专家视点

面向卫星通信的低速率语音编码技术综述

魏晨光，许珈艺，郭勐，杨蕾
（中国移动通信有限公司研究院，北京 100053）

摘　要：随着天地一体化信息网络的建设，卫星直连手机终端逐步普及。如何在卫星链路资源受限的情况下实现稳定清晰的语音通信，成为卫星语音通信业务发展的核心挑战。由于卫星信道具有带宽受限、路径损耗大、时延高等特点，地面蜂窝网络的语音编码难以直接适用，低速率语音编码技术是实现卫星语音服务的关键。基于此，系统总结了面向卫星通信的低速率语音编码技术，介绍了主流技术路线的原理、特点及性能评估，分析各方法的优缺点，并展望未来研究方向。

关键词：卫星通信；低速率语音编码器；语音质量
中图分类号：TN927+.2；TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2026098

A review of low bit-rate speech codec for satellite communication

Wei Chenguang, Xu Jiayi, Guo Meng, Yang Lei
China Mobile Research Institute, Beijing 100053, China

Abstract: With the advancement of the space-integrated-ground network, device-to-satellite communication is transitioning from concept to reality. Achieving stable and clear voice communication with limited satellite link resources is a key challenge for the industry. Due to the bandwidth limitations, high path loss, and high transmission delays of satellite channels, speech codecs used in terrestrial networks are not directly adaptable to satellite communication scenarios. Therefore, low bit-rate speech codec is crucial for satellite voice services. Based on this, the low bit-rate speech codec technologies for satellite communication were systematically summarized, the principles, characteristics, and performance evaluations of mainstream technical routes were introduced, the advantages and disadvantages of each method were analyzed, and future research directions were prospected.

Key words: satellite communication, low bit-rate speech codec, speech quality

收稿日期：2026−01−04；**修回日期：**2026−01−19
通信作者：杨蕾，yangleiyj@chinamobile.com
基金项目：国家自然科学基金资助项目（No.U21B2004）
Foundation Item: The National Natural Science Foundation of China (No.U21B2004)

0 引言

随着天地一体化通信体系的加速构建，卫星通信产业正迎来重要的发展机遇[1]。其中，手机直连卫星语音业务，不仅丰富了个人消费市场的应用场景，促进了通信与汽车、船舶、物联网等多行业的融合创新，也成为连接偏远地区、保障灾害应急通信，以及推动未来6G发展的重要实现路径[2]。美国卫星产业协会（Satellite Industry Association，SIA）于2025年6月发布的第28版卫星产业状况年度报告显示[3]，2024年卫星移动通信业务收入为25亿美元，增长率约为6%。增长动力来自基于卫星移动业务频段的端到端移动语音和数据（包括物联网）服务、商业物联网和批发容量服务的需求。

然而，与地面通信相比，卫星通信系统的链路资源相对受限，这对传统语音编码技术构成了挑战。从现有卫星通信系统的实际应用来看，如国际海事卫星组织（International Maritime Satellite Organization，Inmarsat）推出的Inmarsat-C[4]，是一种典型的低速率、双向全球卫星移动数据通信系统，其支持的语音速率仅为1.2 kbit/s。在国内，“天通一号”卫星语音系统也支持0.8 kbit/s、1.2 kbit/s等多个低速率语音编码档位[5]，而目前地面通信中广泛使用的语音编码器（如自适应多速率（adaptive multi-rate，AMR）编码[6]、增强语音服务（enhanced voice service，EVS）编码[7]等）最低码率仍需要4.75 kbit/s以上，难以直接适配卫星窄带链路。因此，低速率语音编码器成为卫星语音业务开展的基础，其性能直接影响用户的通话清晰度和整体体验。此外，由于卫星信道传输距离远、信道条件复杂，且干扰和衰落现象较为严重，往往接收信号的信噪比极低，这对语音编码器的设计与性能提出了更为苛刻的要求，需要在实现高效压缩编码的同时，兼顾降噪、抗丢包等增强功能。

本文的主要贡献在于：首次围绕卫星信道对语音传输的特定技术要求，全面梳理了面向卫星通信的低速率语音编码技术的研究现状，内容涵盖主流技术路线、性能评估方法、系统集成挑战等方面。在此基础上，本文分析了现有技术的特点与不足，并基于对技术趋势的理解，提出了未来研究方向。

