# 面向边缘设备的轻量化神经语音压缩方法

鲁 昱[1],付永健[2,3*],丁 典[1],潘 昊[1],薛广涛[1],任 炬[3]
(1. 上海交通大学计算机学院,上海 200240;2. 中南大学计算机学院,湖南长沙 410083;
3. 清华大学计算机科学与技术系,北京 100842)

**摘 要:** 近年来,神经网络驱动的音频压缩方法在低比特率语音重建方面表现出显著优势,但其高计算开销与部署复杂度限制了在边缘设备上的实际应用. 为此,本文面向移动终端等资源受限场景,提出一种轻量化的神经语音压缩系统. 该系统在Funcodec框架基础上,对编码器模块进行优化设计,构建了基于卷积神经网络的简化结构,并引入融合感知对齐、频谱约束和对抗训练的知识蒸馏策略,有效迁移教师模型的表征能力. 实验结果表明,所提出的卷积神经网络编码器在保持压缩质量接近原系统的前提下,大幅降低模型复杂度与推理延迟,可在边缘设备上实现毫秒级音频压缩处理. 进一步地,针对原始量化索引中存在的冗余问题,本文提出基于哈夫曼树的变长编码方法,在不影响重建精度的条件下节省约5%的存储空间,提升系统的传输效率. 综合实验结果表明,所提出方案在压缩质量、计算开销与工程部署可行性之间实现了良好平衡,具备在实际语音采集与感知系统中广泛推广的潜力.

**关键词:** 音频压缩;哈夫曼编码;蒸馏学习;边缘计算
**基金项目:** 国家自然科学基金(No.62432004)
**中图分类号:** TN912.3 **文献标识码:** A **文章编号:** 0372-2112(2025)10-3483-14
**电子学报URL:**http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250524

# A Lightweight Neural Speech Compression Method for Edge Devices

LU Yu[1], FU Yong-jian[2,3*], DING Dian[1], PAN Hao[1], XUE Guang-tao[1], REN Ju[3]
(1. *School of Computer Science*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*;
2. *School of Computer Science and Engineering*, *Central South University*, *Changsha*, *Hunan* 410083, *China*;
3. *Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100842, *China*)

**Abstract:** Neural audio compression methods have shown remarkable performance in low-bitrate speech reconstruction, but their high computational cost and deployment complexity limit their practical use on edge devices. To address this issue, this paper proposes a lightweight neural speech compression system tailored for resource-constrained scenarios such as mobile terminals. Based on the Funcodec framework, we redesign the encoder module using a streamlined convolutional neural network architecture and introduce a multi-objective knowledge distillation strategy that integrates perceptual alignment, spectral constraints and adversarial training. Experimental results demonstrate that the proposed convolutional neural network encoder significantly reduces model complexity and inference latency while maintaining comparable compression quality, enabling millisecond-level real-time speech encoding on edge devices. Furthermore, to improve transmission efficiency, we present a Huffman coding-based entropy optimization method that adaptively encodes residual quantization outputs, achieving an average storage reduction of approximately 5% without compromising reconstruction quality. Overall, the proposed system strikes a favorable balance between compression fidelity, computational efficiency and deployability, making it well-suited for real-world speech acquisition and processing applications on edge platforms.

**Key words:** audio compression; Huffman coding; knowledge distillation; edge computing
**Foundation Item(s):** National Natural Science Foundation of China (No.62432004)

收稿日期:2025-06-19;录用日期:2025-09-25;责任编辑:朱梅玉
*通讯作者:付永健

## 1 引言

随着移动互联网的迅猛发展与智能终端的广泛普及,音频信号作为自然直观的人机交互媒介,在语音通信[1~3]、语音识别[4~6]、语音合成[7~9]及语音存储[10~12]等多个领域中得到了广泛应用.在智能语音助手[13]、远程医疗[14]、智慧城市监测[15]、语音行为分析[16]等典型场景中,往往需要边缘设备(如智能手机、可穿戴设备)[17]持续采集大量音频数据,并将其上传至云端服务器以执行自动语音识别(Automatic Speech Recognition,ASR)[4~6]、语音情感识别(Speech Emotion Recognition,SER)[18~20]、说话人识别[21~23]等复杂下游任务.然而,高采样率语音数据通常具有较大的存储与传输开销,尤其在带宽受限和资源有限的边缘计算环境下,直接传输原始语音将导致较高的能耗与通信负担.因此,如何在保证语音重建质量的前提下,实现高效、低延迟、适配边缘设备的语音压缩方案,已成为语音处理与边缘智能系统设计中的关键问题.

传统语音压缩方法主要依赖于手工设计的信号处理编码器,例如MP3(MPEG-1 audio layer III)、AAC(Advanced Audio Coding)[24]、Opus[25]与EVS(Enhanced Voice Services)[26]等标准化方案.这些方法通过频域变换、量化与熵编码等模块对语音信号进行逐层压缩,长期应用于语音通信与音频存储任务.以MP3和AAC为代表的感知编码器在中高比特率下能够维持良好的音质,但在低比特率(如小于16 kbps)下往往存在感知失真和伪影现象.Opus编码器在语音通信领域取得了重要应用,但其在极低比特率和复杂语音场景下仍面临重建质量瓶颈.EVS编码器虽在窄带语音通信标准中展现出较好的鲁棒性,但其硬件适配性和模型灵活性有限.

为突破传统方法在建模能力与端到端优化方面的局限,近年来兴起了一系列基于神经网络的音频编解码方法,如SoundStream[27]、EnCodec[28]、DAC(Descript Audio Codec)[29]与Funcodec[30]等.这些方法借助端到端训练的神经网络模型,能够自动学习潜在语音特征空间,并结合残差量化、多尺度建模与判别器感知损失,有效提升了在低比特率下的语音重建质量.然而,由于这类模型普遍参数规模较大、计算复杂度高,难以直接部署于资源受限的边缘设备上,其实际应用仍面临推理速度慢、功耗高、存储压力大等挑战.因此,设计一种轻量化、高效率、可部署的神经语音压缩框架,成为推动该类技术工程落地的研究重点.

为应对上述挑战,本文提出一种面向边缘设备的轻量化神经语音压缩系统,重点优化编码器模块的模型结构与压缩效率.在原有Funcodec[30]框架基础上,本文设计了一种完全基于卷积神经网络(Convolutional Neural Network,CNN)[31]的编码器,以替代计算复杂度较高的SEANet[32]编码器.考虑到CNN模型在移动端硬件(如SoC、NPU[33])上的良好适配性,该结构可显著降低模型体积与运算开销,并在保持编码精度的前提下实现毫秒级处理延迟.为了有效保留原始编码器的表征能力,本文引入一种基于知识蒸馏[34]的两阶段训练方法:首先通过多种蒸馏监督信号引导CNN编码器对潜在表征进行特征对齐,随后联合优化系统各模块以提升整体重建性能.

此外,为进一步提升压缩率并降低传输成本,本文针对原始定长10比特量化方式存在的冗余问题,提出基于哈夫曼编码(Huffman coding)的熵编码[35]优化策略.该策略通过统计残差矢量量化(Residual Vector Quantization,RVQ)[36]码字的概率分布,自适应构建哈夫曼树,实现变长编码,有效节省约5%的平均存储空间.

在部署架构上,本文采用端-云协同方式:将编码器与量化模块部署于智能手机等边缘设备端,实现实时语音压缩;将解码器置于服务器侧,支持多种语音下游任务处理.在Librispeech(英文)[37]与AISHELL-1(中文)[38]两个公开语音数据集上的实验结果表明,本文方法在语音重建质量、压缩效率与边缘推理延迟等方面均优于现有主流神经压缩方案,具有良好的实际部署潜力与推广价值.

本文的主要贡献如下:(1)本文基于Funcodec框架,设计了结构简洁、计算开销低的纯卷积编码器,有效降低了模型推理复杂度,使其可在移动端实现毫秒级语音压缩,满足实时性和资源受限部署需求;(2)本文引入频谱约束、判别器感知特征对齐与蒸馏等多重损失项,在保持模型紧凑性的同时,增强了学生模型对潜在特征空间的表达能力,提升语音重建质量;(3)针对定长量化索引存在的冗余问题,本文利用残差量化码字的统计分布构建自适应哈夫曼编码器,在不牺牲解码质量的前提下,平均节省约5%的码字存储空间;(4)本文实现了编码器与量化器在边缘设备端的实时部署,解码器在云端支持多种下游语音任务,实验结果表明本方法在语音质量、压缩效率和部署实用性方面均优于现有主流方案.

## 2 相关工作

### 2.1 传统语音压缩方案

在神经网络压缩方法兴起之前,语音与音频压缩主要依赖于基于信号处理和感知模型的传统编解码标准.此类方法通常通过频域变换、量化、熵编码以及人工设计的感知模型来实现压缩与重建[39],广泛应用于语音通信、多媒体传输与音频存储等场景.

早期典型的音频压缩标准包括MP3和AAC[24],其

主要基于离散余弦变换(DCT)[40]与感知编码理论[35],通过剔除人耳不敏感的频率成分来有效降低比特率.在中高比特率(如64~128 kbps)下,该类方法可实现较高的音频保真度;但在低比特率(如低于16 kbps)条件下,常出现明显的失真与压缩伪影,特别是在语音信号高频部分的重建效果显著下降.

