音频压缩与解码技术

杨岢瑞

Nov 21, 2026

在日常生活中，人们通过手机听歌、参加视频会议或使用在线音乐平台，都离不开音频压缩技术的支持。原始音频数据往往体积庞大，难以在有限的存储空间和网络带宽中高效传输，因此压缩成为必然的选择。核心问题在于如何在最小失真与最小体积之间找到平衡，既保证听感质量，又尽可能降低数据量。本文将从数字音频信号基础出发，逐步剖析无损与有损压缩的原理与实现，再深入主流编解码器的内部机制，并探讨解码端优化、软硬件协同及新兴趋势。

1　数字音频基础

音频信号的数字化过程始于采样定理，即奈奎斯特-香农采样定理。该定理指出，如果信号的最高频率为 f_{max}，则采样频率必须至少达到 $2f_{max}$，才能无失真地重建原始信号。实际应用中，CD 音频采用 44.1 kHz 采样率，对应最高可重建频率约 22.05 kHz。量化则是将连续幅值映射到离散电平的过程，位深决定了量化精度。16 位量化可提供约 96 dB 的理论动态范围，而 24 位则可达 144 dB。声道数进一步影响数据量：立体声是单声道的两倍，多声道环绕声则呈指数级增长。举例而言，一张标准 CD 以 44.1 kHz、16 位、立体声录制，码率为 $44100 \times 16 \times 2 = 1.4112$ Mbps；若升级至 96 kHz、24 位、5.1 声道，码率将飙升至约 13.8 Mbps，凸显压缩的必要性。

2　无损与有损：两种哲学

无损压缩追求完全可逆的重建，核心在于熵编码与预测编码。熵编码如霍夫曼编码利用符号出现概率分配短码字，算术编码则将整个符号序列映射为一个区间内的实数，从而逼近理论极限。Rice/Golomb 编码在处理拉普拉斯分布的残差信号时尤为高效。预测编码通过线性预测（LPC）或自适应差分（ADPCM）减少相邻样本间的相关性，残差再经熵编码进一步压缩。FLAC、ALAC、WavPack 等格式均采用上述策略，在典型音乐内容上可获得 40% – 60% 的压缩比，且完全可逆。

有损压缩则主动引入可控失真，借助心理声学模型实现更大压缩。掩蔽效应包括频率掩蔽与时间掩蔽：强音可掩蔽临近频率的弱音，以及前掩蔽与后掩蔽。临界频带将人耳听觉范围划分为约 24 个 Bark 尺度子带，变换域编码在此基础上进行。MDCT（改进离散余弦变换）将时域信号映射到频域，提供良好的频谱分辨率与时域混叠抵消特性。噪声成形将量化噪声推向掩蔽阈值以下的区域，比特分配策略则根据掩蔽阈值动态分配比特。MP3、AAC、Opus、Vorbis 等格式均建立在此框架之上。

3 主流有损编解码器深度拆解

MP3（MPEG-1 Layer III）采用混合滤波器组：32 路伪正交镜像滤波器（PQMF）将信号分为 32 个子带，随后每子带再进行 18 点 MDCT，实现 576 线频谱分辨率。心理声学模型计算掩蔽阈值后，比特分配模块利用双缓冲 Huffman 表与比特库机制，在帧内动态分配比特。之所以 128 kbps 被视为及格线，是因为该码率下立体声音乐的平均信噪比与掩蔽裕度仍能满足多数听众的主观需求，但低码率时高频细节易受损。

AAC（MPEG-2/4）在 MP3 基础上引入更精细的心理声学模型与时域噪声整形（TNS）。TNS 在频域对信号进行线性预测，可有效控制前回声。HE-AAC 进一步集成频谱带复制（SBR）与参数立体声（PS）。SBR 通过高频重建算法，将 0 – 8 kHz 的低频内容映射生成 8 – 16 kHz 的高频成分，从而在 48 kbps 码率下仍能提供接近 96 kbps AAC 的主观质量。PS 则通过少量参数描述声像与相位关系，节省大量比特。

Opus（RFC 6716）采用 SILK 与 CELT 两种编码模式自适应切换。SILK 基于线性预测与长期预测，针对语音优化；CELT 基于 MDCT 与感知矢量量化，针对音乐优化。切换决策由实时信号特征与目标延迟共同驱动，可在 2.5 ms 至 60 ms 帧长之间灵活选择。端到端低延迟设计使其成为实时通信的首选。

