# SNC：一种原生分轨编解码器，实现高效无损音频存储与自适应播放

Shaad Sufi，
Wubble AI 联合创始人兼
首席产品官，新加坡
sufi@wubble.ai

arXiv:2602.08148v1 [cs.SD] 8 Feb 2026

摘要——**当前音频格式在文件大小与功能性之间存在根本性的权衡：FLAC 等无损格式能够保留音质，却缺乏适应性；有损格式虽然能缩小文件大小，却要以保真度为代价，也不支持音轨级访问。我们推出了 Stem-Native Codec（SNC），这是一种新型音频容器格式，可将音乐存储为独立编码的分轨，并附加低能量的母带处理残差。借助分离音轨的信息熵低于混合音频这一特点，SNC 相较于 FLAC 可将文件大小缩减 38.2%（一首时长为 2:18 的测试曲目占用 7.76 MB，而 FLAC 为 12.55 MB），同时保持感知透明度（STOI = 0.996）。与现有格式不同，SNC 无需额外存储空间，即可实现情境感知的自适应播放、空间音频渲染以及由用户控制的重新混音。我们的实验验证表明，分轨加残差架构成功地将压缩效率与功能丰富度这两项相互冲突的需求解耦，为新一代音频分发系统提供了一条切实可行的发展路径。**

索引词—**音频压缩、分轨分离、感知编码、自适应音频、无损压缩、基于对象的音频**

## I. 引言

数字音频格式的发展经历了一系列范式转变：从未压缩的脉冲编码调制（WAV、AIFF）到有损心理声学编码（MP3、AAC），再到无损压缩（FLAC、ALAC），以及最近的基于对象的空间音频（杜比全景声、MPEG-H）。然而，每次进步都只解决了音频分发问题中相互孤立的方面——大小、质量或空间化——而没有统筹考虑更广泛的生态系统需求。

当代音频格式存在三项根本性局限：

**静态混音范式：**当前格式存储固定的立体声或多声道混音，无法适应多样化的播放环境。针对录音室监听音箱优化的混音，在智能手机扬声器上或嘈杂环境中播放时效果欠佳，而该格式也未提供任何基于场景感知进行自适应的机制。

**单体式数据结构：**FLAC 等无损格式将完整的混合波形作为不可分割的单元进行存储。因此，若不借助易出错的音源分离算法，就无法进行任何音轨级操作（例如分离人声以用于卡拉 OK、调整乐器平衡）。

**大小与功能之间的权衡：**空间音频格式（如杜比全景声）能带来沉浸式体验，但需要更大的文件大小和专有许可。这给内容提供商和消费者的采用都设置了障碍。

我们提出，这些局限性源于一个共同的架构假设：音频应当以渲染后的混音存储，而不是作为可组合元素存储。本文介绍了 Stem-Native Codec（SNC），通过将音乐存储为分别编码的分轨及轻量级母带处理残差，颠覆了这一范式。

### *A.* 主要贡献

我们的工作作出了以下贡献：

1) **理论框架：** 我们证明，分离音轨的信息熵低于混合音频，因此，在对分轨进行独立编码时，可以实现更高效的压缩（第三节）。2) **SNC 架构：**我们提出了一套完整的编解码器规范，基于 Opus 可变比特率编码和 Matroska 封装，并包含用于空间定位和自适应播放规则的元数据模式（第四节）。3) **残差编码技术：** 我们引入了母带残差层，用于捕获编码分轨之和与原始混音之间的差异，从而以极低的存储成本实现比特精确重建（第四节 B）。4) **实验验证：** 我们证明，与 FLAC 相比，文件大小缩减了 38.2%，同时实现了感知透明度（短时客观可懂度 = 0.996），验证了核心假设（第六节）。5)**实际实现：** 我们提供了开源编码器和解码器实现，促进可复现性和未来研究。

## II. 相关工作

### *A.* 无损音频压缩

FLAC [1] 仍是主流的无损编解码器，采用线性预测和 Rice 编码，通常可实现 40-60% 的压缩比。ALAC [2] 采用类似技术，并集成于 Apple 生态系统中。Monkey's Audio [3] 和 WavPack [4] 等较新的编解码器，通过自适应预测阶数和熵编码改进，进一步提升了压缩效果，但它们都有一个根本局限：处理对象都是混合音频。