在语音通信领域,多种专用语音编码标准得到了广泛应用.例如,Opus编码器[25]融合了线性预测编码(Linear Predictive Coding, LPC)[41]与频域变换编码技术,支持多带宽自适应与低延迟传输,已成为互联网语音与视频通信的主流选择之一;而EVS[26]编码器作为3GPP标准之一,在语音通信网络中进一步提升了编码鲁棒性与语音保真度,特别是在窄带和超宽带(WB/SWB)条件下表现出良好的压缩性能与抗噪能力.

尽管上述方法在过去几十年中支撑了大多数语音与音频压缩任务,其核心设计仍主要依赖专家规则与线性建模,在非线性特征建模、模型灵活性及端到端压缩优化方面存在一定局限.尤其在极低比特率和复杂声学环境下,传统方法难以同时兼顾编码效率与重建质量.随着边缘计算场景下对超低比特率传输、实时处理和多任务协同能力的需求不断增强,传统方法在扩展性、适应性及深度感知能力方面的瓶颈日益凸显,促使近年来基于神经网络的音频压缩方法加速发展,成为音频压缩领域的重要研究方向.

### 2.2　基于神经网络的语音压缩方案

近年来,随着深度学习技术的快速发展,基于神经网络的音频压缩方法已成为语音与音频编码领域的重要研究方向.相较于传统依赖手工特征设计与信号处理流程的编解码方法,神经网络技术能够通过端到端学习自适应地建模复杂声学空间中的非线性特征与长程依赖关系,尤其在极低比特率条件下展现出更优越的语音重建能力.

SoundStream[27]是Google提出的代表性端到端神经音频压缩模型,采用全卷积式编码器-解码器结构,并引入残差矢量量化(RVQ)[36]机制,以提升量化精度与压缩效率.该方法通过对编码器、RVQ量化器与解码器进行联合训练,在比特率低至3 kbps的条件下即可实现接近传统高比特率编解码器(如Opus)的语音重建效果,同时具备低延迟特性,适用于边缘设备部署.

Meta提出的EnCodec[28]框架进一步扩展了神经音频压缩技术,采用流式卷积网络与分层向量量化策略,并结合多尺度频谱损失与感知对抗训练,有效缓解了音频信号在高频重建阶段易出现的伪影与失真问题.EnCodec支持灵活的比特率控制,在多种采样率和压缩等级下均能提供高保真的语音与音乐还原能力,已在多个实际应用场景中展现出良好的通用性.

Descript提出的DAC[29]在全带宽音频压缩任务中展现出优异的性能.该方法通过优化残差量化机制、引入向量量化嵌入空间正则化与对抗损失设计,在44.1 kHz采样率下可于8 kbps甚至更低比特率下实现高质量音频重建,适用于语音与非语音类音频压缩任务.

为统一上述方法的研究范式,阿里巴巴提出的Funcodec[30]框架集成了多种神经编解码技术,支持包括SoundStream与EnCodec在内的多个压缩架构的训练流程.该系统在引入残差量化、判别器感知损失与特征匹配损失的基础上,采用模块化设计增强了模型的训练稳定性与结构可扩展性,已广泛应用于语音编码、语音合成与识别等任务的前端特征压缩模块中.

尽管神经网络驱动的语音压缩方法在压缩率与重建质量方面取得了显著进展,但当前主流方法普遍存在模型规模大、计算资源消耗高、边缘设备部署困难等问题.特别是在实时性与能效要求较高的边缘计算场景中,如何在保持语音质量的前提下实现模型轻量化与推理高效化,仍是神经语音压缩研究亟待突破的关键挑战.

## 3　系统设计

### 3.1　基于CNN的语音编码器

为满足边缘设备在语音处理任务中对实时性与资源消耗的严格约束,本文设计了一种完全基于CNN[31]的轻量级语音编码器.该编码器采用多层卷积与残差块的堆叠结构,用于高效提取与压缩语音特征,充分发挥CNN在局部建模与计算并行性方面的优势.整体架构在保证压缩性能的同时显著降低了模型计算复杂度,使其具备在终端设备上实现实时语音编码的能力.

如图1所示,该编码器基于一系列卷积操作和残差连接,输入音频信号$s \in \mathbf{R}^{1\times T}$($T$为语音采样点个数,例如对于16 kHz采样的1 s语音,$T=16\,000$)首先经过一维卷积层(初始卷积层),该层采用大小为7的卷积核提取音频的低级特征,如音高和共振峰等.接下来,网络引入核心模块EluResnetBlock,每个EluResnetBlock首先通过两层卷积操作,卷积层间通过ELU激活函数[42]引入非线性特征.每个残差块的核心是通过卷积提取特征,并将其与单个残差卷积层处理的输入信号进行加法运算(即残差连接),从而确保信息流在多层中不会丢失,并有效避免了梯度消失问题.每个EluResnetBlock包含的卷积层采用了不同大小的卷积核(如$3\times3$和$1\times1$),通过这些卷积层逐步提取不同层次的特征.随着网络深度的增加,每个EluResnetBlock输出的通道数逐渐增大,从32个通道逐步提升到64、128和256个通道,这样可以捕获深度的语音特征.在每个块中,使

用1×1卷积作为残差卷积层来优化特征传递，减少计算开销. 随后，本文利用更大步幅的降采样卷积层来进行特征压缩，减小特征的长度维度，通过一系列卷积操作，网络能够输出一个压缩后的低维度潜变量特征表示 $r \in \mathbf{R}^{N\times t_c}$，其中 $N$ 为潜变量表征的特征维度（默认为128），$t_c$ 为潜变量表征的时间维度（默认1 s语音在16 kHz采样率下有 $t_c=50$）.

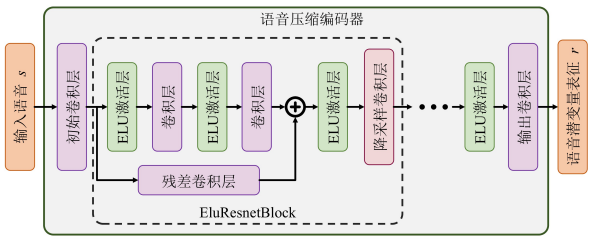



图1 基于CNN的语音编码器架构图

对比Funcodec中SEANet编码器，我们的纯CNN方案以线性复杂度 $O(T)$ 替代LSTM或者自注意力的二次复杂度，显著降低端侧推理延迟；同时卷积算子在PyTorch或是TensorFlow等框架下拥有最成熟的指令级优化，可直接利用cuDNN或oneDNN等实现浮点高速并行，可以在保持音频质量一致的前提下提供更轻量、更高效、更硬件友好的语音编码器.

### 3.2 基于蒸馏学习的两阶段训练框架

为在保证音频压缩质量的前提下实现模型的轻量化与边缘侧可部署性，本文提出一种基于知识蒸馏[34]的两阶段训练架构. 该训练流程如图2所示，旨在充分继承原Funcodec框架中编码器的建模能力，并将其迁移至计算复杂度更低的纯CNN结构中. 具体而言，第一阶段通过蒸馏机制引导学生模型（CNN编码器）学习教师模型（原Funcodec编码器）的潜在特征表征能力，实现特征提取能力的有效迁移；第二阶段将CNN编码器替换至Funcodec框架中，并与残差量化模块和解码器共同进行端到端联合优化，进一步提升各模块之间的协同性能，从而构建出一套压缩率高、推理开销低、可在边缘设备部署的神经语音压缩系统.

#### 3.2.1 编码器蒸馏学习

首先，本文介绍基于蒸馏学习的第一阶段训练过程. 如图2所示，本阶段冻结以Funcodec框架训练好的语音编码器SEANet[32]编码器，残差向量量化模块，语音解码器SEANet解码器，以及多尺度短时傅里叶变换（Short-Time Fourier Transform，STFT）[43]判别器模块，仅对CNN语音编码器进行蒸馏训练，从而使得学生模型CNN语音编码器具有和教师模型SEANet编码器相当的语音编码能力. 具体而言，假如输入语音信号 $s$，通过SEANet编码器和CNN语音编码器进行特征编码，分别得到语音潜变量表征 $r_{\mathrm{SEA}} \in \mathbf{R}^{N\times t_c}$ 和 $r_{\mathrm{CNN}} \in \mathbf{R}^{N\times t_c}$. 为了使学生模型尽可能逼近教师模型在表示空间中的建模能力，该模块引入均方误差（Mean Squared Error，MSE）损失作为蒸馏目标函数，定义如下：

$$L_{\mathrm{distill}} = \left\| r_{\mathrm{SEA}} - r_{\mathrm{CNN}} \right\|_2^2 \tag{1}$$

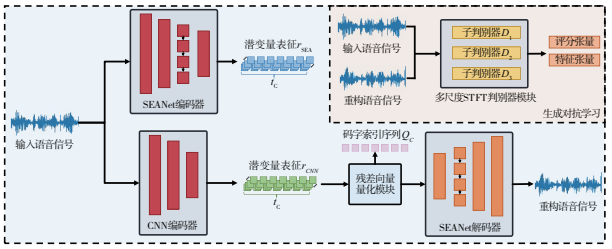



图2 基于蒸馏学习的训练框架

同时，为增强学生编码器对末端语音重建路径的感知能力，本文进一步引入端到端重建路径监督机制. 具体地，借助Funcodec框架中已训练完成的残差矢量量化模块、SEANet解码器以及多尺度STFT判别器模块，对学生模型输出的潜在表示进行感知层级的重建反馈. 在训练过程中，将学生CNN编码器的输出记作 $r_{\mathrm{CNN}}$，并将其作为输入传递至上述冻结模块，生成重建语音信号 $\hat{s}_{\mathrm{CNN}}$，并计算如下损失函数作为附加训练信号.