1 卫星通信对语音编解码器的技术要求

1.1 手机直连卫星语音通话流程

一个基于IP多媒体子系统（IP multimedia subsystem，IMS）网络架构的典型手机直连卫星语音通话系统如图1所示，其端到端流程如下。

（1）卫星终端（user equipment，UE）（主叫）：将语音信号实时编码为低速率语音编码格式，以适应卫星信道带宽窄、时延高的传输特性。

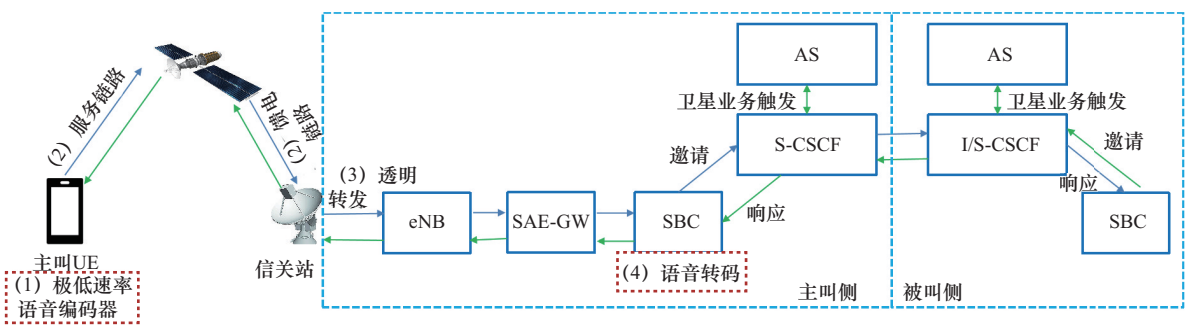

图1 基于IMS网络架构的典型手机直连卫星语音通话系统

（2）卫星链路：编码后的语音数据包经上行链路传输至卫星，再通过下行馈电链路发送至地面信关站。

（3）地面信关站：负责对卫星信号进行透明转发，不改变数据内容。

（4）基站与核心网：语音数据经由基站（即演进型Node B（evolved Node B，eNB））接入，通过系统架构演进网关（system architecture evolution gateway，SAE-GW）送至IMS核心网，会话边界控制器（session border controller，SBC）进行实时转码，将卫星侧的低速率语音流转换为地面蜂窝网络，如长期演进语音承载（voice over long-term evolution，VoLTE）通用的标准语音编码格式。

（5）地面终端（被叫）：转码后的语音流经由地面网络路由至被叫终端，完成端到端通信。

图1中，核心网元包括SBC、查询呼叫会话控制功能（interrogating-call session control function，I-CSCF）及服务呼叫会话控制功能（serving-call session control function，S-CSCF）。其中，SBC负责IMS会话起始协议（session initiation protocol，SIP）信令处理、媒体资源分配、媒体编解码转换及终端位置信息获取，与I-CSCF、S-CSCF交互完成IMS注册与呼叫等流程，并将业务触发至应用服务器（application server，AS）。

1.2　基于卫星通信的低速率语音编码器技术要求

根据卫星信道的特点[8]，其对语音编码器的参考技术要求如下。

（1）语音速率：高轨卫星的语音通话技术要求见表1。在第三代合作伙伴计划（3rd Generation Partnership Project，3GPP）Rel-20技术报告TR 22.887[8]中，高轨卫星的单路语音通话带宽被定义为1～3 kbit/s。对于低轨卫星，也可以通过优化现有语音编码器的编码速率，进一步提升系统的并发容量，从而支持更大规模的用户接入。

（2）语音时延：卫星语音通信的端到端时延由多个部分构成，包括信号传输时延、编码时延、转码时延、缓冲时延、抖动排队时延以及其他处理时延。其中，3GPP对于高轨卫星传播时延的定义为280 ms[8]，语音编码算法时延为20～100 ms，转码时延为10～20 ms，缓冲时延为20～60 ms，抖动排队时延为10～50 ms，其他处理如协议封装、回声消除等算法处理时延为10～30 ms，因此，高轨卫星链路通常具有800～1 000 ms的往返时延。

（3）语音处理复杂度：卫星语音的编解码器需要具备实时处理和并发能力，其复杂度设计应适配移动终端设备中的中央处理器（central processing unit，CPU）或数字信号处理器（digital signal processor，DSP）的资源限制，其中模型参数量是衡量编码器语音处理复杂度的关键指标之一。此外，若内存占用过高，将导致更频繁的动态随机存储器（dynamic random access memory，DRAM）访问，从而产生显著的功耗开销，影响设备的整体性能与续航能力。

（4）语音质量：在实际语音质量评估中，通常采用国际电信联盟电信标准化部门（ITU-T）P.800标准定义的平均意见分（mean opinion score，MOS）[9]评价方法。根据VoLTE通话语音质量规范，要求地面语音MOS值不低于3.5分，对于卫星语音MOS值，其分值应在3分以上。

表1　高轨卫星的语音通话技术要求

场景	UE类型	传输速率		通话建立时间/s
		上行链路/(kbit·s^{-1})	下行链路/(kbit·s^{-1})	
IMS高轨语音通话	手持终端	1～3	1～3	≤30

（5）互联互通要求：搭载低速率语音编码器的卫星终端可通过网络侧转码与现有普通终端实现连接。为确保互通性，低速率语音编码器需要与传统编解码器进行互联，其中转码时延包括传统编解码器本身的编码时延（如AMR/AMR-WB编码为5 ms[6]，EVS编码为12 ms[7]），以及转码处理所需的额外缓冲时延（约为2 ms）。因此，在使用AMR/AMR-WB转码时，总时延增加7 ms，使用EVS转码时，总时延增加14 ms。

2 低速率语音编码算法的主流技术路径

语音编码器的核心在于通过压缩技术去除冗余信息，在保证可接受语音质量的同时，降低传输或存储所需的带宽资源[10]。依据编码速率，语音编码器可分为以下几个类别：速率高于32 kbit/s的属于高速率编码器，如G.711[11]、IVAS[12]等；速率在4～32 kbit/s的属于中速率编码器，如AMR[6]、EVS[7]、Opus[13]等；速率低于4 kbit/s的则属于低速率编码器，而速率低于1.2 kbit/s的可进一步归类为超低速率或极低速率编码器。目前，国际、国内语音编码标准多集中在中、高速率范围，低速率语音编码领域已成为技术创新的重要前沿，其核心设计挑战是在提升压缩率的同时，仍保持足够的可懂度和自然度。

低速率语音编码算法的主流技术路径如图2所示。从技术路径看，目前主流的低速率语音编码技术可分为三大类：基于声道模型的传统语音编码方法、基于AI/神经网络模型的语音编码方法，以及基于语义的语音编码方法。其中，基于神经网络模型的方法还可进一步划分为混合式编码和端到端神经网络语音编码。

