此，概率密度 $p_{\hat{\boldsymbol{s}}|\tilde{\boldsymbol{z}}}$ 可以近似为 $p_{\boldsymbol{s}|\tilde{\boldsymbol{z}}}$ 并转换为 $p_{\boldsymbol{y}|\tilde{\boldsymbol{z}}}$。类似地，$\boldsymbol{z}$ 被添加一个服从均匀分布的偏移量而不是标量量化，传输速率被按比例限制为 $-\log P_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z})$。

第三项表示恢复 $\boldsymbol{x}$ 的对数似然。

根据以上分析，RD 函数可以简化为

$$\begin{aligned} L_{RD} = \mathbb{E}_{\boldsymbol{x}\sim p_x}\big[&-\eta_y \log p_{\boldsymbol{y}|\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z}) - \eta_z \log p_{\tilde{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \\ &+ \lambda_{MSE} d_{MSE}(\boldsymbol{x},\hat{\boldsymbol{x}}) + \lambda_{Perc} d_{Perc}(\boldsymbol{x},\hat{\boldsymbol{x}})\big] \end{aligned} \tag{9}$$

其中拉格朗日乘子 λ 决定了无线传输总带宽成本 R 和端到端失真 D 之间的权衡。η_y 和 η_z 是平衡 $\boldsymbol{y}$ 和 $\boldsymbol{z}$ 的信道带宽消耗的两个超参数。$d(\cdot,\cdot)$ 是原始信号和重建信号之间的失真。

在 DSST 方案中，在时域和频域中都使用了失真函数 d。首先，利用原始信号 $\boldsymbol{x}$ 和重建信号 $\hat{\boldsymbol{x}}$ 之间的均方误差（MSE）来评估时域中的重建误差，其权衡系数为 λ_{MSE}。然后，计算原始信号和重建信号的 MFCC[12]，并在频域中使用 MFCC 向量之间的 l_2 距离来追求更好的人类感知质量[13]。考虑到频谱倾斜现象[14]，如果在频域中使用 MSE 作为损失函数，高频的相对误差会更加显著，导致高频信息的丢失。因此，最终应用归一化均方误差（NMSE）作为损失函数，其权衡系数为 λ_{Perc}。这有助于模型针对重建高频进行优化，并

赋予重建语音更高的人类感知质量。

3. 网络架构和实现

接下来对本提案中 DSST 方案里的各个模块的实现方案进行介绍。

以语音信号为例。经过归一化、分帧、加窗等预处理操作后，输入帧$\boldsymbol{x}$被建模为向量$\boldsymbol{x}\in\mathbb{R}^{B\times C\times L}$，其中$B$是帧数，$C$是声道数，$L$是帧长度。

- 语义解析变换和语义综合变换g_a,g_s：语义解析变换和语义综合变换采用与 Kankanahalli 网络[15]类似的架构，它包含多层一维卷积块和残差网络，可以有效地提取语音波形的语义特征。两次下采样后，潜在特征维度为$\boldsymbol{y}\in\mathbb{R}^{B\times M\times\frac{L}{4}}$，$M$是通道维度。如下图 3 所示：