（1）重构损失. 该损失直接评估学生模型经教师解码路径后的重建效果：

$$L_{\mathrm{recon}} = \left\| \hat{s}_{\mathrm{CNN}} - s \right\|_1 \tag{2}$$

（2）多尺度窗口梅尔谱和功率谱重建损失. 该损失在不同时间窗口设置下对输入语音信号和重构语音信号进行梅尔滤波变换[44]，以计算不同频率分辨率下重建音频与原始音频在梅尔谱和功率谱上的差异，从而增强模型在感知域（特别是中高频区域）的建模能力：

$$\begin{aligned} L_{\mathrm{mel}} = \sum_i \Big( & \left\| M(\hat{s}_{\mathrm{CNN}};\ \mathrm{win}_i) - M(s;\ \mathrm{win}_i) \right\|_1 \\ & + \left\| M(\hat{s}_{\mathrm{CNN}};\ \mathrm{win}_i) - M(s;\ \mathrm{win}_i) \right\|_2^2 \\ & + \left\| P(\hat{s}_{\mathrm{CNN}};\ \mathrm{win}_i) - P(s;\ \mathrm{win}_i) \right\|_1 \\ & + \left\| P(\hat{s}_{\mathrm{CNN}};\ \mathrm{win}_i) - P(s;\ \mathrm{win}_i) \right\|_2^2 \Big) \end{aligned} \tag{3}$$

其中，$M(\cdot;\ \mathrm{win}_i)$ 和 $P(\cdot;\ \mathrm{win}_i)$ 中的 $\mathrm{win}_i$ 指以 $\mathrm{win}_i$ 为窗口长度，以帧移来进行STFT变换，$M$ 指进行梅尔变换得到梅尔谱，而 $P$ 指功率谱. 默认情况下，$\mathrm{win}_i$ 的长度为2的幂次，有 $\mathrm{win}_i \in [128, 256, 512, 1\,024]$.

（3）对抗损失. 本文直接复用Funcodec中已预训练完成的多尺度STFT判别器作为固定判别网络，用于对学生编码器路径下生成的重建语音进行感知质量评价与训练指导. 如图2所示，该多尺度判别器由多个子判别器组成，每个子网络在不同的STFT参数配置下运

行,具备不同的时间-频率分辨率,从而能够捕捉语音信号在多个尺度下的时频结构特征. 通过引入该判别器提供的多尺度感知反馈,学生编码器在训练过程中能够获得更细粒度的语音质量监督,从而有效提升其重建语音的感知保真度. 本文将由学生编码器路径生成的重建语音信号 $\hat{s}_{\mathrm{CNN}}$ 输入该多尺度判别器模块,判别器输出对应的评分张量,表示其对当前输入是否为真实语音的判断置信度. 在训练过程中,通过最小化对抗损失,即最大化判别器对学生输出判定为“真实”的概率,来优化学生编码器参数. 具体而言,令每个子判别器的输出评分张量逼近常数1,从而引导学生模型生成更符合真实语音分布的输出:

$$L_{\mathrm{adv}}=E_{\hat{s}_{\mathrm{CNN}}}\left[\frac{1}{k}\sum_{i=1}^{k}\max\left(0,1-D_i(\hat{s}_{\mathrm{CNN}})\right)\right] \tag{4}$$

其中,$D_i$ 表示第 $i$ 个子判别器,在多尺度STFT判别器中共存在 $k$ 个子判别器. 该过程的核心目标在于从感知层面对语音的纹理细节与频谱一致性进行约束,弥补传统重构损失(如 $L1$)在主观听感方面的不足,进一步提升语音压缩后在听觉上的自然度与质量.

(4)特征匹配损失. 除了对抗损失本身带来的感知优化效果,为了进一步稳定训练过程、提升学生模型对真实语音分布的拟合能力,本文引入特征匹配损失作为附加监督信号. 该损失项的核心思想是,使学生模型生成的语音在判别器内部的多层中间特征表示与真实语音尽可能一致,从而从判别器的判别过程中获得更细粒度的训练反馈. 具体而言,将原始语音信号 $s$ 与重建语音信号 $\hat{s}_{\mathrm{CNN}}$ 分别输入预训练的多尺度STFT判别器模块. 该模块由 $k$ 个子判别器(默认为3个)组成,每个子判别器设计为具有 $l_i$ 层的多层结构(其中 $i=1,2,\cdots,k$,$l_i$ 默认为6),用于在不同时间–频率分辨率下提取语音的感知表示. 在前向传播过程中,每一层均会输出一个中间特征张量表征该层对语音信号的响应. 本文将每一对输入(即真实语音与重建语音)在各层中对应的特征张量之间的 $L1$ 差作为特征匹配损失,并对所有子判别器和所有层进行加权求和,定义如下:

$$L_{\mathrm{fm}}=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{l_i}\sum_{j=1}^{l_i}\left\| D_i^j(\hat{s}_{\mathrm{CNN}})-D_i^j(s)\right\|_1 \tag{5}$$

其中,$D_i^j(\cdot)$ 表示第 $i$ 个子判别器的第 $j$ 层特征输出. 该损失项的引入可以缓解对抗训练过程中梯度不稳定的问题,并通过强制重建语音在判别器内部表达空间中与真实语音保持一致,进一步增强学生编码器的建模能力与生成音频的结构一致性.

综上所述,损失函数由多个部分组成:

$$L_{\mathrm{stage1}}=\lambda_1 L_{\mathrm{distill}}+\lambda_2 L_{\mathrm{recon}}+\lambda_3 L_{\mathrm{mel}}+\lambda_4 L_{\mathrm{adv}}+\lambda_5 L_{\mathrm{fm}} \tag{6}$$

其中,$\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5$ 为超参数来平衡各项损失,该多维度损失联合策略有效地提升了学生CNN编码器的建模能力,使其在压缩精度、听感质量及模型复杂度之间达成优良平衡,并为下一阶段微调提供了坚实基础.

### 3.2.2　联合优化微调

在完成第一阶段的蒸馏训练后,Funcodec原始架构中的SEANet编码器模块被替换成训练得到的CNN语音编码器,构建出完整的轻量级语音压缩系统. 该系统保留Funcodec中已训练完成的残差向量量化模块(RVQ)与音频解码器(SEANet解码器)结构,同时保持判别器模块(多尺度STFT判别器)不变. 在此阶段,除判别器以外的模块不再被冻结,而是采用较低的学习率对系统各个组成部分进行端到端的联合微调,进一步优化模块间的协同表现与整体重建质量. 本文将系统划分为生成器部分与判别器部分进行分别优化(如图2中的两个部分).

(1)生成器部分. 包括CNN语音编码器、残差向量量化模块(RVQ)以及SEANet解码器,构成完整的音频压缩路径. 本文在第一阶段损失函数的基础上,继续使用重构损失、频域感知损失(如Mel谱损失与功率谱损失)、对抗损失与特征匹配损失,作为主要的训练信号. 此外,针对RVQ模块的训练,本文引入承诺损失 $L_{\mathrm{commit}}$,用于约束编码器输出与量化向量之间的距离,从而提升量化稳定性与压缩精度. 设RVQ模块共包含 $o$ 个子量化器(默认设置为32个),每个子量化器对应的输入向量为 $q_i\in\mathbf{R}^N$,编码输出为 $\hat{q}_i$,则有如下关系:

$$q_{i+1}=q_1-\sum_{j=1}^{i}\hat{q}_j,\hat{q}_{i+1}=\mathrm{VQ}_{i+1} \tag{7}$$

其中,$\mathrm{VQ}_i$ 表示向量量化操作,从第 $i$ 个码本中寻找与输入向量最具有最小误差距离(如欧氏距离或者余弦相似度)的向量. 所有编码输出构成最终RVQ模块输出 $\mathrm{RVQ}(q_1)=\sum_{j=1}^{o}\hat{q}_j$,并且,取每个向量对应的索引为 $\mathrm{id}_1,\mathrm{id}_2,\cdots,\mathrm{id}_o$,则单个向量 $q_1$ 对应的量化索引结果为 $\phi=[\mathrm{id}_1,\mathrm{id}_2,\cdots,\mathrm{id}_o]$,而 $r_{\mathrm{CNN}}=\left[q_1^1,q_1^2,\cdots,q_1^{t_c}\right]$,则通过RVQ模块,可以得到最终的量化索引结果为 $Q_{\mathrm{c}}=[\phi^1,\phi^2,\cdots,\phi^{t_c}]$. 承诺损失被定义如下:

$$L_{\mathrm{commit}}=\sum_{j=1}^{t_c}\left(\begin{array}{c}\left\|q_1^j-\mathrm{RVQ}\left(q_1^j\right)\right\|_1\\+\dfrac{1}{o}\displaystyle\sum_{i=1}^{o}\left\|q_i^j-\mathrm{VQ}_i\left(q_i^j\right)\right\|_1\end{array}\right) \tag{8}$$

该损失项鼓励编码器在特征空间中靠近其对应的量化向量,从而提升整体量化表示的稳定性与收敛性. 因此,此时生成器部分的损失函数如下:

$$L_{\mathrm{g}}=\mu_1 L_{\mathrm{recon}}+\mu_2 L_{\mathrm{mel}}+\mu_3 L_{\mathrm{adv}}+\mu_4 L_{\mathrm{fm}}+\mu_5 L_{\mathrm{commit}} \tag{9}$$

其中,$\mu_1,\mu_2,\mu_3,\mu_4,\mu_5$ 为超参数来平衡各项损失.