2.1 基于声道模型的传统语音编码方法

基于声道模型的传统语音编码器是发展周期长、技术非常成熟的一类语音编码方案。其技术原理源于对人类发音器官（如声带、口腔等）的物理特性进行数学建模，从中提取基频、共振峰等关键声学参数，并基于这些参数进行编码传输[14]。基于声道模型的传统语音编码器的典型参考流程如图3所示。

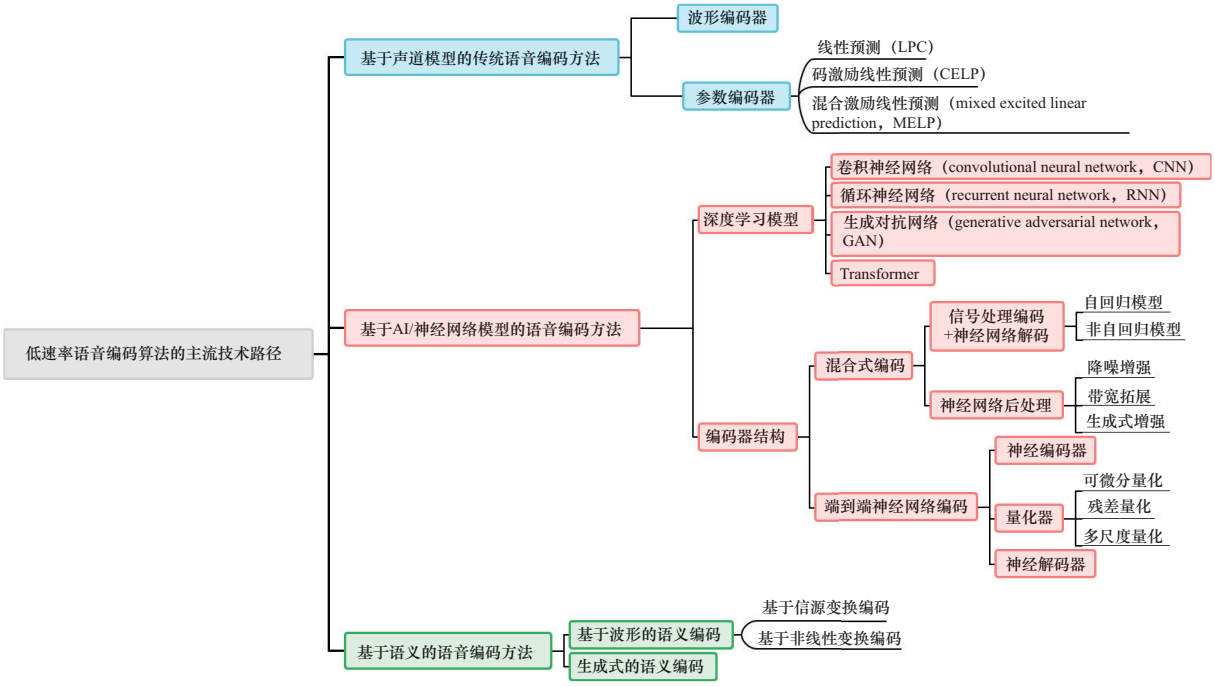

图2 低速率语音编码算法的主流技术路径

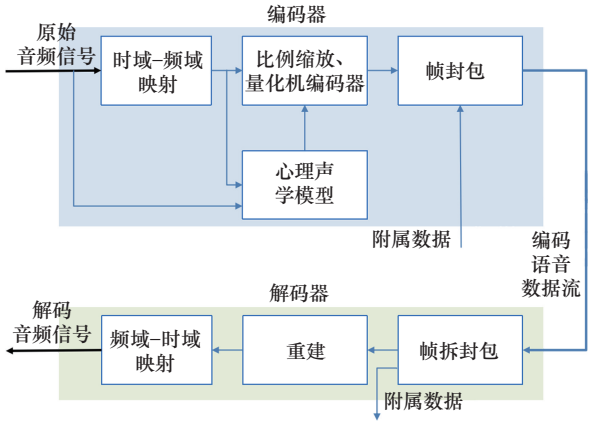

图3　基于声道模型的传统语音编码器的典型参考流程

早期语音编码器将语音视为普通波形信号，直接对采样后的信号进行量化与编码，这类方法通常被称为波形编解码器，典型代表如基于脉冲编码调制（pulse code modulation，PCM）的G.711等。它们在中高码率下表现优异，但在低于8 kbit/s的速率下会产生明显的量化噪声与音质劣化，难以满足卫星通信等对低码率有严格要求的语音传输场景[15]。基于声道模型的传统语音编码器的MOS值–码率关系曲线[16]如图4所示。

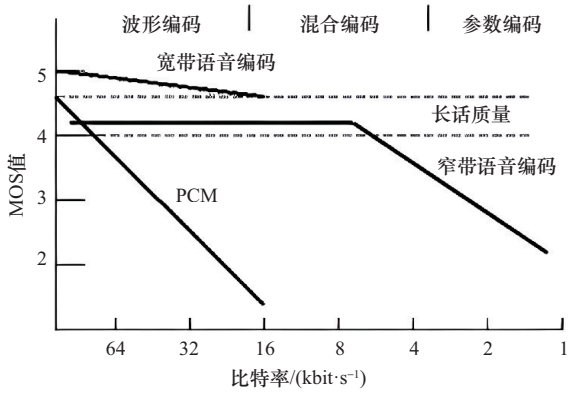

图4　基于声道模型的传统语音编码器的MOS值-码率关系曲线[16]