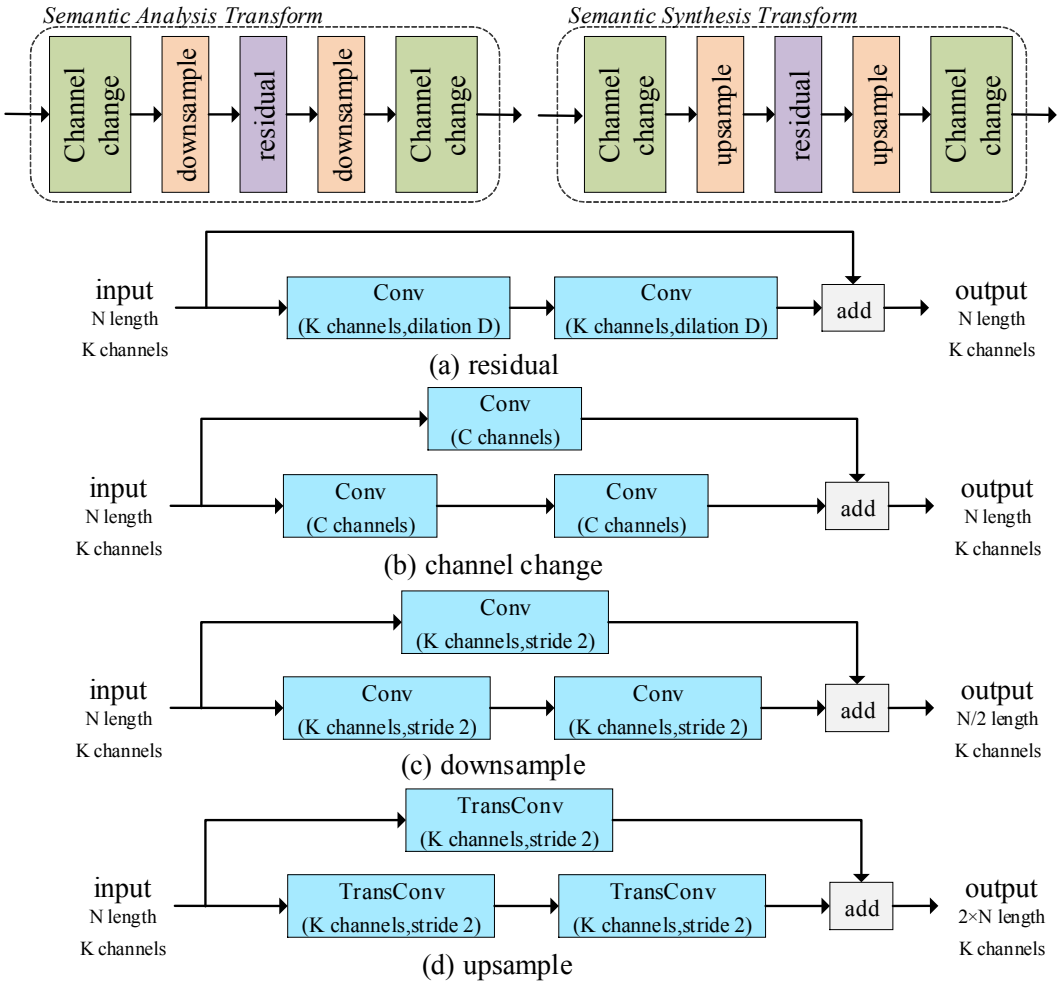

图 2 语义解析变换和语义综合变换的结构

- 超先验编解码器h_a,h_s：如图 3 所示，编码器h_a的结构由 3 个N通道的卷积块组成。第一个块包含内核大小为 3 的卷积层和 ReLU 激活函数。其他块由下采样层组成，该下采样层由跨步卷积组成，内核大小为 5，步幅为 2，后跟 ReLU 激活函数。解码器

会议名称	IMT-2030(6G)语义通信任务组第 4 次会议	会议地点	上海
提交单位	北京邮电大学	会议时间	2024 年 5 月 16 日

h_s 与编码器 h_a 呈镜像结构。超先验模型计算了 $\boldsymbol{z} \in \mathbb{R}^{B \times N \times \frac{L}{16}}$ 中的均值和标准差的分布。它有效地捕获语音波形的语义时间关系。