(2)判别器部分. 判别器仍采用多尺度STFT判别器,为了提升判别器区分真实语音与模型重建语音的

能力，以更有效地为生成器（即编码器-RVQ-解码器路径）提供频域判别反馈，本文使用Hinge Loss[45]形式的损失函数对判别器进行优化，如下所示：

$$\begin{aligned}L_d=&E_{\hat{s}_{\mathrm{CNN}}}\left[\frac{1}{k}\sum_{i=1}^{k}\max\left(0,1+D_i(\hat{s}_{\mathrm{CNN}})\right)\right]\\&+E_s\left[\frac{1}{k}\sum_{i=1}^{k}\max\left(0,1-D_i(s)\right)\right]\end{aligned}\tag{10}$$

该损失函数分别对真实样本与生成样本计算，使每个子判别器的输出评分张量分别逼近常数1和−1，从而使得判别器具有准确的判断能力.

通过系统级的联合优化与微调训练，进一步激发了轻量化CNN编码器的表征能力，使整个语音压缩系统在保持低编码计算开销的前提下，依然具备优越的语音压缩性能.

### 3.3 哈夫曼编码驱动的码字熵压缩优化

在本文提出的语音压缩系统中，音频信号首先通过轻量化CNN编码器进行特征提取，生成一组潜变量表示$r_{\mathrm{CNN}}$. 随后，这些连续表示被输入至多层残差向量量化模块进行量化. RVQ模块将高维的连续向量逐层分解并量化为离散码字索引序列，最终得到码字索引列表$Q_{\mathrm{c}}$，该序列将作为系统压缩后输出的编码结果传输或存储.

具体而言，对于每一级量化器$\mathrm{VQ}_i$而言，假设其码本大小为$S_{\mathrm{VQ}_i}$，则其输出的索引$\mathrm{id}_i\in\{0,1,\cdots,S_{\mathrm{VQ}_i}-1\}$. 在传统的向量量化压缩框架中，码本中的每一个索引都会被统一地映射为固定长度的二进制编码. 例如，若码本大小为1 024（即$S_{\mathrm{VQ}_i}=2^{10}$），则每个码字索引都需用10 bit进行表示，从而确保能完整覆盖所有可能的码字. 这种定长编码方案虽然实现简单、便于解码，但存在显著的问题，其编码效率完全忽略了码字分布的非均匀性.

在实际语音压缩任务中，输入语音信号通常具有较强的结构性，导致量化后的码字分布呈现显著的不均衡特性. 具体而言，部分码字在多数样本中频繁出现，而另一些码字则极少被触发，呈现出明显的统计偏差. 在此情况下继续采用定长编码将造成信息熵利用率下降，即大量比特被用于表示低频甚至冗余的码字索引，降低整体压缩效率. 为解决上述问题，本文引入基于概率建模的熵编码策略，通过构建数据驱动的变长编码方案，在保证重建质量的前提下进一步压缩码字存储开销，从而有效提升语音编码系统的整体压缩性能.

在RVQ模块中，为了提高编码效率，本文对每一层量化器输出的索引序列进行统计分析，并计算其在整个训练集上的经验概率分布. 具体来说，对于包含$S_{\mathrm{VQ}_i}$个码字的第$i$层量化器的码本，记该层中某个码字$c_i\in\{0,1,\cdots,S_{\mathrm{VQ}_i}-1\}$的出现频率为$p(c_i)$，则可以建立该层索引序列的离散概率模型.

以Librispeech训练集为例，我们将该数据集的音频进行逐个裁剪到定长的音频段，得到音频段集合$\Phi_{\mathrm{Libri}}=\{s_1,s_2,\cdots,s_{\mathrm{NL}}\}$，其中NL为数据集中的音频段个数. 随后，我们对集合中的音频段进行逐个编码，得到他们对应的量化索引结果$\{Q_{\mathrm{c}}^1,Q_{\mathrm{c}}^2,\cdots,Q_{\mathrm{c}}^{\mathrm{NL}}\}$，接下来，我们遍历这些量化索引结果，例如，对于$Q_{\mathrm{c}}^1=[\phi^1,\phi^2,\cdots,\phi^{t_c}]$，我们遍历每个$\phi^i=[\mathrm{id}_1,\mathrm{id}_2,\cdots,\mathrm{id}_o]$，则对于码本$1,2,\cdots,o$，我们依次对值为$\mathrm{id}_1,\mathrm{id}_2,\cdots,\mathrm{id}_o$的码字的出现频率加1. 最终我们得到每个码本中每个码字的出现频数$\mathrm{Num}_i$，随后，对于每一个码本，我们计算每个码字的出现频率：

$$p(c_i)=\frac{\mathrm{Num}_i}{\sum_{i=1}^{o}\mathrm{Num}_i}\tag{11}$$

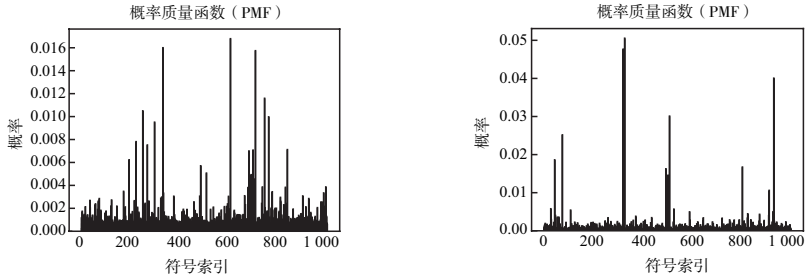
对于每个码本，我们可以以$x$轴为码字的符号索引（即0~1 023），$y$轴为该码字的出现频率，画出每个码本的概率质量函数. 图3展示了对各音频样本的编码结果$Q_{\mathrm{c}}$进行统计后的结果，分别显示了前四个码本的码字索引的概率分布（图3($a$)~($d$)）. 通过计算这些索引的熵值，可以得到对每个码字的理论期望码长$l_{\mathrm{b}}$，该值表示了在给定概率分布$p$下编码单个索引所需的最小比特数. 具体地，理论期望码长$l_{\mathrm{b}}$由式(12)给出：

$$l_{\mathrm{b}}=E_{c\sim p}\left[-\log_2 p(c)\right]\tag{12}$$

以码本1的概率质量函数为例，通过计算可以得到$l_{\mathrm{b}}=9.45$ bit. 这表明，相较于当前采用固定长度的10 bit编码方式，采用熵编码方法可将编码长度减少约0.55 bit，节省约5%的编码空间. 例如，对于单个码本，假设其包含50个码字，且1 s的音频信号经过编码后，理论编码长度可以从500 bit降至472.5 bit.

基于上述计算得到的概率模型，本文进一步采用哈夫曼编码对每个码字索引进行变长编码. 哈夫曼编码是一种无损的熵编码方法，它通过构建最优前缀码，使得频繁出现的码字分配较短的编码，而较少出现的码字则分配较长的编码，从而提高编码效率. 对于每一层量化器，根据其索引的概率分布构建相应的哈夫曼树，并将原始的定长索引序列$Q_{\mathrm{c}}$映射为不定长的哈夫曼编码序列. 由于该过程是基于概率模型和哈夫曼树的共同优化，解码时只需要共享相同的码本和哈夫曼树即可实现无损还原，确保语音重建的精度不受影响.

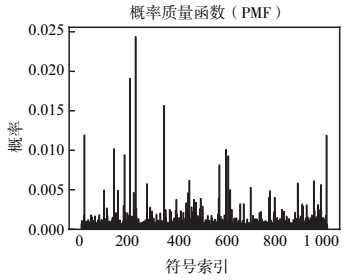
基于上述概率分布建模，本文将RVQ模块的量化输出由固定比特率编码转换为基于数据统计特性的自适应熵编码形式，从而显著提升了编码效率，有效降低



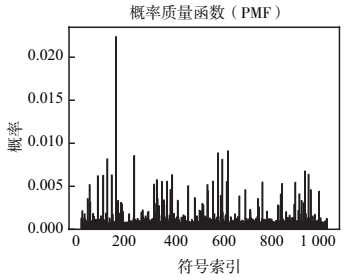

(*a*) 码本1的概率质量函数




(*b*) 码本2的概率质量函数




(*c*) 码本3的概率质量函数




(*d*) 码本4的概率质量函数

图3　不同码本中索引的概率质量函数

了语音压缩过程中的存储开销与传输带宽需求.