参数编解码器通过结合参数化建模与波形逼近技术，形成了一系列成熟的主流方案，如码激励线性预测（code excited linear prediction，CELP）[17]系列、AMR[6]和EVS[7]等编码标准。相关的实验结果表明，在9.6～13.2 kbit/s速率范围内，此类编码器（如AMR-WB、EVS-SWB、EVS-WB）表现优异，其MOS值介于3.5～4.0；在5.9～7.2 kbit/s速率下，EVS-WB的MOS值仍可保持在3.2分以上[18]。此类编码方案具有可控的编码时延（10～30 ms）和适中的算法复杂度，但其码率难以进一步降低，否则会导致音质显著下降。开源语音编码器Codec2[19]，虽不完全遵循CELP架构（如未使用固定码本激励），其核心仍依赖于线性预测和参数量化技术，能够实现700 bit/s、1 200 bit/s、2 400 bit/s等级别的极低码率传输。Codec2[19]编解码时延低于50 ms，在ARM Cortex-M4处理器上单帧处理时间不足1 ms，资源占用极低。然而，Codec2因主要针对语音信号优化，对非语音成分的重建能力较弱，主观听感偏机械，MOS值较低，但仍能保持语义可懂。此外，其参数易受噪声影响，抗干扰能力相对有限。

综上所述，传统语音编码器具有实现简单、计算效率高等优点，但表达能力受限，仅能够在低码率下维持语音可懂度，重构语音仍带有明显机械感，且通常仅适用于纯语音信号场景。

2.2　基于神经网络模型的语音编码方法

随着深度学习的发展，基于神经网络模型的语音编码器成为当前语音技术演进的重要方向。其技术原理在于，利用大量语音数据训练神经网络模型，自动学习语音信号的高效紧凑表示及高质量重建过程，并在实际运行时基于预先训练好的参数进行前向推理，实现大量线性与非线性的拟合[20]。目前，相关技术主要分为以下两类。

（1）与传统编码器结合的混合式编码器。这类编码器通常保留传统语音编码器的核心模块（如线性预测、变换编码等），用于处理语音信号的结构化特征，同时引入神经网络模型来优化传统方法中的关键环节（如系数量化、后处理增强等）。例如，LPCNet以传统的线性预测编码（linear prediction coding，LPC）[21]模拟声道滤波器，预测语音样本的大致轮廓，并引入循环神经网络

（recurrent neural network，RNN）来生成语音样本。此外，三星研究院提出的X-Net架构[22]也属于混合式编码，其核心设计是在发送端与接收端分别部署Scale-down与Scale-up模块，发送端通过Scale-down模块对高带宽语音进行下采样，提供低带宽输入供后续编码；编码后的数据经信道传输至接收端后，Scale-up模块对解码输出的低带宽语音进行上采样，恢复为原始高带宽语音输出，从而在带宽受限的环境下实现了高质量的语音传输。该方案兼顾了传统方法的稳定性与深度学习在建模复杂特征、提升重建质量方面的优势，从而实现了在有限码率下的高性能语音编码。然而，混合式编码方法在码率支持上仍相对有限，且通常仅适用于较低的语音采样频率。

（2）端到端的神经网络语音编码器。端到端的神经网络语音编码器已成为语音编码领域主要的研究方向。该类编码器摒弃了传统编码器中基于信号处理模块提取特征和依赖人工设计参数的编码思路，采用神经网络编码器—可学习量化器—神经网络解码器的结构。基于神经网络模型的语音编码器的参考模型如图5所示，编码器将原始音频（波形或频谱）映射为高维连续潜变量，量化器将编码器得到的连续特征离散化以实现数据压缩，解码器则从量化索引中重建音频信号。这类编码器均在大量语音数据上进行训练，利用神经网络自动学习语音的声学特征（如频谱包络、基频、韵律），取代了传统基于人工定义的声学模型，并通过将连续潜变量转化为离散码本索引，实现了语音数据的高效压缩与重建。

目前主流的端到端神经网络语音编码器及其性能特点见表2。其中具有代表性的算法如谷歌公司于2021年提出的SoundStream[23]算法，其核心创新在于将残差矢量量化器（residual vector quantizer，RVQ）引入语音编码中，有效解决了传统矢量量化码本规模过大、量化失真的问题。RVQ由多层量化器级联组成，每层仅对前一层的残差误差进行量化，并通过调整量化器数量实现码率的控制。此外，为提高重建音频的感知质量，该算法引入语音合成中的对抗训练思想，并增加两种判别器，用于区分解码生成的音频与真实音频。SoundStream算法已在谷歌发布的Lyra V2编码器中得到应用，可支持3.2 kbit/s、6 kbit/s和9.2 kbit/s这3种不同码率，并提供相应的语音质量。Lyra V2将端到端时延从100 ms降至20 ms。在Pixel 6 Pro智能手机上的实测结果为：该编解码器可在0.57 ms内完成对20 ms语音样本的编码与解码，其处理速度是实时流传输所需速率的35倍。

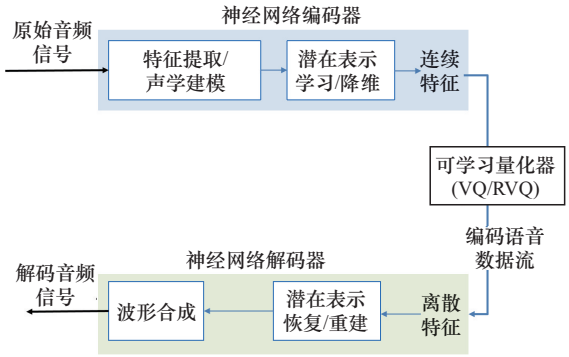

图5　基于神经网络模型的语音编码器的参考模型

另一项代表性算法是Facebook于2022年提出的EnCodec[24]，它采用流式编解码架构，结合RVQ与一种新颖的基于多尺度短时傅里叶变换（multi-scale short-time Fourier transform，MS-STFT）的判别器进行对抗训练。该算法引入了损失平衡机制以提升训练稳定性，并在编码基础上增加了轻量级Transformer模型进熵编码，可在不降低音质的前提下，进一步将带宽压缩25%～40%，即将3 kbit/s模型的带宽降至1.9 kbit/s[24]。性能测试表明，在24 kHz单声道条件下，EnCodec在1.5 kbit/s和12 kbit/s等多个码率上的多激励隐藏参考基准与锚点测试（multi-stimulus test with hidden reference and anchor，MUSHRA）[25]的评分均显著优于Opus、EVS和Lyra V2。在实