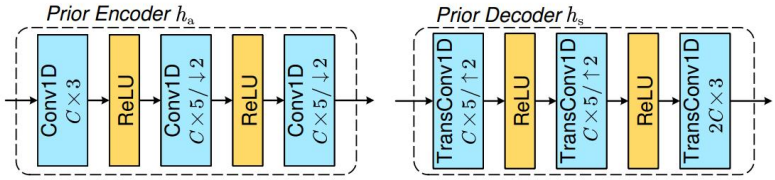

图 3 超先验网络的结构。

Conv1D$C\times k$ 是一个具有 C 个通道和 k 个滤波器的一维（1D）卷积层，↓表示降维，步长为 2。

- 自适应速率 Deep JSCC 编解码器 f_e, f_d：Deep JSCC 编解码器的结构如图 4 所示。

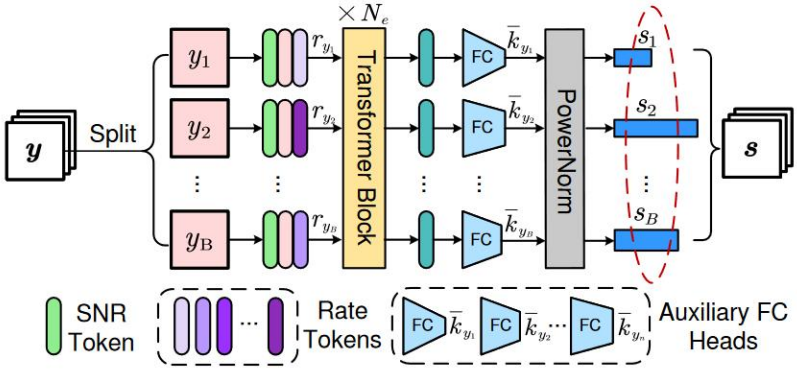

图 4 自适应速率 Deep JSCC 编码器的结构。Deep JSCC 解码器的结构采用镜像架构。

编码器 f_e 包含 N_e 个 Transformer 块和 FC 层，以实现速率分配。具体来说，$\boldsymbol{y}$ 首先被分为块嵌入序列$\{y_1, y_2, \cdots, y_B\}$。为了自适应地将 y_i 映射到 $\overline{k}_{y_i}$ 维信道输入向量 $\boldsymbol{s}_i$，DSST 使用了一个速率标记向量集 $\mathcal{R} = \{r_{y_1}, r_{y_2}, \cdots, r_{y_n}\}$ 来表示速率信息和一个量化值集合 $\mathcal{V} = \left\{\overline{k}_{y_1}, \overline{k}_{y_2}, \cdots, \overline{k}_{y_n}\right\}$ 来表示输出维度。通过这种方法，每个 y_i 与熵 $-\log_{y_i|\boldsymbol{z}}(y_i|\boldsymbol{z})$ 对应的速率标记 r_{y_i} 合并，然后被发送到 Transformer 块和 FC 层以转换为 $\overline{k}_{y_i}$ 维向量。特别地，DSST 采用了一组具有不同输出维度的 FC 层，其输出维度由速率标记 r_{y_i} 指导。

此外，DSST 在这个过程中进行了 SNR 自适应。DSST 方案假设发射机和接收机可以接收 SNR 反馈信息以获得更好的性能。如图 4 所示，每个块 y_i 都与 SNR 标记 $C \in \mathbb{R}^B$ 连接，使得 Transformer 可以利用自注意力机制来学习 SNR 信息。因此，当在随机 SNR 下训练时，单个模型最终的性能至少可以与单独针对每个 SNR 训练的模型一样好。

- 可选的传输链路：如果 DSST 传输 $\boldsymbol{z}$ 来获得解码增益，其会先对 $\boldsymbol{z}$ 进行量化，然后对其进行熵编码和信道编码，以保证可靠传输。在接收机处，超先验解码器重建边信息 $\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}}$ 并将其馈送到 Deep JSCC 解码器。

会议名称	IMT-2030(6G)语义通信任务组第 4 次会议	会议地点	上海
提交单位	北京邮电大学	会议时间	2024 年 5 月 16 日

4. 实验结果

4.1 实验准备

- 数据集：本提案中的 DSST 模型训练采用 TIMIT[16]数据集中的 16kHz 音频，该数据集总共包含 6300 个句子，由来自美国 8 个主要方言地区的 630 个说话者每人说出 10 个句子组成。在每个 Batch 中包含$B=100$帧，每帧长度为$L=512$ 个样本。
- 对比方案：采用经典的信源信道分离编码方案 AMR-WB+5G LDPC 和 Opus+5G LDPC[17-19]和标准神经网络 JSCC 模型 DeepSC-S。对于传统的分离编码方案，本提案根据自适应调制编码（adaptive modulation coding, AMC）原理[20]选择不同的编码和调制方案。
- 评估指标：本提案使用客观评估指标并进行主观实验评估重建语音质量。关于客观质量评估，本提案测量语音质量感知评估[21]（perceptual evaluation of speech quality, PESQ）分数以反映重建的语音质量。此外，本提案实施 MUSHRA 主观测试[22]来由人类评估者评估重建音频质量，其从测试数据集中随机选择十个重建的波形信号。