### 3.4　系统部署

在所提出的系统架构中(如图4所示),边缘设备与服务器端协同工作,共同完成语音信号的高效压缩与解码任务. 边缘侧主要承担实时语音压缩任务,其中压缩模块包括轻量化CNN编码器与残差矢量量化(RVQ)模块,负责对采集到的音频数据进行快速压缩,从而显著降低存储与传输成本. 在部署流程上,模型训练完成后,首先将模型导出为ONNX格式[46],并通过Android Studio平台[47]将其部署至智能手机终端. 针对移动设备的资源约束,模型在部署前进行了量化优化,有效降低了存储占用与推理延迟. 考虑到智能终端在存储容量与计算资源方面的局限性,所提出方法通过将音频压缩至较低比特率,在确保语音重建质量的同时大幅降低本地数据存储需求,使得多个音频文件可在设备端高效存取与处理,进一步提升了系统整体运行效率与边缘设备的实用性.


图4　系统部署架构

另一方面,压缩后的音频数据通过网络传输到服务器端,服务器上的解码器负责将压缩后的音频数据进行解压重建. 由于服务器端拥有更强大的计算资源,解码器能够高效地解码压缩数据并恢复出高质量的音频信号. 解码后的音频信号可用于执行各种下游任务,例如语音识别、情感分析、语音合成等.

这种分布式的架构充分利用了边缘设备的计算能力进行数据预处理和压缩,减少了数据传输的负担,同时将解码和复杂的下游任务处理集中在计算能力更强的服务器上,确保系统的高效性和实时性. 这种方式不仅提高了存储利用率,还能够适应实时语音处理的需求,尤其适用于需要低延迟、高效存储的应用场景,如语音交互、实时翻译和语音监控等.

## 4　系统评估

### 4.1　实验设置

#### 4.1.1　硬件平台

模型训练与推理测试均在高性能计算与移动终端两类硬件平台上完成. 训练环节采用配备NVIDIA A800 80 GB GPU的服务器,该平台具备充裕显存与计算能力,可高效处理大规模语料并显著缩短深度模型的训练与调优周期. 在部署与推理阶段,为评估编码压缩模块在边缘设备上的实际性能,本文选取七款主流智能手机:Huawei Mate 40、Huawei Nova 6 SE、Huawei Nova 12、Google Pixel 6 Pro、Honor X40 GT、Vivo X80以及Xiaomi 11 Pro. 通过在各设备上部署经训练的编码器模块,对实时推理延迟、峰值内存占用和能耗等关键指标进行系统测试,以全面验证所提方法在多种实际终端场景下的可部署性与性能稳定性.

#### 4.1.2 训练细节

在本实验中，本文采用Pytorch[48]框架下的Adam[49]优化器进行训练. 在蒸馏学习阶段，本文仅对CNN编码器进行训练，冻结其他网络组件的参数. 学习率设置为$3\times10^{-4}$，并训练10个周期. 所有输入语音信号均进行重采样至16 kHz，并裁剪为固定长度40 960（即2.56 s），确保每个样本具有一致的时长. 在微调训练阶段，本文对所有模型组件进行联合训练，学习率降低至$3\times10^{-5}$，并训练20个周期. 此阶段的目标是进一步优化各个模块的协同效果，提升整体压缩性能与语音质量. 训练时，批次大小设定为32，所有语音信号的输入前均进行重采样与裁剪.

在本研究的训练流程中，我们先执行蒸馏学习（式(6)），再进行联合优化微调（式(9)）. 为使各损失项在反向传播时的梯度贡献处于同一数量级，我们在100个小批量上统计每项损失的平均梯度范数，并按梯度对齐策略对权重进行归一化. 蒸馏阶段重点突出教师-学生对齐，固定$\lambda_1=1.0$，为其他损失参数乘以0.1的参数权重，并根据梯度比例得到最终的$\lambda_2=0.1$，$\lambda_3=0.1$，$\lambda_4=0.011$，$\lambda_5=1.111$；在微调阶段沿用相同原则，得到$\mu_1=1.0$，$\mu_2=1.0$，$\mu_3=0.11$，$\mu_4=11.11$，$\mu_5=1.0$. 验证集结果表明，该权重配置能够兼顾模型收敛速度与音质、感知两方面指标的平衡.

#### 4.1.3 数据集与指标

为评估所提出语音压缩系统的整体性能，本文选取两个常用的公开语音数据集进行实验验证. 其中，英文语音部分采用Librispeech[37]数据集，中文语音部分则选用AISHELL-1[38]数据集，分别代表不同语种与语音场景下的编码需求，以全面测试系统在多语言环境下的压缩效果与通用性.

Librispeech是一个广泛使用的英语语音识别数据集，专为语音识别任务而设计. 该数据集由OpenSLR提供，包含约1 000 h的英语语音数据，涵盖了多种口音和噪声环境，适用于训练和评估自动语音识别（ASR）系统. Librispeech数据集的音频来自公开的有声书和录音，经过精确的转录，并按照清晰度分为多个子集（如清晰、干净、背景噪声等）. 我们使用其train-clean-100子集作为训练集（包括28 539条录音数据），并在其中划分出十分之一的数据作为验证集，同时，使用其test-clean子集作为测试集，包括2 620条录音数据. 其标准化的结构使其成为语音压缩与语音识别任务中广泛使用的基准数据集. Librispeech数据集的广泛应用和开放性使得它成为许多语音处理系统的基准，特别是在语音压缩和音频生成的评估中具有重要价值.

AISHELL-1是一个中文语音识别数据集，专为中文语音识别与处理任务而设计. 该数据集包含178 h的中文普通话语音，涵盖了多种口音和语音变种，适合于语音识别、语音合成及语音压缩等任务. AISHELL-1数据集中的语音样本来源于真实对话场景，并经过专业的转录和标注. 数据集的训练集包含120 418条音频数据，而验证集和测试集分别包含14 331和7 176条音频数据. AISHELL-1数据集的特点是其高质量的标注和对中文普通话语音的充分覆盖，使其在中文语音处理任务中成为一个重要的基准数据集. 该数据集广泛应用于中文语音识别技术的研发和评估，也为中文语音压缩技术提供了有价值的测试素材.

此外，为全面评估所提出语音压缩系统在音频重建质量方面的性能，本文选取两种常用的语音客观质量评价指标：PESQ（Perceptual Evaluation of Speech Quality）[50]和ViSQoL（Virtual Speech Quality objective Listener）[51].

PESQ是一种广泛应用于语音质量评估的主观指标，其设计旨在模仿人耳的听觉感知特性. PESQ的评分范围通常从−0.5~4.5，4.5表示最佳的语音质量，−0.5表示最差. 该指标主要用于评估语音通信系统（如语音编码、压缩和传输系统）的重建质量. PESQ基于主观听感质量，通过与人工听感评分的相关性来计算语音的失真程度，能够有效地反映出压缩后语音的可听度和自然性. 其广泛应用于各类语音编码和压缩算法的性能评估中，是评估语音质量的重要标准.

ViSQoL是一种新型的客观语音质量评估模型，旨在模拟人类听觉系统对语音质量的感知. 与PESQ不同，ViSQoL采用基于神经网格相似性指数度量（Neurogram Similarity Index Measure，NSIM）的时频域相似性度量方法，能够更好地处理语音信号中的时延、抖动和频率漂移等问题，特别适用于语音通信中的网络退化情况. ViSQoL的评分范围通常为1~5，数值越高表示语音质量越好. 研究表明，ViSQoL在评估语音质量时，比PESQ更能准确反映人类听感，尤其在VoIP（Voice over IP）等网络环境下的语音质量评估中表现优越.

#### 4.1.4 基准方法

在本实验中，本文选取四种具有代表性的主流神经音频压缩方法作为对比基线，包括Funcodec[30]、DAC[29]、EnCodec[28]和SoundStream[27]. 上述方法在语音压缩领域均具有较高的影响力与广泛的应用基础，分别采用不同的网络结构与优化策略，在多个比特率设置下展现出优秀的语音重建性能. 通过与这些方法的系统性能对比，能够全面验证本文所提方法在压缩质量与计算效率方面的综合优势.

（1）Funcodec是一款由阿里巴巴达摩院提出的神经语音编解码工具包，旨在为语音压缩和生成任务提供高质量、可复现且易于集成的模型架构. 该工具包扩

展了开源语音处理工具包FunASR,提供了包括SoundStream和EnCodec在内的最新神经语音编解码模型的训练和推理脚本. Funcodec的设计强调模块化和可扩展性,支持多种下游任务,如语音识别、个性化语音合成等. 实验结果表明,在相同压缩比下,Funcodec在语音重建质量上优于其他工具包和已发布的模型.

(2)DAC,即DescriptAudioCodec,是一款高保真通用神经音频压缩模型,能够以8 kbps的比特率压缩44.1 kHz的音频,约为MP3[24]的90倍压缩率,同时保持出色的音频质量. DAC结合了图像领域改进的矢量量化技术、对抗性训练和重构损失,适用于语音、环境声音、音乐等多种音频类型. 该模型的优势在于其通用性和高效性,适用于带宽受限的应用场景,如实时通信和音频流媒体.

(3)EnCodec是Meta提出的高保真神经音频压缩模型,采用流式编码器-解码器架构,并在量化潜空间中进行训练. 该模型通过引入多尺度频谱对抗训练和损失平衡机制,显著减少了音频重建中的伪影,提升了音频质量. EnCodec支持多种带宽和采样率的音频压缩,适用于语音、音乐等多种音频类型. 实验结果表明,EnCodec在多个带宽下的音频质量均优于传统编码器,如Opus[25]和MP3.