表2　主流的端到端神经网络语音编码器及其性能特点

编码器	时间	作者单位	采样频率/kHz	速率/（kbit·s^{-1}）	模型参数量/×10^6
SoundStream[23]	2021.07	谷歌	24	3～18	2.4、8.4
EnCodec[24]	2022.10	Facebook	24、48	1.5、3、6、12、24	15
AudioDec[26]	2023.05	Meta	48	12.8	29
AcademiCodec[27]	2023.05	北京大学&腾讯	16	2、3	64
DAC[28]	2023.06	Descript	16、24	0.5、1.0、1.5…	76
SpeechTokenizer[29]	2023.08	复旦大学	16	4	108
FunCodec[30]	2023.09	阿里	8、16、24	0.25～8、0.5～16	0.52～57.83
X-Codec[31]	2024.08	香港科技大学&微软	16	4	31.45
WavTokenizer[32]	2024.08	浙江大学	16、24、48	0.5、0.9	85
SemantiCodec[33]	2024.05	萨里大学&上海交通大学	16	0.31～1.40	1033
MimiCodec[34]	2024.10	Kyutai	24	1.1	82
FocalCodec[35]	2025.03	康考迪亚大学	16	0.16～0.65	142～145
ALMTokenizer[36]	2025.04	香港中文大学	24	0.41	87～174
XY-Tokenizer[37]	2025.07	复旦大学	24	1	259
LongCat-Audio-Codec[38]	2025.10	美团	16	0.43、0.65、0.87	650

注：关于编码器的语音质量和时延，目前尚未有基于统一测试数据集和标准实验环境所获得的测试数据。

时性方面，24 kHz版本的流式EnCodec算法时延约为13 ms，编解码速度远超实时处理需求，适用于实时通信场景。引入熵编码后，该算法的处理速度略有降低，但仍能满足流媒体等实际应用的需求。

相较于传统语音编码器，基于神经网络模型的语音编码方法在捕捉语音中的高层语义结构方面表现更优。在低码率条件下，当传统基于声道模型的编码器已出现严重失真甚至“失声”时，基于神经网络模型的语音编码器仍能生成清晰、可懂且自然的语音。然而，这类技术目前仍处于探索阶段，在实际部署中仍存在以下几方面的问题。

（1）复杂度高。基于神经网络模型的低速率语音编解码器虽已实现实时处理，但其模型通常涉及数亿次运算，这导致其在手机等移动设备上运行时能耗高，不利于设备续航。因此，需要综合运用模型剪枝、量化，以及设计高效神经网络架构等技术进行模型轻量化，旨在确保编码器核心性能不发生显著下降的前提下将模型复杂度压缩，使其能运行于手机终端芯片。

（2）噪声场景下，鲁棒性与泛化性不足。基于神经网络模型的低速率语音编解码器对噪声的敏感性较高，在应急通信、车载通信等典型的卫星语音通信场景中，电磁杂波、突发冲击等多种类型噪声持续干扰，不仅会遮蔽语音主频带（2～4 kHz）内的关键信息，还会导致重建后的语音可懂度急剧下降。此外，神经网络模型性能依赖于训练数据的质量与覆盖范围。然而，现实通话环境复杂多变，如地域方言、个人口音、环境噪声以及信道干扰等因素，均对编码器的鲁棒性构成挑战。如果模型只在“干净”的实验室语音数据上训练，其在真实复杂场景下的表现将难以保证。因此，应构建大规模、多场景、带噪声的语音数据集，并在训练过程中注重数据的广泛性并持续优化学习方法，以确保编解码器在多样化场景与干扰下仍能保持稳定的语音质量。

（3）迭代更新困难。通信技术是持续演进的，语音编码器也需要不断更新迭代，以适配新需求与新场景。传统语音编码器由模块化算法构

成，局部优化和替换相对容易，而神经网络整个模型高度耦合，一旦部署，对其进行局部升级或替换相对困难。因此，需要探索模块化的神经网络设计，以实现基于神经网络模型的低速率语音编解码器版本更替和升级。

2.3 基于语义的语音编码方法

语义编码器是一种面向未来的新兴通信范式，其原理是不再传输声波的物理细节，而仅编码传输语音中的高层语义单元（如文本、韵律），由接收端进行语音重建。例如，基于非线性变换的语音语义信源信道联合编码系统在相同主客观感知质量下[39]，相较于传统语音编码方案（如AMR-WB或Opus编码器），所需带宽显著降低，且在衰落信道环境下表现出更优越的鲁棒性；北京邮电大学提出的一种基于语义的语音编码（semantic speech codec，SSC）方法[40]，实现了106 bit/s的超低码率语音编码压缩，接近语音信息率的理论上限（约100 bit/s）。SSC系统通过以下3个模块，在极致压缩的同时保持了语音质量：语义向量量化编解码器通过融合频谱和音高特征，并利用混合注意力机制，实现了高效的语义信息提取和压缩；低数据开销声纹编码器捕获时间不变的说话人特征，只需要一次性传输，即可实现个性化语音合成，几乎没有额外的数据开销；条件扩散模型及其相关的扩散损失机制被用来替换传统解码器，通过渐进式的恢复细节，显著增强了合成语音的自然度和真实感。相关实验结果表明，SSC方法在150 bit/s的主观MUSHRA的评分上超越了750 bit/s的EnCodec，验证了其在带宽受限场景下的性能。