4.2 结果分析

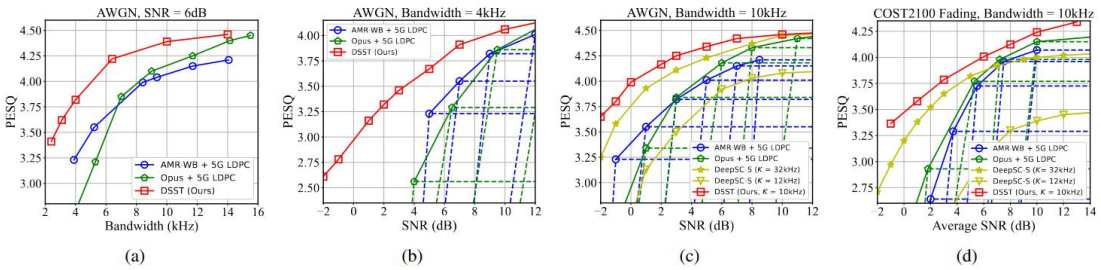

图 5 (a)在 AWGN 信道上 SNR=6dB 时，PESQ 性能与信道带宽成本的关系 (b)在 AWGN 信道上，带宽成本（K）为 4kHz，PESQ 性能与信道 SNR 的关系 (c)(d)在 AWGN 信道和 COST2100 5.3GHz 室内衰落信道上，PESQ 性能与信道 SNR 的关系，K 为 10kHz，但 Deep SC-S 的 K 为 12kHz 和 32kHz（绿线）

图 5(a)显示了 AWGN 信道上各种传输方法的速率-质量结果，信道 SNR 为 6dB。DSST 方案在所有带宽成本情况下都优于传统的 AMR-WB+5G LDPC 和 Opus+5G LDPC。这主要是因为 DSST 方案能准确地估计语音信源的语义特征分布，从而减轻了信源编码和信道编码之间的不匹配从而获得了压缩增益。此外，DSST 方案展示了其灵活性。该方案可以实现任意传输速率下的高效语音传输。传统方案中，AMR-WB 和 Opus 方案都无法有效地在低带宽下传输语音。

图 5(b)和图 5(c)显示了各方案在不同 SNR 下的性能。与上述结果类似，DSST 方案在所有 SNR 上都非常明显地优于传统的分离编码方案。此外，如图 5(a)和图 5(c)所示，当

PESQ 为 4.3 且 SNR 为 6dB 时，DeepSC-S 使用 32kHz 带宽，而 DSST 方案仅使用 8kHz 带宽，这节省了约 75%的信道带宽成本。这种增益来自语义转换模块和 Deep JSCC 模块。前者丢弃了语音信号中的大部分冗余，后者更合理地分配不同的编码速率，充分利用了编码资源。

图 5(d)为在实际的 COST2100 衰落信道上的实验结果。CSI 样本在 5.3GHz 频段室内场景采集，所有方案均采用一次性传输。如图 5(d)所示，DSST 方案对信道变化表现出很强的鲁棒性。特别是在低信噪比的情况下，DSST 方案明显地优于其他方案。与现有的神经网络方法 DeepSC-S 相比，当使用相似的带宽资源时，DSST 模型在 PESQ 性能方面实现了 30%以上的提升。

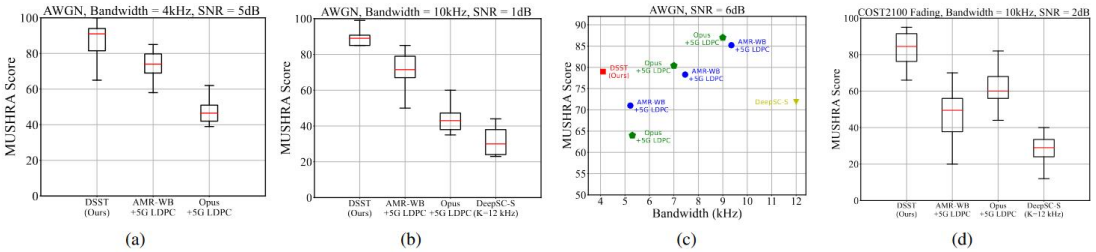

图 6 在以下情况的 MUSHRA 分数 (a)K=4kHz, SNR=5dB, AWGN 信道 (b)K=10kHz, SNR=1dB, AWGN 信道 (c)DSST@4kHz 与经典的传统分离编码的编解码器，AWGN 信道 (d) K=10kHz, SNR=5dB, COST2100 衰落信道