(4)SoundStream是Google提出的端到端神经音频编解码器,能够高效压缩语音、音乐和一般音频,在语音定制编解码器通常针对的比特率范围内运行. SoundStream采用完全卷积的编码器/解码器网络和残差矢量量化器,联合训练以实现高质量的音频重建. 该模型支持在低延迟下进行实时推理,适用于智能手机等边缘设备. 实验结果表明,SoundStream在3 kbps的比特率下,音频质量超过了12 kbps的Opus编码器,并接近9.6 kbps的EVS编码器[26].

### 4.2　基准实验

为验证所提出语音压缩系统在压缩效率方面的优势,本文设计了基准对比实验,分别在英文语音数据集Librispeech与中文语音数据集AISHELL-1上进行测试. 实验主要对比本文方法与四种主流神经音频压缩方法(Funcodec、DAC、EnCodec和SoundStream)在不同压缩比率下的性能表现. 压缩效率的核心评估指标包括编码后码字长度(即压缩比)以及对应的语音重建质量,后者采用PESQ与ViSQoL两项客观评估指标进行量化分析.

在Librispeech数据集上的实验结果如图5(*a*)和图5(*b*)所示. 从PESQ和ViSQoL两项语音质量评估指标可以看出,所提方法在不同比特率下均展现出良好的语音重建性能. 在PESQ评估结果中,随着比特率的增加,各方法的评分整体呈现递增趋势. 其中,Funcodec始终保持了较高的语音质量,在所有比特率下均处于领先位置;DAC、SoundStream和EnCodec的PESQ分数整体偏低,尤其在低比特率段(小于6 kbps)重建质量受限较为明显. 相比之下,本文提出的方法已明显优于SoundStream、DAC和EnCodec,并接近于Funcodec的性能,显示出良好的压缩适应性与编码效率. 在ViSQoL评估结果中,对于基准方法而言,Funcodec仍保持最优性能,而本文方法在全比特率段上均优于SoundStream、EnCodec和DAC,说明本文方法在感知音质和频谱一致性方面具有更好的鲁棒性. 此外,本文方法与Funcodec已基本持平,验证了所提系统在高效压缩的同时,仍能保持优异的语音重建质量. 整体而言,本文方法在低比特率下的编码长度压缩优势与在语音质量上的平衡能力得到有效验证,展现出良好的综合压缩性能.

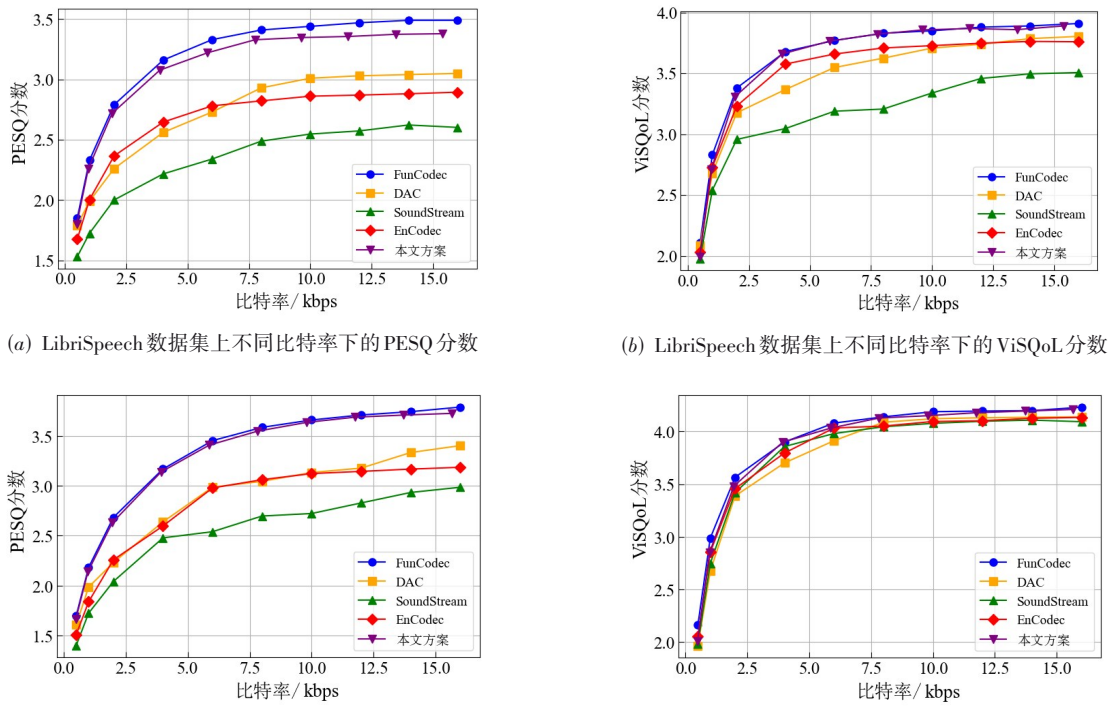
在AISHELL-1中文语音数据集上的测试结果如图5(*c*)和图5(*d*)所示. 从PESQ和ViSQoL两项指标来看,整体趋势与Librispeech上保持一致,但在不同比特率下,各方法之间的性能差异有所体现. 在PESQ评估结果中,本文方法在各比特率段均表现出明显优势. 相比之下,Funcodec表现良好,SoundStream则整体评分偏低,在各个比特率下均低于其他方法. 在ViSQoL评估结果中,所有方法在高比特率下均取得了接近饱和的评分(约4.0分),整体差距较小,说明各方法在高比特率下均能较好地恢复感知一致性. 然而在低比特率(如2 kbps以下)时,仍能观察到本文方法与SoundStream、DAC相比在感知质量上的轻微优势,表明其在极低码率下具备更好的鲁棒性与频谱一致性能力. 整体来看,在AISHELL-1中文语音数据集上的测试结果进一步验证了所提方案在多语言、多场景下的通用性与扩展性.

为了进一步从模型结构的角度验证本文所提出编码器的有效性,我们对比了不同主流神经音频编码器在参数量、浮点运算次数(Floating Point Operations, FLOPs)以及乘加运算次数(Multiply-Accumulate Operations, MACs)上的差异,结果如表1所示. 可以看出,本文方案在保持较低计算复杂度的同时,参数量仅为3.22 M,显著低于Funcodec(7.42 M)、SoundStream(4.79 M)、DAC(21.51 M)等代表性方案,同时在FLOPs和MACs上也具有最小的计算开销. 这表明我们所设计的轻量化编码器在实现低能耗与低延迟的同时,也具有参数效率上的明显优势,进一步印证了其在资源受限设备上的应用潜力.

### 4.3　移动设备部署表现

#### 4.3.1　智能手机性能基准测试

为评估不同智能终端在实时推理任务中的性能差异,本文基于安兔兔评估平台[52]对六款市售主流智能手机进行了硬件性能测试. 测试涵盖四个关键指标,包




(*a*) LibriSpeech数据集上不同比特率下的PESQ分数 (*b*) LibriSpeech数据集上不同比特率下的ViSQoL分数

(*c*) AISHELL-1数据集上不同比特率下的PESQ分数 (*d*) AISHELL-1数据集上不同比特率下的ViSQoL分数

图5 中英文数据集上各种音频压缩算法在不同比特率下的PESQ和ViSQoL分数

**表1 不同编码器方案参数量和浮点计算次数对比**

| 编码器方案 | 浮点运算次数/G | 乘加运算次数/G | 参数量/M |
|---|---|---|---|
| Funcodec | 1.99 | 0.99 | 7.42 |
| DAC | 24.55 | 12.28 | 21.51 |
| SoundStream | 12.16 | 6.08 | 4.79 |
| Encodec | 1.99 | 0.99 | 7.42 |
| 本文方案 | 1.57 | 0.78 | 3.22 |

括中央处理器性能(Central Processing Unit,CPU)、图形处理器性能(Graphics Processing Unit,GPU)、内存性能(MEMory performance,MEM)以及用户体验(User eXperience,UX). 其中,CPU指标反映设备在执行计算密集型任务中的处理能力,GPU性能主要衡量其在图形与音视频渲染方面的表现,MEM指标评估设备的内存读写速率与多任务处理能力,而UX指标则综合体现设备在真实使用场景中的响应速度与系统流畅性. 安兔兔平台对上述指标提供统一的测试标准,并以评分形式输出各项性能结果,得分越高表示性能越强. 通过对这些指标的量化分析,本文可在统一评估体系下对比不同硬件配置对语音压缩编码器推理效率的影响,为实际部署提供依据.

如表2中的结果显示,Vivo X80在所有设备中表现最为出色,特别是在CPU、GPU和UX上的高得分,表明其在处理计算密集型任务和图形渲染时的优势. 而Huawei Mate 40 Pro和Xiaomi 11 Pro紧随其后,虽然它们在CPU和GPU的性能上具有竞争力,但相较于Vivo X80,其MEM和UX得分较低,影响了整体性能. 相对较弱的Huawei Nova 6 SE和Google Pixel 6 Pro在GPU和MEM上的得分较低,限制了其在压缩和解压任务中的处理能力. 这些性能差异将直接影响设备在运行压缩编码器和执行实时推理时的速度和效率,尤其是在存储和图形处理的要求较高的应用场景中.