目前，基于语义的语音编码的相关研究工作还较少，这种方式理论上能够实现极高的压缩效率，并对信道噪声和丢包具有较强的鲁棒性。然而，语义编码的非保真特性可能会导致说话人身份、语调及情感等个性化信息的丢失，这也是该技术走向实用过程中需要重点解决的问题。

3 性能质量评估方法

评估卫星通信场景下的语音通话质量，既要关注语音信号的保真度与人类听觉体验，也要考虑极端环境下语音的可懂性。同时，还需要兼顾在实际部署中的需求，如实时性、计算复杂度、存储开销及终端功耗等多方面指标。其中，语音编码器的质量评估通常分为主观评价与客观评价两类。主观评价基于听音人对编码后语音的听觉感受进行评分，而客观评价则依据特定算法，量化计算解码信号与原始信号之间的质量差异。

3.1 主观评测方法

根据ITU-T P.800标准，语音质量的主观评价主要采用MOS[9]。该标准不仅定义了MOS的评价框架，还就语音质量主观评测的通用方法和测试环境提出了指导性建议，为各类语音主观测试提供了基础规范。

在进行主观语音评价时，常用方法包括绝对等级评定（absolute category rating，ACR）、损伤等级评定（degradation category rating，DCR）和比较等级评定（comparison category rating，CCR），评分范围为5～1分，依次对应优、良、可接受、较差和很差这5个等级。

MUSHRA方法[25]是ITU-R BS.1534标准中定义的音频主观评价方法，主要用于评估音频系统或处理算法的音质表现。该方法的核心设计是在测试中隐藏参考信号和锚点信号，以模拟真实听觉场景，从而提升评价结果的客观性与可靠性。评分采用百分制，分数越高代表音质越好：80～100分表示音质接近参考信号，无明显瑕疵；50～80分表示存在轻微失真或噪声；0～50分则表示音质显著下降或不可接受。

3.2 客观评测方法

客观语音质量评价方法主要可分为时域、频域和听觉感知域这3类。时域和频域方法依赖信号指标（如信噪比、误码率、频谱失真等）来间

接反映语音质量，但这些参数往往无法准确对应人耳的主观听感。因此，基于心理声学模型通过计算感知失真来评估语音质量的方法逐渐成为研究重点。

感知客观听力质量评估（perceptual objective listening quality analysis，POLQA）[41]与语音质量感知评估（perceptual evaluation of speech quality，PESQ）[42]是两种基于感知模型的客观评价方法，均是ITU标准化的专业语音质量评估工具。PESQ适用于50～7 000 Hz音频范围，支持8 kHz和16 kHz的采样频率，但对经过噪声抑制、回声消除等处理的语音评估效果有限。POLQA覆盖范围更广，支持窄带（50～1 400 Hz）和超宽带（320～3 400 Hz）模式，采样频率可扩展至48 Hz，更适合现代宽带及高码率语音场景。

虚拟语音质量客观评价（virtual speech quality objective listener，ViSQOL）工具[43]是谷歌开源的基于机器学习的语音质量评估工具。它通过比较参考音频与测试音频在频谱及时域特征上的相似性，生成感知语音质量客观评分（mean opinion score-listening quality objective，MOS-LQO），以模拟人耳的感知质量。该工具分别提取参考信号和测试信号中与听觉相关的频谱、时长、响度等特征，再通过机器学习模型计算二者差异，并映射为0～5分的质量评分，分数越高表示语音质量越接近原始信号。

研究表明，现有的客观评测方法对传统编码器效果较好，但在评估基于神经网络模型的语音编码方法时，其客观得分与主观听感之间存在较大差异，仍需要进一步研究。

4　面临的技术挑战

传统基于声道模型的编码方法在低码率下性能已趋近极限，基于神经网络模型的低速率语音编码技术虽表现出明显优势，但在实际部署中仍面临诸多亟待解决的技术挑战，总结如下。

（1）数据集缺失。当前大多数语音编码器的性能评估和模型训练都依赖于较为纯净的实验室语音数据，缺乏覆盖卫星通信典型噪声环境（如电磁干扰、多径衰落、突发冲击噪声等）的大规模、高质量、场景化的语音数据集。

（2）客观评测体系尚不完善。现有语音质量评估指标（如PESQ、POLQA）主要针对传统编码设计，对神经网络生成的语音的评价存在局限，尤其在低码率卫星通信下，亟须建立更符合人耳感知、适用于基于神经网络模型编码器的客观评价体系，并推动相关标准的研究与制定。

（3）算法框架缺乏统一性与兼容性。目前，各类低速率语音编码算法在系统结构、量化方案、训练策略上差异较大，缺乏统一的设计范式，导致系统集成与算法兼容存在障碍。因此，推动模块化、可扩展的编码框架的发展，并实现与传统编码器的向后兼容，是迈向产业化应用的关键。

（4）轻量化与嵌入式部署的挑战。尽管已有部分轻量级网络（如Lyra V2）与模型压缩研究，但在手机终端、物联网设备等资源严格受限的场景中，基于神经网络模型的编码器仍面临计算复杂度高、内存占用大、能耗突出等问题。AI模型轻量化的核心思路是在维持模型性能基本不变的前提下，降低模型的参数量与计算复杂度。目前，主流技术手段包括模型剪枝、量化压缩和知识蒸馏等[44]。