图 6 显示了 MUSHRA 主观测试用户评分。前三张图展示了 AWGN 信道下的结果。图 6(a)展示了低带宽成本条件下的性能。DSST 方案显著地超越了两种传统的分离编码方案。图 6(b)展示了低 SNR 情况下的性能。DSST 方案在相似带宽下获得的 MUSHRA 分数是 DeepSC-S 的三倍。图 6(c)比较了不同带宽成本下的各种传输方案。为了匹配 DSST 的语音质量，Opus+5G LDPC 需要使用 7kHz 带宽，而 AMR-WB+5G LDPC 至少需要 7.5kHz 带宽。DSST 方案能节省带宽资源 45%以上。此外，DSST 方案使用低至 4kHz 的带宽重建了高质量的语音，其质量明显优于 12kHz 的 DeepSC-S。最后，图 6(d)展示了 COST2100 衰落信道上不同方案的性能。DSST 方案在真实衰落信道中仍然优于其他方案。

5. 总结

本提案介绍了一种基于非线性变换编码的深度语音语义传输方案：DSST。DSST 方案采用了非线性变换方法、可学习的熵模型和自适应速率的 Deep JSCC 方法来获得更高的编码增益。实验结果表明，DSST 方案在各种传输速率下都取得了更好的性能，并且在 AWGN 信道和 COST2100 无线信道下均优于传统的分离编码方案和现有神经网络方案

DeepSC-S。值得强调的是，DSST 方案在恶劣的信道条件下，也能成功重建高质量的语音信号。DSST 方案的提出，为高效可靠的语音语义传输提供了新的解决方案，也为相关领域的研究者和工程师提供了宝贵的技术支持和理论参考。

参考文献

[1] Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[2] Bourtsoulatze E, Kurka D B, Gündüz D. Deep joint source-channel coding for wireless image transmission[J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(3): 567-579.

[3] Kurka D B, Gündüz D. Bandwidth-agile image transmission with deep joint source-channel coding[J]. IEEE Transactions on Wireless Communications, 2021, 20(12): 8081-8095.

[4] Farsad N, Rao M, Goldsmith A. Deep learning for joint source-channel coding of text[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 2326-2330.

[5] Weng Z, Qin Z. Semantic communication systems for speech transmission[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2434-2444.

[6] Ballé J, Chou P A, Minnen D, et al. Nonlinear transform coding[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 15(2): 339-353.

[7] Ballé J, Laparra V, Simoncelli E P. End-to-end optimized image compression[J]. arXiv preprint arXiv:1611.01704, 2016.

[8] Ballé J, Minnen D, Singh S, et al. Variational image compression with a scale hyperprior[J]. arXiv preprint arXiv:1802.01436, 2018.

[9] Dai J, Wang S, Tan K, et al. Nonlinear transform source-channel coding for semantic communications[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(8): 2300-2316.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[11] Liu L, Oestges C, Poutanen J, et al. The COST 2100 MIMO channel model[J]. IEEE Wireless Communications, 2012, 19(6): 92-99.

[12] Muda L, Begam M, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. arXiv preprint arXiv:1003.4083, 2010.

[13] Blau Y, Michaeli T. The perception-distortion tradeoff[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6228-6237.

[14] Chen J H, Gersho A. Adaptive postfiltering for quality enhancement of coded speech[J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 59-71.

[15] Kankanahalli S. End-to-end optimized speech coding with deep neural networks[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 2521-2525.

[16] Garofolo J S. Timit acoustic phonetic continuous speech corpus[J]. Linguistic Data Consortium, 1993, 1993.

[17] Bessette B, Salami R, Lefebvre R, et al. The adaptive multirate wideband speech codec (AMR-WB)[J]. IEEE transactions on speech and audio processing, 2002, 10(8): 620-636.

[18] Richardson T, Kudekar S. Design of low-density parity check codes for 5G new radio[J]. IEEE Communications Magazine, 2018, 56(3): 28-34.

[19] Valin J M, Vos K, Terriberry T. Definition of the opus audio codec[R]. 2012.

[20] Peng F, Zhang J, Ryan W E. Adaptive modulation and coding for IEEE 802.11 n[C]//2007 IEEE Wireless Communications and Networking Conference. IEEE, 2007: 656-661.

[21] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221). IEEE, 2001, 2: 749-752.

[22] Series B. Method for the subjective assessment of intermediate quality level of audio systems[J]. International Telecommunication Union Radiocommunication Assembly, 2014.