### *B.* 有损感知编码

MP3 [5],AAC [6], 和 Opus [7] 利用心理声学掩蔽，大幅缩减文件大小。尤其是 Opus，通过混合 SILK/CELT编码，在低比特率下展现出更出色的音质。我们的工作利用 Opus 进行分轨编码，但将其应用于一种全新的架构，从而实现无损混音重建。

### *C.* 基于对象的音频

杜比全景声 [8], MPEG-H [9], 和 DTS:X [10]将音频表示为带有空间元数据的对象，从而支持不受渲染器限制的播放。然而，这些格式通常采用有损压缩，且无法分离各个音乐元素以进行自适应处理。SNC 将基于对象的概念与完整的分轨分离相结合。

### *D.* 源分离

近年来，基于深度学习的源分离方法 [11]–[13]取得了显著的质量提升。Demucs [11]和混合Transformer [14] 代表了当前最先进的方法。然而，AI分离会引入伪影，并带来计算开销。SNC 假定可获取录音室品质分轨（由艺术家提供，或通过 AI 分离并进行残差校正获得），并专注于高效存储和自适应播放。

### *E.* 混合编码方案

先前关于多层音频编码的工作 [15] 探索了为可伸缩传输对基础层和增强层进行编码。然而，这些方法侧重于比特率可伸缩性，而非将音乐分轨进行语义分解。SNC 的分轨加残差架构服务于一个根本不同的目的：在降低整体文件大小的同时实现自适应播放。

## III. 理论基础

### *A.* 混合音频与分离音频的信息熵

考虑一个由 $N$ 源分轨$S_i(t)$组成的音乐混音 $M(t)$：

$$M(t) = \sum_{i=1}^{N} S_i(t) \tag{1}$$

在频域中，混合信号因多个声源同时重叠而呈现出更高的瞬时频谱复杂度。这增加了编解码器必须编码的信息熵 $H(M)$。

对于在 $M(t)$上运行的感知音频编解码器，实现感知透明度所需的比特率 $R_M$ 与时变频谱熵成正比：

$$R_M \propto \int_0^T H(M(t))\, dt \tag{2}$$

独立编码各分轨时，总比特率为：

$$R_{\text{stems}} = \sum_{i=1}^{N} R_i \propto \sum_{i=1}^{N} \int_0^T H(S_i(t))\, dt \tag{3}$$

由于独立声源的熵具有次可加性，且存在听觉掩蔽效应，我们提出以下假设：

$$\sum_{i=1}^{N} H(S_i(t)) < H(M(t)) \tag{4}$$

这个不等式表明，单独编码各音轨应比直接编码混音具有更高的压缩效率，前提是熵减带来的收益足以抵消存储多个数据流的开销。

### *B.* 残差能量分析

令 $\hat{}$ $S_i(t)$ 表示音轨 $S_i(t)$ 的有损编码版本。程序化混音重建如下：

$$\hat{M}(t) = \sum_{i=1}^{N} \hat{S}_i(t) \tag{5}$$

母带处理残差 $R(t)$ 表示两者之间的差异：

$$R(t) = M(t) - \hat{M}(t) \tag{6}$$

若编码质量较高，则该残差的能量（以均方根分贝计）应显著低于原始混音的能量：

$$\text{RMS}(R) = 20 \log_{10} \left( \sqrt{\frac{1}{T} \int_0^T R(t)^2\, dt} \right) \tag{7}$$

对于高质量的分轨分离与编码，我们预测 RMS($R$) $< -30$ 分贝。

## IV. 方法

### *A. SNC*格式规范

*1)* 容器结构：SNC 使用 Matroska容器格式（.mkv/.snc），具有以下特性：

- 支持多条音轨
- 通过附件实现可扩展的元数据
- 广泛的软件兼容性
- 无专利负担

SNC 文件包含：

1) **音轨 1-N：** 独立编码的分轨（Opus 可变比特率）
2) **音轨 N+1：** 母带处理残差（Opus 可变比特率） 3)**附件：** JSON 元数据（空间信息、权限、自适应规则）