（5）噪声处理问题。当前，大多数语音编码方案将降噪作为独立的语音前处理模块，并依赖不同终端设备内置的降噪能力。然而，基于神经网络模型的语音编码器对环境噪声较为敏感[45]，因此，是否需要在编码器结构中集成专用的降噪模块，成为一个值得探讨的问题。内置降噪模块有助于提升语音可懂度，但不可避免地会增加模型的计算复杂度与系统开销，因此，需要在性能提升与资源消耗之间作出权衡。

（6）标准化与产业化推进不足。目前，低速

率语音编码尚未在3GPP、ITU-T等国际标准组织中形成完整体系，导致产业生态体系仍不完善，不同编码方案之间难以实现有效互联互通，这也在一定程度上制约了技术的规模化应用与推广。当前3GPP已立项极低速率语音编码器标准[18]，旨在推动卫星语音业务基础产业生态体系的构建。

5 结束语

低速率语音编码是突破有限带宽瓶颈、实现卫星语音业务商用的关键技术，未来有望应用于地面网络带宽受限的场景，如大型线上会议等。本文系统梳理并评述了适用于卫星通信的低速率语音编码的技术要求、主流技术路径以及性能评估方法，分析了技术演进方向和研究进展。在后续工作中，面向卫星通信的低速率语音编码重点研究方向包括3个方面：一是构建面向卫星通信场景的语音数据集及客观评测体系；二是突破模型轻量化、降噪处理等工程实现关键技术；三是推动算法框架的统一及相关技术标准的建立。

参考文献：

[1] 杨岭才. 关于快速形成我国天地一体通信运营能力的思考[J]. 电信科学, 2022, 38(4): 1-10.
Yang L C. Thoughts on the rapid formation of China’s space-ground integrated communication operation capability[J]. Telecommunications Science, 2022, 38(4): 1-10.

[2] 陈山枝. 星地融合移动通信系统与关键技术: 从5G NTN到6G的卫星互联网发展[M]. 北京: 人民邮电出版社, 2024.
Chen S Z. Integrated satellite-terrestrial mobile communication systems and key technologies: from 5G NTN to 6G satellite internet development[M]. Beijing: Posts & Telecom Press, 2024.

[3] 李铁骊. 2025年《卫星产业状况报告》发布[J]. 卫星应用, 2025(9): 51-57.
Li T L. The report on satellite industry in 2025 was released[J]. Satellite Application, 2025(9): 51-57.

[4] Rodionov V V. Data services in the Inmarsat communication system[C]//Proceedings of the 3rd International Conference on Satellite Communications. Piscataway: IEEE Press, 2002: 67-70.

[5] 王晓雪, 杨新聪. 天通卫星的技术应用与市场前景分析[J]. 数字通信世界, 2025(8): 193-195, 204.
Wang X X, Yang X C. The technical application and market prospects of Tiantong satellite[J]. Digital Communication World, 2025(8): 193-195, 204.

[6] 3GPP TS 071: 1999 Mandatory speech Codec speech processing functions AMR Speech Codec; General description[S].

[7] 3GPP TS 26.441: 2014 Codec for enhanced voice services (EVS); General overview (Release 12)[S].

[8] 3GPP TR 22.887: 2024 Feasibility study on satellite access. Phase 4 (Release 20)[S].

[9] Streijl R C, Winkler S, Hands D S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives[J]. Multimedia Systems, 2016, 22(2): 213-227.

[10] 周波, 许萌. 数字语音编码技术研究[J]. 科技情报开发与经济, 2008(3): 165-167.
Zhou B, Xu M. Research on digital speech coding technology[J]. Sci-Tech Information Development & Economy, 2008(3): 165-167.

[11] Hiwasaki Y, Ohmuro H. ITU-T G.711.1: extending G.711 to higher-quality wideband speech[J]. IEEE Communications Magazine, 2009, 47(10): 110-116.

[12] 3GPP TS 26.250: 2024 Codec for immersive voice and audio services (IVAS); General overview (Release 18)[S].

[13] RFC 6716: 2012 Definition of the Opus audio Codec[S].

[14] 朱丽, 郭从良. 心理声学模型在数字音频中的应用[J]. 电声技术, 2002, 26(8): 11-14.
Zhu L, Guo C L. Application of psycho-acoustic model in digital audio[J]. Audio Engineering, 2002, 26(8): 11-14.

[15] 赵仁仲. VoIP系统中语音编码算法研究[D]. 成都: 电子科技大学, 2011.
Zhao R Z. Research on voice coding algorithms in VoIP systems[D]. Chengdu: University of Electronic Science and Technology of China, 2011.

[16] Salami R, Laflamme C, Bessette B, et al. ITU-T G.729 Annex A: reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data[J]. IEEE Communications Magazine, 1997, 35(9): 56-63.

[17] Schroeder M, Atal B. Code-excited linear prediction(CELP): high-quality speech at very low bit rates[C]//Proceedings of the ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2003: 937-940.

[18] 3GPP TR 26.940: 2025 File structure for ultra-low bitrate coding (FS_ULBC) (V0.5.0)[S].

[19] Wisayataksin S. An efficient hardware architecture of Codec2

low bit-rate speech decoder[C]//Proceedings of the 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST). Piscataway: IEEE Press, 2019: 1-4.

[20] 王晶, 徐亮, 陈晓娇, 等. 基于神经网络的低码率语音编码技术研究综述[J]. 信号处理, 2024, 40(12): 2261-2280.
Wang J, Xu L, Chen X J, et al. Research review on low bit rate speech coding technology based on neural networks[J]. Journal of Signal Processing, 2024, 40(12): 2261-2280.

[21] Valin J M, Skoglund J. LPCNet: improving neural speech synthesis through linear prediction[C]//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 5891-5895.