**表2 实机测试智能手机的硬件性能基准测试结果**

| 智能手机 | CPU↑ | GPU↑ | MEM↑ | UX↑ | SUM↑ |
|---|---|---|---|---|---|
| Huawei Mate 40 Pro | 209 489 | 238 773 | 171 636 | 160 574 | 780 472 |
| Huawei Nova 6 SE | 135 102 | 91 791 | 99 468 | 98 715 | 425 076 |
| Huawei Nova 12 | 201 788 | 138 934 | 140 850 | 135 491 | 617 063 |
| Google Pixel 6 Pro | 183 604 | 174 109 | 148 578 | 169 700 | 675 991 |
| Honor X40 GT | 192 461 | 216 011 | 144 437 | 184 854 | 737 763 |
| Vivo X80 | 368 452 | 361 166 | 222 312 | 259 556 | 1 211 486 |
| Xiaomi 11 Pro | 238 110 | 218 164 | 182 153 | 230 205 | 868 632 |

### 4.3.2 智能手机部署评估

为验证所提出轻量化CNN编码器在实际移动终端部署场景下的性能优势,本文在所选智能手机上分别部署了本文设计的CNN编码器与Funcodec框架中的原始SEANet编码器,并系统性评估两者在语音压缩任务中的端侧运行性能. 实验从压缩延迟、内存占用与能耗开销这三个维度对模型进行性能分析,重点考察不同

编码器在移动设备上实际运行时的计算资源消耗及部署友好性,为模型在资源受限环境中的可用性提供量化依据.

考虑到单次处理1 s音频所产生的计算延迟与能量消耗较小,难以形成稳定的测量基准,本文采用长序列批量测试策略以获得更具代表性和可量化的评估结果.具体地,在每台测试设备上持续压缩总时长为10 000 s的语音数据,记录整个压缩过程的总运行时延与总能量消耗.随后,通过总值与语音总时长的均值计算,得到每秒音频的平均压缩延迟与单位能量开销.内存占用则通过系统监测工具在模型运行过程中实时采样,并统计整个测试期间的峰值使用量,以反映模型在移动端执行过程中的内存资源占用上限.

实验在前述七款主流智能手机设备上(1:Huawei Mate 40 Pro,2:Huawei Nova 6 SE,3:Huawei Nova 12,4:Google Pixel 6 Pro,5:Honor X40 GT,6:Vivo X80,7:Xiaomi 11 Pro)进行部署测试,涵盖了不同硬件平台与计算能力.测试过程中,两组模型使用完全一致的输入数据与推理配置,均以16 kHz采样率、1 s长度的音频片段作为压缩输入,确保不同模型间的测试结果具有直接可比性.

各设备在实际部署下的编码器运行时延、内存占用与能量开销如图6~图8所示.总体来看,所提出的CNN编码器在三项指标上均表现出明显优于Funcodec的SEANet编码器的性能优势,充分验证了本文设计在端侧实时部署中的轻量化效果.

从图6所示的时延结果来看,在所有设备上,CNN编码器的平均压缩时延均明显低于SEANet编码器.其中在部分高性能设备上(如设备3、4、6、7),CNN编码器基本能够将单秒音频的压缩时延控制在30 ms范围内,而SEANet编码器在相同设备下时延普遍在40 ms以上.特别是在计算能力相对有限的设备(如设备2),两者差距更加显著,CNN编码器的时延接近SEANet编码器的一半.这表明CNN编码器能够有效适配不同硬件条件下的实时计算需求.

在内存占用方面(图7),CNN编码器同样展现出良好的资源控制能力.大多数设备上,CNN编码器内存使用量在100~200 MB范围内,相比SEANet编码器平均降低了15%~30%的内存占用.其中在设备1、3、6等设备中,CNN编码器的内存消耗优势尤为明显,进一步提升了模型在中低端移动终端上的可部署性与运行稳定性.

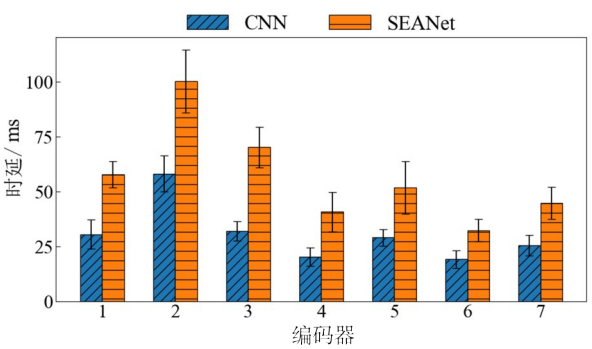


图6　不同编码器在智能手机上压缩1 s音频的编码时延

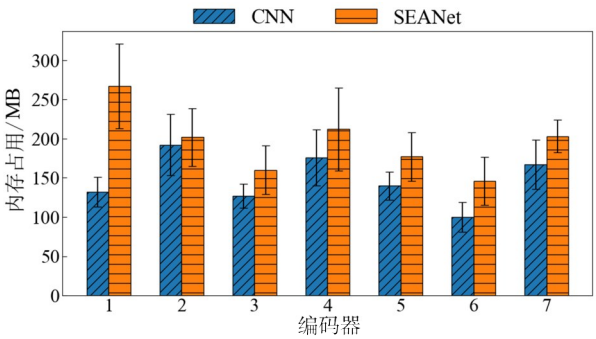


图7　不同编码器在智能手机上压缩1 s音频的内存占用

在能量开销方面(图8),CNN编码器在所有设备上的单位音频能耗均显著低于SEANet编码器.在不同硬件平台上,CNN编码器平均节省约40%~60%的能量消耗,有效延长了移动设备在长时序语音采集与压缩任务中的续航能力.尤其在持续高频运行场景下,能耗优势对于移动终端实际应用具有重要意义.

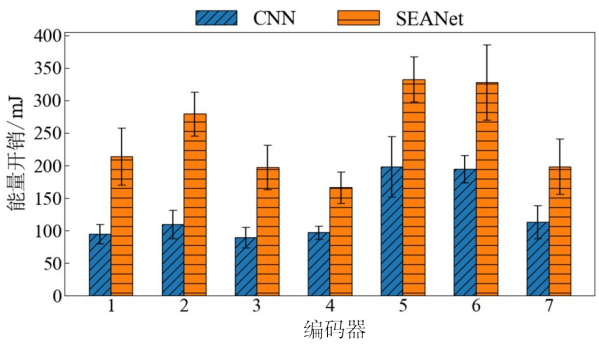


图8　不同编码器在智能手机上压缩1 s音频的能量开销

综上所述,本文提出的CNN编码器在智能手机等边缘设备上具备良好的实时性、资源友好性与能效表现,能够有效支撑高频语音采集任务下的大规模部署需求,并充分缓解了传统复杂模型在端侧实时压缩任务中的计算与能耗瓶颈.

## 5　结论

本文针对当前神经语音压缩方法在边缘设备部署中普遍存在的计算复杂度高、实时性不足等问题,提出了一种轻量高效的神经语音压缩系统.该系统在Funcodec框架基础上,设计了纯卷积结构的CNN编码器,并结合基于蒸馏学习的两阶段训练策略,实现了对原始模型建模能力的有效迁移,同时显著降低了模型复杂度与推理开

销. 为进一步提升压缩效率,本文引入基于码字分布统计的哈夫曼熵编码机制,在保证语音重建质量的前提下,有效减少了存储与传输所需比特数. 在系统部署方面,本文将 CNN 编码器与残差向量量化模块部署至多款主流智能手机端,并开展了系统性实测评估,验证所提方法在压缩延迟、内存占用及能量消耗等关键指标上相较于传统复杂模型具备明显优势. 此外,在 Librispeech 与 AISHELL-1 两个公开语音数据集上进行的实验表明,本文方法在语音质量与压缩比之间实现了良好权衡,整体性能达到或优于现有先进神经压缩模型. 综合而言,本文所提出的轻量化神经语音压缩系统在保持语音重建保真度的同时兼具高压缩率与低资源开销,具备在移动边缘语音采集与智能终端环境中实际部署与推广的应用潜力.

## 参考文献


[1] MENG T, LI W F, YUAN C, et al. AsTree: An audio subscription architecture enabling massive-scale multi-party conferencing[C]//22nd USENIX Symposium on Networked Systems Design and Implementation. Berkeley: USENIX Association, 2025: 653-666.

[2] 张聿晗, 李艳雄, 江钟杰, 等. 基于联合学习框架的音频场景聚类[J]. 电子学报, 2021, 49(10): 2041-2047.
ZHANG Y H, LI Y X, JIANG Z J, et al. Audio scene clustering based on joint learning framework[J]. Acta Electronica Sinica, 2021, 49(10): 2041-2047. (in Chinese)

[3] 白海钏, 鲍长春, 刘鑫. 基于局部最小二乘支持向量机的音频频带扩展方法[J]. 电子学报, 2016, 44(9): 2203-2210.
BAI H C, BAO C C, LIU X. Audio bandwidth extension method based on local least square support vector machine[J]. Acta Electronica Sinica, 2016, 44(9): 2203-2210. (in Chinese)

[4] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ACM, 2023: 28492-28518.

[5] SHA F, SAUL L K. Large margin hidden markov models for automatic speech recognition[M]//Advances in Neural Information Processing Systems 19. Cambridge: The MIT Press, 2007: 1249-1256.

[6] PENG Y F, DALMIA S, LANE I, et al. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding[EB/OL]. (2022-07-06)[2025-08-03]. https://arXiv.org/abs/2207. 02971.