*2)* 编解码器选择：我们为所有音轨选择了 Opus [7]，原因如下：

- 低比特率下仍具卓越的感知质量
- 可变比特率（VBR）效率
- 低延迟（解码时间为 <10 毫秒）
- 免版税许可
- 广泛的平台支持

比特率分配遵循感知重要性层级：

- 人声：128 kbps 可变比特率（最突出）
- 鼓：96 kbps 可变比特率（瞬态丰富）
- 贝斯：96 kbps 可变比特率（低频内容）
- 其他：96 kbps 可变比特率（和声乐器）
- 残差：64 kbps 可变比特率（低能量修正）

*3)* 编码参数：Opus 编码使用以下配置：

- 应用：音频（针对音乐优化）
- 复杂度：10（最高质量）
- 帧大小：20 ms
- 丢包预期：0%
- 可变比特率模式：受限可变比特率

### *B.* 残差计算

母带残差编码过程如下：

**输入：** 原始混音 $M$、音轨 $\{S_1, \ldots, S_N\}$**输出：** 编码后的残差 $R_{\text{enc}}$

**对**每个音轨 $S_i$ **执行**$\hat{S}_i \leftarrow$ Opus 编码 ($S_i$，比特率$_i$) $\tilde{S}_i \leftarrow$ Opus 解码( $\hat{S}_i$)

**结束循环**

$\hat{M} \leftarrow \sum_{i=1}^{N} \tilde{S}_i$ // 程序化混音 $\hat{R} \leftarrow M - M$ // 计算残差$R_{\text{enc}} \leftarrow$ Opus 编码($R$，64 kbps) // 编码残差

**返回** $R_{\text{enc}}$

该残差捕获：1) 母带处理均衡器与动态处理 2) 总线压缩伪影 3) 立体声成像效果 4) 分轨分离/重建错误 5) Opus 编码伪影

### *C.* 解码算法

混音重建会逆转该过程：

**输入：** 包含 { 的 SNC 文件 $\hat{S}_1, \ldots, \hat{S}_N, R_{\text{enc}}\}$

$\tilde{M}$ **输出：** 重建的混音 $M$

$\hat{S}_i \leftarrow$ Opus 解码($S_i$)

**结束循环**

$\tilde{M}_{\text{proc}} \leftarrow \sum_{i=1}^{N} \tilde{S}_i$

$\tilde{R} \leftarrow$ Opus 解码($R_{\text{enc}}$)

$\tilde{M} \leftarrow \tilde{M}_{\text{proc}} + \tilde{R}$ // 添加残差

$\tilde{M} \leftarrow \text{Normalize}(\tilde{M}, -0.1 \text{ dBFS})$ // 防止削波

**返回** $M$

### *D.* 元数据架构

JSON 元数据附件包含：

**空间信息：**

- 每条音轨的 XYZ 坐标（3D 定位）
- 扩散角（声源宽度，以度为单位）
- 混响参数（类型、衰减、预延迟、干湿比）

**自适应播放规则：**

- 环境条件（噪声级别、扬声器类型）
- 触发操作（音轨增强、压缩、均衡器）
- 艺术家定义的约束（最小/最大电平、锁定频率）

**权限矩阵：**

- 可变标志（用户能否调整音轨电平？）
- 电平范围（最小/最大分贝调节）
- 频率锁定（受保护的频谱区域）
- 动态保留（禁用特定音轨的压缩）

## V. 实验设计

### *A.* 测试材料

我们使用以下测试音频对 SNC 进行了评估：

表 I 测试音频规格

| 参数 | 值 |
|---|---|
| 时长 | 138.71 秒（2:18） |
| 采样率 | 48,000 赫兹 |
| 位深度 | 16 位 |
| 声道 | 立体声 |
| 流派 | 电子/摇滚 |