[22] Li Y Y, Wang Z Y, Yin L, et al. X-Net: a dual encoding-decoding method in medical image segmentation[J]. The Visual Computer, 2023, 39(6): 2223-2233.

[23] Zeghidour N, Luebs A, Omran A, et al. SoundStream: an end-to-end neural audio Codec[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 495-507.

[24] Défossez A, Copet J, Synnaeve G, et al. High fidelity neural audio compression[PP]. arXiv (2022-10-24)[2026-01-04]. arXiv: arXiv.2210.13438.

[25] ITU-R BS: 2002 Multi stimulus test with hidden reference and anchor (MUSHRA)[S].

[26] Wu Y C, Gebru I D, Marković D, et al. Audiodec: an open-source streaming high-fidelity neural audio codec[C]//Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[27] Yang D C, Liu S X, Huang R J, et al. HiFi-codec: group-residual vector quantization for high fidelity audio codec[PP]. V2. arXiv (2023-05-07)[2026-01-04]. arXiv: arXiv.2305.02765.

[28] Kumar R, Seetharaman P, Luebs A, et al. High-fidelity audio compression with improved rvqgan[J]. Advances in Neural Information Processing Systems, 2023, 36: 27980-27993.

[29] Zhang X, Zhang D, Li S M, et al. SpeechTokenizer: unified speech tokenizer for speech large language models[PP]. V2. arXiv (2024-01-23)[2026-01-04]. arXiv: arXiv.2308.16692.

[30] Du Z H, Zhang S L, Hu K, et al. FunCodec: a fundamental, reproducible and integrable open-source toolkit for neural speech codec[C]//Proceedings of the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 591-595.

[31] Ye Z, Sun P W, Lei J H, et al. Codec does matter: exploring the semantic shortcoming of codec for audio language model[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(24): 25697-25705.

[32] Ji S P, Jiang Z Y, Wang W, et al. WavTokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling[PP]. V3. arXiv (2025-02-25)[2026-01-04]. arXiv: arXiv.2408.16532.

[33] Liu H H, Xu X N, Yuan Y, et al. SemantiCodec: an ultra low bitrate semantic audio codec for general sound[J]. IEEE Journal of Selected Topics in Signal Processing, 2024, 18(8): 1448-1461.

[34] Défossez A, Mazaré L, Orsini M, et al. Moshi: a speech-text foundation model for real-time dialogue[PP]. V2. arXiv (2024-10-02)[2026-01-04]. arXiv: arXiv.2410.00037.

[35] Della Libera L, Paissan F, Subakan C, et al. FocalCodec: low-bitrate speech coding via focal modulation networks[PP]. V2. arXiv (2025-10-24)[2026-01-04]. arXiv: arXiv.2502.04465.

[36] Yang D C, Liu S X, Guo H H, et al. ALMTokenizer: a low-bitrate and semantic-rich audio codec tokenizer for audio language modeling[PP]. arXiv (2025-04-14)[2026-01-04]. arXiv: arXiv.2504.10344.

[37] Gong Y T, Jin L, Deng R F, et al. XY-tokenizer: mitigating the semantic-acoustic conflict in low-bitrate speech codecs[PP]. V2. arXiv (2025-07-09)[2026-01-04]. arXiv: arXiv.2506.23325.

[38] Zhao X H, Xiang H Y, Ye S Z, et al. LongCat-audio-codec: an audio tokenizer and detokenizer solution designed for speech large language models[PP]. arXiv (2025-10-17)[2026-01-04]. arXiv: arXiv.2510.15227.

[39] 张平, 戴金晟, 张育铭, 等. 面向语义通信的非线性变换编码[J]. 通信学报, 2023, 44(4): 1-14.
Zhang P, Dai J S, Zhang Y M, et al. Nonlinear transform coding for semantic communications[J]. Journal on Communications, 2023, 44(4): 1-14.

[40] Jia R J, He Z Q, Niu K, et al. SSC: 106 bit/s ultra-low bitrate semantic speech coding[C]//Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2025: 1-5.

[41] Beerends J G, Schmidmer C, Berger J, et al. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I-temporal alignment[J]. Audio Engineering Society, 2013, 61(6): 366-384.

[42] 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]//Proceedings of the 2001 IEEE International Conference on Acoustics,

Speech, and Signal Processing. Proceedings. Piscataway: IEEE Press, 2002: 749-752.

[43] Hines A, Skoglund J, Kokaram A C, et al. ViSQOL: an objective speech quality model[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2015, 2015(1): 13.

[44] 高杨, 曹仰杰, 段鹏松. 神经网络模型轻量化方法综述[J]. 计算机科学, 2024, 51(增刊1): 11-21.
Gao Y, Cao Y J, Duan P S. Lightweighting methods for neural network models: a review[J]. Computer Science, 2024, 51(S1): 11-21.

[45] Tseng W C, Harwath D. Probing the robustness properties of neural speech codecs[J]. V2. arXiv (2025-05-30)[2026-01-04]. arXiv: arXiv.2505.24248.

[作者简介]

魏晨光（1972–），女，中国移动通信有限公司研究院副院长、高级工程师，主要研究方向为移动媒体技术、5G新通话、新型智能业务、移动信息智库等。

许珈艺（1993–），女，中国移动通信有限公司研究院研究员，主要研究方向为卫星通信、沉浸式媒体。

郭勐（1981–），男，博士，中国移动通信有限公司研究院高级工程师，主要研究方向为视频编码、沉浸式媒体、计算机视觉以及高性能计算。

杨蕾（1983–），女，博士，中国移动通信有限公司研究院业务研究所副所长、高级工程师，主要研究方向为模式识别、图像与视频处理、计算机视觉以及移动通信业务。