[7] 章子旭, 简志华. 采用双重交换表示分离的任意说话人语音转换[J]. 电子学报, 2024, 52(6): 2141-2150.
ZHANG Z X, JIAN Z H. Any-to-any voice conversion using double exchange representation separation[J]. Acta Electronica Sinica, 2024, 52(6): 2141-2150. (in Chinese)

[8] MIN D C, LEE D B, YANG E, et al. Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation[EB/OL]. (2021-06-16)[2025-08-30]. https://arXiv.org/abs/2106.03153.

[9] LE M, VYAS A, SHI B W, et al. Voicebox: Text-guided multilingual universal speech generation at scale[EB/OL]. (2023- 10-19)[2025-08-30]. https://arXiv.org/abs/2306.15687.

[10] G N B, ANEES M, G T Y. Speech coding techniques and challenges: A comprehensive literature survey[J]. Multimedia Tools and Applications, 2024, 83(10): 29859-29879.

[11] 唐昆, 崔慧娟, 刘志勇, 等. 高质量 4 ~ 8kb/s 变速率有限状态 ACELP 语音编码算法研究[J]. 电子学报, 2000, 28(1): 21-25.
TANG K, CUI H J, LIU Z Y, et al. 4~8kb/s variable rate-finite state-algebraic code excited linear prediction speech coding algorithm[J]. Acta Electronica Sinica, 2000, 28(1): 21-25. (in Chinese)

[12] TAN K, WANG D L. Towards model compression for deep learning based speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1785-1794.

[13] CHENG P, ROEDIG U. Personal voice assistant security and privacy: A survey[J]. Proceedings of the IEEE, 2022, 110(4): 476-507.

[14] CHEN T, YANG Y J, FAN X R, et al. Exploring the feasibility of remote cardiac auscultation using earphones[C]//Proceedings of the 30th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2024: 357-372.

[15] AKRAM M F, WANG S G, ANWAR M R, et al. A comprehensive survey on MEC enabled tactile internet: Applications, challenges, and efficient resource handling[J]. Chinese Journal of Electronics, 2025, 34(5): 1449-1463.

[16] DU J, ZOU X, HAO J, et al. The efficiency of ICA-based representation analysis: Application to speech feature extraction[J]. Chinese Journal of Electronics, 2011, 20(2): 287-292.

[17] CHEN Q L, YE A Y, ZHANG Q, et al. A new edge perturbation mechanism for privacy-preserving data collection in IOT[J]. Chinese Journal of Electronics, 2023, 32(3): 603-612.

[18] HUANG Z W, DONG M, MAO Q R, et al. Speech emotion recognition using CNN[C]//Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM, 2014: 801-804.

[19] LATIF S, RANA R, KHALIFA S, et al. Survey of deep

representation learning for speech emotion recognition[J]. IEEE Transactions on Affective Computing, 2023, 14(2): 1634-1654.

[20] YE J X, WEN X C, WEI Y J, et al. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[21] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset[C]//Interspeech 2017. Singapore: ISCA, 2017: 2616-2620.

[22] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[C]//Interspeech 2018. Singapore: ISCA, 2018: 1086-1090.

[23] LIU T C, LEE K A, WANG Q Q, et al. Disentangling voice and content with self-supervision for speaker recognition[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: ACM, 2023: 50221-50236.

[24] BRANDENBURG K. MP3 and AAC explained[C]//Audio Engineering Society Conference: 17th International Conference on High-Quality Audio Coding. New York: Audio Engineering Society, 1999: 1-12.

[25] VALIN J M, VOS K, TERRIBERRY T. RFC 6716: Definition of the Opus audio codec[EB/OL]. (2012-09-01)[2025-08-20]. https://dl.acm.org/doi/book/10.17487/RFC6716#sec Authors.

[26] DIETZ M, MULTRUS M, EKSLER V, et al. Overview of the EVS codec architecture[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2015: 5698-5702.

[27] 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, 2021, 30: 495-507.

[28] DÉFOSSEZ A, COPET J, SYNNAEVE G, et al. High fidelity neural audio compression[EB/OL]. (2022-10-24)[2025-08-20]. https://arXiv.org/abs/2210.13438.

[29] KUMAR R, SEETHARAMAN P, LUEBS A, et al. High-fidelity audio compression with improved RVQGAN[C]// Proceedingsofthe 37th International Conference on Neural Information Processing Systems. New York: ACM, 2023: 27980-27993.

[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]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 591-595.

[31] O'SHEA K, NASH R. An introduction to convolutional neural networks[EB/OL]. (2015-12-02)[2025-08-20]. https://arXiv.org/abs/1511.08458.

[32] TAGLIASACCHI M, LI Y P, MISIUNAS K, et al. SEANet: A multi-modal speech enhancement network[EB/OL]. (2020-10-01)[2025-08-20]. https://arXiv.org/abs/2009.02095.

[33] JANG J W, LEE S, KIM D, et al. Sparsity-aware and reconfigurable NPU architecture for samsung flagship mobile SoC[C]//Proceedings of the 48th Annual International Symposium on Computer Architecture. New York: ACM, 2021: 15-28.

[34] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[35] VITTER J S. Design and analysis of dynamic Huffman codes[J]. Journal of the ACM, 1987, 34(4): 825-845.

[36] BARNES C F, RIZVI S A, NASRABADI N M. Advances in residual vector quantization: A review[J]. IEEE Transactions on Image Processing, 1996, 5(2): 226-262.

[37] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: An ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2015: 5206-5210.

[38] BU H, DU J Y, NA X Y, et al. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Piscataway: IEEE, 2018: 1-5.

[39] PAN D. A tutorial on MPEG/audio compression[J]. IEEE MultiMedia, 1995, 2(2): 60-74.

[40] AHMED N, NATARAJAN T, RAO K R. Discrete cosine transform[J]. IEEE Transactions on Computers, 1974, C-23(1): 90-93.

[41] LIU P, LI S B, WANG H Q. Steganography integrated into linear predictive coding for low bit-rate speech codec[J]. Multimedia Tools and Applications, 2017, 76(2): 2837-2859.

[42] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (ELUs)[EB/OL]. (2016-02-22)[2025-08-20]. https://arXiv.org/abs/1511.07289.

[43] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform[J]. IEEE Transactions on Acoustics,

Speech, and Signal Processing, 1984, 32(2): 236-243.
[44] STEVENS S S, VOLKMANN J, NEWMAN E B. A scale for the measurement of the psychological magnitude pitch[J]. The Journal of the Acoustical Society of America, 1937, 8(3): 185-190.
[45] GENTILE C, WARMUTH M K. Linear hinge loss and average margin[C]//Proceedings of the 12th International Conference on Neural Information Processing Systems. New York: ACM, 1998: 225-231.
[46] YANG H J, FRITZSCHE M, BARTZ C, et al. BMXNet: An open-source binary neural network implementation based on MXNet[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 1209-1212.
[47] HAGOS T. Learn Android Studio 3 with Kotlin: Efficient Android App Development[M]. Berkeley: Apress, 2018.
[48] IMAMBI S, PRAKASH K B, KANAGACHIDAMBARESAN G R. PyTorch[M]//Programming with TensorFlow. Cham: Springer International Publishing, 2021: 87-104.
[49] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2017-01-30)[2025-08-20]. https://arXiv.org/abs/1412.6980.
[50] 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 code-cs[C]//2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2002: 749-752.
[51] HINES A, SKOGLUND J, KOKARAM A, et al. ViSQOL: The virtual speech quality objective listener[C]//IWAENC 2012 International Workshop on Acoustic Signal Enhancement. VDE, 2012: 1-4.
[52] GUO Y, XU Y N, CHEN X Q. Freeze it if you can: Challenges and future directions in benchmarking smartphone performance[C]//Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications. New York: ACM, 2017: 25-30.

## 作者简介

**鲁 昱** 男,2001年6月出生于四川省简阳市. 现为上海交通大学计算机学院计算机科学与技术专业博士生. 主要研究方向为移动计算与智能感知.
E-mail: yulu01@sjtu.edu.cn

**付永健** 男,1999年2月出生于四川省成都市. 2021年本科毕业于中南大学物联网工程专业. 现为清华大学计算机系访问博士生,中南大学计算机科学与技术专业在读博士生. 主要研究方向为边缘智能.
E-mail: yongjianwork@gmail.com

**丁 典** 男,1994年1月出生于江苏省海安市. 现为上海交通大学计算机学院博士后. 主要研究方向为移动计算与智能感知.
E-mail: dingdian94@sjtu.edu.cn

**潘 昊** 男,1994年11月出生于江苏省高邮市. 2022年毕业于上海交通大学计算机系,现为上海交通大学计算机学院长聘教轨副教授. 主要研究方向为无线通信、无线感知、可穿戴医疗等.
E-mail: panh09@sjtu.edu.cn

**薛广涛** 男,1976年5月出生于江苏省徐州市. 2004年于上海交通大学获计算机软件与理论专业博士学位,现为上海交通大学计算机学院特聘教授. 主要研究方向为物联网技术、智能感知、分布式计算系统、数据流通与治理等.
E-mail: gt_xue@sjtu.edu.cn

**任 炬** 男,1987年12月出生于湖南省汨罗市. 博士,清华大学计算机与技术系长聘副教授. 国家级人才项目获得者. 主要研方向域为边缘智能计算与智能协作、边缘智能安全与隐私保护.
E-mail: renju@tsinghua.edu.cn