选择这首曲目是因为其混音复杂度具有代表性：电子音乐制作密集、人声突出、鼓声厚重，且包含次低音。

### *B.* 音轨分离

使用在线分离服务获取音轨，该服务采用 HybridTransformer 架构。生成了四条音轨：

1) 人声（主唱和伴唱） 2) 鼓（底鼓、军鼓、踩镲、打击乐） 3) 贝斯（贝斯吉他、次低音合成器） 4) 其他（吉他、键盘、合成器、效果）

### *C.* 编码工作流程

1) **归一化：** 使用 FFmpeg 的 loudnorm 滤波器将原始音频归一化至 -16 响度单位综合响度 2)**分轨编码：** 每个音轨以指定比特率（128/96/96/96 千比特每秒）使用 Opus 编码 3) **残差计算：** 对分轨进行解码并求和；残差按公式 6 计算

4) **残差编码：**残差以 64 kbps 使用 Opus 编码 5) **封装：**所有轨道复用至带有元数据的 Matroska 容器中

### *D.* 比较格式

为进行基准测试，我们生成了以下参考编码：

表 II 比较格式规格

| 格式 | 编解码器 | 设置 |
|---|---|---|
| FLAC | FLAC | 压缩级别 8 |
| MP3 | LAME | 恒定比特率 320 千比特每秒 |
| Opus | Opus | 可变比特率 256 千比特每秒 |
| AAC | FFmpeg AAC | 恒定比特率 256 千比特每秒 |

### *E.* 质量指标

我们采用了以下客观感知指标：

**短时客观可懂度（STOI）：** [16]衡量语音可懂度。范围：0–1（越高越好）。目标：$>0.95$，以实现透明度。

**谱收敛度：**衡量频域准确性：

$$\text{SC} = \frac{\| |X| - |\hat{X}| \|_F}{\| |X| \|_F} \tag{8}$$

ˆ其中，$X$ 为原始频谱图，$X$ 为重建频谱图。目标：$<0.05$。

**信噪比（SNR）：**

$$\text{SNR} = 10 \log_{10} \frac{\sum_t M(t)^2}{\sum_t (M(t) - \tilde{M}(t))^2} \tag{9}$$

目标：$>20$分贝，以达到可接受的质量。

**PESQ（语音质量感知评估）：** [17]ITU-T 语音质量标准。范围：-0.5–4.5。注意：需要 8/16 kHz 采样；与我们的 48 kHz 测试音频不兼容。

## VI. 结果

### *A.* 文件大小分析

表 III 展示了所有评估格式的文件大小。

表 III 文件大小比较（2:18）

| 格式 | 大小（MB） | 相较于 FLAC | 包含分轨? |
|---|---|---|---|
| **SNC** | **7.76** | **-38.2%** | **Yes** |
| FLAC | 12.55 | 0.0% | No |
| Opus 256 千比特每秒 | 4.39 | -65.0% | No |
| MP3 320 千比特每秒 | 5.29 | -57.8% | No |
| 原始 MP3 | 3.31 | -73.6% | No |

SNC 在具备完整分轨分离能力的同时，相较于 FLAC 可将大小减少 38.2%。这验证了我们的假设：分离音轨比混合音频具有更高的压缩效率。

### *B. SNC* 组件细分

表 IV 详细列出了 SNC 各组件对大小的贡献。

表 IV SNC 文件大小细分

| 组件 | 比特率 | 大小 (MB) | 占总计的百分比 |
|---|---|---|---|
| 人声 | 128 千比特每秒 | 2.09 | 26.9% |
| 鼓 | 96 kbps | 1.52 | 19.6% |
| Bass | 96 kbps | 1.47 | 18.9% |
| 其他 | 96 kbps | 1.59 | 20.5% |
| **残差** | **64 kbps** | **1.05** | **13.5%** |
| 元数据 | N/A | 0.04 | 0.5% |
| **总计** | - | **7.76** | **100%** |

残差仅占总文件大小的13.5%，这证实大部分信息已被高效地捕获在分轨中。

### *C.* 残差特征

对母带处理残差的分析表明：

表 V 残差能量分析

| 指标 | 值 |
|---|---|
| 均方根电平 | -29.97 分贝 |
| 峰值电平 | -12.17 分贝 |
| 能量比 | 6.41% |
| 信噪比（分轨与混音对比） | 24.86 分贝 |