杜比 AC-4 与 MPEG-H 3D Audio 引入对象音频与场景音频概念。元数据描述声源位置、轨迹与声学环境，解码端根据听音者位置与设备能力进行个性化渲染。高阶 Ambisonics 通过球谐函数分解声场，压缩时仅需传输少量系数即可重建三维声场。

4 无损压缩的工程实现

FLAC 的子帧结构由帧头、子帧头与编码残差组成。子帧头包含声道数、采样率、位深及预测阶数等信息。残差编码采用改进的 Rice 编码，参数由子块统计特性动态选择。编码器首先对整帧进行互相关分析，选择最优预测器；随后对残差进行分区，针对每个分区独立计算 Rice 参数并编码。解码端只需按相同参数解码即可完全重建。多线程编码通过帧级并行与 SIMD 指令集加速。AVX2 可在单指令周期内完成 8 个 32 位整数的并行预测与熵编码。内存带宽成为瓶颈时，可采用分块处理与预取策略。文件容器方面，OGG、Matroska、MP4 均可封装 FLAC 流，并通过 Vorbis Comment、ID3 或 APIC 存储元数据与封面图片。

5 解码端关键技术

解码流程依次包括比特流解析、逆量化、IMDCT/滤波器合成与重叠相加。比特流解析需处理 Huffman 码表、比例因子与全局增益；逆量化将量化索引映射回频域系数；IMDCT 将频域系数变换回时域窗函数；重叠相加消除块间边界效应。定点实现需在 16 位或 32 位定点数与浮点之间权衡，嵌入式平台常采用 24 位定点以兼顾精度与功耗。错误隐藏与丢包补偿（PLC）通过外推相邻帧的 LPC 系数或频谱包络，生成平滑的替代帧，降低实时通信中的听感损伤。

6 软硬件协同：从算法到 SoC

固定功能 DSP 针对特定算法硬连线，功耗最低但灵活性差；可编程 DSP 通过微码或指令集实现多种编解码，平衡功耗与灵活性；神经网络加速器则针对端侧神经音频编解码提供矩阵乘法与激活函数加速。典型移动 SoC 音频子系统包含多核 DSP、共享 SRAM 与低功耗 DMA。功耗约束要求在 1 – 10 mW 量级完成实时解码，内存带

宽则需满足每秒数百万次访存。实时性约束要求端到端延迟低于 20 ms，以避免唇音不同步。

7 质量评估与主观测试

客观指标包括信噪比（SNR）、感知客观差异等级（ODG）与 PEAQ。PEAQ 通过模拟人耳听觉模型，输出 0 – 5 的客观差异分数。主观测试遵循 MUSHRA 与 ABX 协议。MUSHRA 要求听音者在 0 – 100 分量表上对多个样本评分，ABX 则测试被试能否可靠区分 A、B 两段。典型双盲测试表明，192 kbps AAC 在多数内容上与无损的差异已接近随机水平，而 320 kbps MP3 在高频瞬态丰富的音乐中仍可被部分听音者辨别。

8 新兴趋势与挑战

端侧神经音频编解码如 SoundStream、EnCodec 与 Lyra 采用 VQ-VAE 与 Transformer 架构，在 3 – 6 kbps 码率下实现接近传统 32 kbps 的语音质量。模型通过残差向量量化与因果卷积，在保持低延迟的同时大幅提升压缩效率。空间音频与 6DoF 沉浸式内容依赖高阶球谐函数压缩，需在有限比特预算内传输声场系数。可持续性方面，更低比特率直接降低传输能耗，但模型推理能耗成为新变量。开放标准 Opus 已获广泛支持，AV1-audio（IAMF）正处于标准化阶段，有望成为下一代沉浸式音频的开放方案。

从早期「让人听得见」到如今「让人听得爽」，音频压缩技术经历了从波形编码到心理声学编码，再到神经网络编码的演进。AI 正重塑「编码器-声道-耳朵」链路，通过端到端可微分优化实现感知最优。读者可通过挑选合适格式、进行 ABX 盲听，以及关注 ffmpeg、opus-tools 等开源项目，深入理解并实践音频压缩技术。