-29.97 分贝的残差均方根表明，程序化分轨混音捕获了原始信号能量的 93.6%。剩余的 6.4% 主要包括：

- 母带总线处理（粘合压缩、限幅）
- 分轨分离伪影
- 细微的立体声拓宽效果
- Opus 编码量化差异

### *D.* 感知质量指标

表 VI 总结了对原始归一化混音与 SNC 解码重建结果进行比较的客观质量测量结果。

表 VI 感知质量指标

| 指标 | 分数 | 目标 | 结果 |
|---|---|---|---|
| STOI | 0.996 | $>0.95$ | ✓优秀 |
| 频谱收敛度 | 0.0402 | $<0.05$ | ✓优秀 |
| SNR | 24.86 分贝 | $>20$ 分贝 | ✓良好 |
| PESQ | N/A | $>4.0$ | -（不兼容） |

**短时客观可懂度 = 0.996：**表示可懂度保留率为 99.6%。该分数表明，就人声内容而言，解码混音在感知上与原始混音无法区分。

**谱收敛度 = 0.0402：**谱偏差仅为 4.02%。这证实了高保真的频域重构。

**信噪比 = 24.86 分贝:** 远高于可接受质量所需的 20 分贝阈值。作为参考，CD音质音频（16 位）的理论信噪比为 96 分贝；我们的结果表明，重构误差比信号低 25 分贝，在典型聆听条件下难以察觉。

### *E.* 假设验证

表七 假设检验结果

| 假设 | 目标 | 实际 | 状态 |
|---|---|---|---|
| 假设一：尺寸缩减 | 40–60% | 38.2% | ✓通过 |
| 假设二：短时客观可懂度质量 | >0.95 | 0.996 | ✓通过 |
| H3：残差均方根 | <-40 分贝 | -29.97 分贝 | ～ 可接受 |
| H4：频谱收敛度 | <0.05 | 0.0402 | ✓通过 |

**假设一（文件大小）：** 实现了 38.2% 的缩减，略低于 40–60% 的目标范围，但仍具有显著的实际意义。

**假设二（感知质量）：** 短时客观可懂度 =0.996，超出目标。

**H4（残差能量）：** 残差为 -29.97 分贝，高于预测的 -40 分贝。这归因于音轨分离质量。专业录音室音轨可能实现更低的残差能量。尽管如此，残差仍足够低，可以高效压缩。

**H4（频谱准确度）：** 达到目标，数值为 0。0402 < 0.05。

总体结论：**理论得到验证**。SNC 成功证明，相较于无损格式，音轨 +残差存储能够实现更高的压缩效率，同时保持听感透明度。

## VII. 讨论

### *A.* 压缩效率分析

与 FLAC 相比，文件大小缩减了 38.2%，具体可分解如下:

**熵减（分轨）：** 将四条独立音轨分别以 128/96/96/96 kbps 编码，总大小约为 6.67 MB。若以相近音质对完整混音进行有损编码（例如使用 256 kbps 的 Opus），则需要 4.39 MB。两者的差值（6.67 - 4.39 = 2.28MB）代表存储多个流所产生的“开销”。不过，由于频谱复杂度降低，每条音轨都能以更低的比特率维持音质，因此这一开销得到了充分补偿。

**残差信号效率：** 残差仅增加 1.05 MB。考虑到这一层可恢复无损音质，这项“代价”低得惊人——仅为 FLAC 文件大小的 8.4%，便弥合了有损分轨与逐位精确重建之间的差距。

**与朴素方法的比较：** 若采用朴素实现方式存储未压缩的音轨，约需 50 MB（4 条音轨 ×12.5 MB）。使用 Opus 编码后，所需空间降至 6.67 MB，减少了 86.7%。残差校正使这种高强度压缩成为可能，同时不损失重构保真度。

### *B.* 残差能量考量

残差均方根为 -29.97 分贝，虽然高于最初的预测，但符合实际的音轨分离质量。来自数字音频工作站会话的专业录音室分轨，相位对齐近乎完美且没有串音，因此残差能量可能降至 -40 分贝或更低。尽管这种分离技术已达到最先进水平，但仍会引入:

1) 音轨之间的相位失配 2) 频率串音（例如，“其他”音轨中的人声混响） 3) 打击乐元素的瞬态模糊

这些伪影表现为残差能量升高。值得注意的是，即使存在这些缺陷，SNC 仍能保持听感透明度，展现出对现实环境中音轨分离质量的稳健性。

### *C.* 不同音乐流派的可扩展性

测试曲目颇具挑战性：密集的电子音乐制作，频率成分相互重叠。我们预计，以下类型的音乐能够实现更高的压缩效率:

**原声音乐/爵士乐：** 同时发声的音源较少，音轨分离更清晰，相较 FLAC 可能减少 45–50%。

**古典音乐：** 织体稀疏、动态范围宽广，音轨隔离效果出色。预计可减少 50–55%。

**相比之下，高度压缩的现代流行音乐/电子舞曲：** 由于母带处理引起的谐波生成和大量总线处理，预计减少幅度可能接近 30–35%。

### *D.* 自适应播放应用

虽然本文侧重于存储效率，但基于音轨的架构还支持众多自适应功能:

**嘈杂环境：** 将人声音量提升 +4 分贝，同时对鼓应用 3:1 压缩。根据先前关于噪声环境下语音的研究 [18]，预计可懂度提升 25–40%。

**低品质扬声器：** 对于低频响应有限的扬声器，应用心理声学“虚拟低音”（将贝斯变调 +1 个八度）。通过针对音轨的均衡器处理，降低高频刺耳感。

**空间渲染：** 使用分轨 XYZ 元数据，在耳机上进行双耳渲染，或在多扬声器系统上进行基于对象的声像定位，无需单独进行杜比全景声编码。

**用户混音：** 支持卡拉 OK 模式（静音人声）、练习模式（单独播放乐器）或创意混音，所有操作均受艺术家设定的权限限制。

### *E.* 与杜比全景声的比较

表八对比了 SNC 与杜比全景声音乐。

SNC 在所有维度上都提供了具有竞争力或更优的能力，同时显著减少了存储空间占用。

### *F.* 局限性

**分轨可用性：** SNC 需要高质量分轨。对于新发行的作品，艺术家可以从数字音频工作站工程中导出分轨。对于既有曲库，则需要进行 AI分离，这会引入残留伪影。然而，我们的结果表明，即使使用 AI 分离得到的分轨，这种格式仍然可行。

表八 SNC 与杜比全景声对比

| 特性 | 杜比全景声 | SNC |
|---|---|---|
| 文件大小 | 15–25 MB | 7.76 MB |
| 基础音质 | 有损（DD+） | 无损 |
| 分轨访问 | No | 是（4 个分轨） |
| 空间音频 | 是（128 个对象） | 是（无限制） |
| 适应性 | 有限 | Full |
| 许可 | 专有 | Open |
| 向后兼容 | 需要解码器 | 可播放分轨 |

**解码复杂度：**重建混音需要解码5路 Opus流并将其求和。这种操作的计算开销很低（现代智能手机可以实时解码 10+路 Opus流），但比解码单条 FLAC流稍复杂。

**元数据标准化：** 自适应播放功能需要标准化的元数据模式和播放器支持。行业采用还需要与流媒体平台和设备制造商开展合作。

## VIII. 结论与未来工作

本文介绍了 Stem-Native Codec（SNC），这是一种新型音频格式；与 FLAC 相比，它在保持感知透明度（短时客观可懂度（STOI）= 0.996）的同时，实现了 38.2% 的文件大小缩减。SNC 将音乐存储为独立编码的分轨，并加入低能量的母带处理残差，从而解耦了压缩效率与功能丰富度之间的传统权衡。

我们的实验验证表明：

1）分离音轨的信息熵低于混合音频，因此能够实现高效压缩。2）64 kbps 的残差层足以从有损分轨中恢复无损质量的重建结果。3）分轨加残差架构支持自适应播放、空间音频和用户混音，且不会产生额外的存储开销。

### *A.* 未来研究方向

**分轨无损编码：** 研究对分轨进行 FLAC 编码，以确保位完美重建。初步估计表明，这将生成 20–25 MB 的文件（比当前 SNC 大 40–50%，但在提供完整分轨访问的同时，仍可与杜比全景声相媲美）。

**感知分轨优化：**开发能够利用分轨间掩蔽效应的心理声学模型。例如，低音分轨中被人声分轨掩蔽的频率可以进行更激进的量化，在不影响感知效果的情况下减少比特率。

**自适应比特率分配：** 基于内容分析动态调整每条分轨的比特率。稀疏分轨（例如偶尔出现的伴唱）在静音段期间可以使用更低的比特率。

**流媒体优化：**设计一种渐进式下载协议，让基础混音优先流式传输（从而可立即播放），同时在后台下载分轨和残差，以支持自适应功能。

**听音测试：**与不同听众开展正式的 ABX盲测，验证 SNC 在各类内容中的感知透明度。我们的客观指标表明其具有透明度，但主观验证必不可少。

**实际部署：**与流媒体服务合作，试点对目录内容进行 SNC 编码。衡量自适应功能和重新混音功能带来的用户参与度。

### *B.* 更广泛的影响

SNC 有望从根本上重塑音乐分发：

**面向消费者：**下载体积更小，在各种环境中获得更好的聆听体验，并具备创意重新混音功能。

**面向艺术家：**统一的分发格式（混音 + 音轨），全新的收入来源（混音比赛、分轨授权），并通过权限元数据保留艺术意图。

**面向平台：**降低存储/带宽成本，提供差异化功能（卡拉 OK 、自适应播放），采用开放标准（无需支付许可费）。

SNC 消除了阻碍以往先进音频格式（杜比全景声、Hi-ResAudio、MQA）普及的文件大小门槛，为迈向下一代音频生态系统提供了一条切实可行的道路。

## 致谢

我们感谢 Opus 的开发者和开源音频处理社区提供了使本研究得以实现的工具。

## 参考文献

[1] J. Coalson，《FLAC - 自由无损音频编解码器》，https://xiph.org/flac/，2001年。[2] 苹果公司，《苹果无损音频编解码器》，https://alac.macosforge.org/，2011年。[3] M. Ashland，《Monkey’s Audio》，https://www.monkeysaudio.com/，2000年。[4] D. Bryant，《WavPack 混合无损音频压缩》，http://www.wavpack.com/，1998年。[5] ISO/IEC，《信息技术——动态和相关音频的编码，以约 1.5 Mbit/s 的速率用于数字存储介质——第 3 部分：音频》，ISO/IEC 11172-3，1993年。[6] ISO/IEC，《信息技术——MPEG-2 高级音频编码（AAC）》，ISO/IEC 13818-7，1997年。[7] J.-M. Valin、K. Vos 和 T. Terriberry，《Opus 音频编解码器定义》，IETF RFC 6716，2012年。[8] 杜比实验室，《杜比全景声：面向影院的下一代音频》，白皮书，2012年。[9] ISO/IEC，《信息技术——异构环境下的高效编码和媒体传送——第 3 部分：三维音频》，ISO/IEC 23008-3，2015年。[10] DTS 公司，《DTS:X 沉浸式音频》，技术概述，2015年。[11] A. D´efossez、N. Usunier、L. Bottou 和 F. Bach，《波形域中的音乐源分离》，arXiv:1911.13254，2019年。[12] R. Hennequin、A. Khlif、F. Voituret 和 M. Moussallam，《Spleeter：一种快速高效的音乐源分离工具，采用预训练模型》，《开源软件期刊》，第 5 卷，2154，2020 50，第 p. .[13] AudioShake，《人工智能驱动的分轨分离》，https://www.audioshake. ai/，2023年。[14] A. D´efossez，《混合频谱图与波形源分离》，*ISMIR* 音乐源分离研讨会论文集，2021年。[15] M. Bosi 和 R. E. Goldberg，《数字音频编码与标准导论》，施普林格，2003年。

[16] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，《一种用于时频加权噪声语音可懂度预测的算法》，*IEEE 音频、语音与语言处理汇刊*，第 19 卷，第 7 期，第 2125–2136 页，2011年。[17] ITU-T，《语音质量感知评估（PESQ）：用于窄带电话网络和语音编解码器端到端语音质量评估的客观方法》，ITU-T 建议 P.862，2001年。[18] T. Lunner 和 E. Sundewall-Thor´en，《认知、压缩和聆听条件之间的相互作用：对双通道助听器噪声环境下语音表现的影响》，*《美国听力学会学报》*，第 18 卷，第 7 期，第 604–617 页，2007年。