ICS 03080
CCSA12

团　　体　　标　　准

T/INFOCA 2—2024

面向虚拟现实场景的音频重构技术指南

Guidelines of audio reconstruction techniques for virtual reality scenes

（征求意见稿）

（本稿完成日期：2024 年 3 月 25 日）

××××-××-××发布　　××××-××-××实施

中关村现代信息消费应用产业技术联盟　发布

目　次

前　言

本文件按照GB/T 1.1——2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》给出的规定起草。

本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由中关村现代信息消费应用产业技术联盟提出并归口。

本文件起草单位：中国传媒大学、广东南方新媒体股份有限公司、河南广播电视台、哈尔滨工业大学、中国科学院大学、中兴通讯股份有限公司

本文件主要起草人：叶龙，程皓楠，刘淑琳，蔡娟娟，胡飞，王兵、王庆宝、陈志业，王春阳，贾中原，张谦，范晓鹏、王兴涛，张新峰，安泓宇，李欢洋，刘成刚，陶长标，史美康

引　　言

随着多媒体与人工智能等领域取得关键理论突破，虚拟现实产业进入高速发展阶段。音频技术作为虚拟现实的基础技术，在不同业务应用场景中具有不同新技术需求。

面向虚拟现实场景的音频重构技术指南标准便是为规范虚拟现实音频应用而起草的。本文件给出了面向虚拟现实场景音频重构技术的相关建议。本标准中提到的音频重构技术面向的是有视觉画面的虚拟现实视听场景。随着数字技术的日益进步，虚拟现实已跨越了科幻的边界，逐步成为日常生活中不可或缺的一部分。在这一领域中，音频重构技术尤显重要，它不仅赋予了虚拟环境以生动的听觉感受，而且极大地增强了用户的沉浸感及互动体验。音频重构技术的核心在于其对声波在三维空间中传播特性的模拟能力。通过精确地再现声音的方向性、距离感、反射和衍射等属性，该技术为用户营造出一个具有高度真实感的听觉空间。这种空间化处理能力不局限于任何特定的应用场景，而是广泛适用于游戏设计、影视制作、教育训练以及心理治疗等多个领域。在电影和游戏产业中，音频重构技术使得音效设计师能够创造出与视觉内容高度匹配的环境音效，为观众或玩家提供一种身临其境的体验。在教育域，该技术能够模拟各类自然环境或人造场景中的声响，从而增强学习者的沉浸感，提升教育训练的有效性。在心理治疗实践中，音频重构被用于营造有助于放松或治疗的听觉环境，辅助患者更好地应对心理压力与焦虑情绪。此外，音频重构技术的发展催生了新的交互模式。例如，在多人在线VR游戏中，玩家可以通过声音定位来判断其他玩家的位置和距离，这不仅增加了游戏的互动性，也提高了其战略性。在虚拟会议应用中，音频重构技术提供了更为自然的沟通体验，促进了远程协作的效率与愉悦度。

本标准制定对虚拟现实音频技术规范性具有指导价值，有助于规范虚拟现实产业空间形态，促进虚拟现实音频产业发展。

面向虚拟现实场景的音频重构技术指南

1 范围

本文件给出了面向虚拟现实场景的音频内容呈现过程中的音频重构技术总体原则和技术建议。

本文件适用于面向虚拟现实场景的音频内容呈现过程中的音频重构方案设计。

2 规范性引用文件

下列文件对于本文件的应用是必部可少的，凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包含所有的修改单）适用于本文件。

GY/T 363—2023 三维声编解码及渲染

GB/T 38258-2019 信息技术 虚拟现实应用软件基本要求和测试方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

虚拟现实 virtual reality

采用以计算机为核心的现代高科技手段生成的逼真的视觉、听觉、触觉、嗅觉、味觉等多感官一体化的数字化人工环境,用户借助一些输人、输出设备,采用自然的方式与虚拟世界的对象进行交互,相互影响,从而产生亲临真实环境的感觉和体验。

[来源：GB/T 38258-2019, 2.1]

3.2

音频重构 audio reconstruction

将给定的音频输入信号建模、转换为适用于终端扬声器配置的、可直接重放的音频格式的过程。

3.3

声道 channel

用于传送到单个扬声器或其他重放设备的一组有序音频样本集合。

[来源：GY/T 363—2023, 3.7]

3.4

双声道立体声 stereo audio

一种音频格式，该格式下，使用两个声道承载有一定相位关系的音频信号，通常通过位于听音者前方的两个对称的扬声器或使用耳机重放，带给听音者更宽的声场感觉。

[来源：GY/T 363—2023, 3.8]

3.5

环绕声 surround sound

一种音频格式，该格式下，使用多个声道承载构成完整音频内容的多路音频信号，通过位于听音者耳部高度层的环绕听音者的多个扬声器重放，给听音者带来被环绕的声场感觉。

[来源：GY/T 363—2023, 3.9]

3.6

三维声 3D audio

一种音频格式，该格式下，多个声道承载构成完整音频内容的多路音频信号，通过环绕听音者的位于不同高度层的多个扬声器直接重放，或经过渲染或映射后重放，提供更高的声像定位空间解析度，并给听音者带来沉浸式的声场感觉。

[来源：GY/T 363—2023, 3.10]

4 缩略语

3D：三维(3-dimension)

6DoF：六自由度(6 degree of freedom)

FFT：快速傅里叶变换(fast fourier transform)

MFCC：梅尔倒谱系数(mel-frequency cepstral coefficients)

HRTF：头相关传递函数(head related transfer functions)

ITD：时间差(interaural time difference)

ILD：声级差(interaural level difference)

5 音频重构技术建议

5.1 系统架构

音频重构系统宜由音频元数据内容建模、音频元数据同步计算、空间声重放三部分组成，系统架构见图 1。

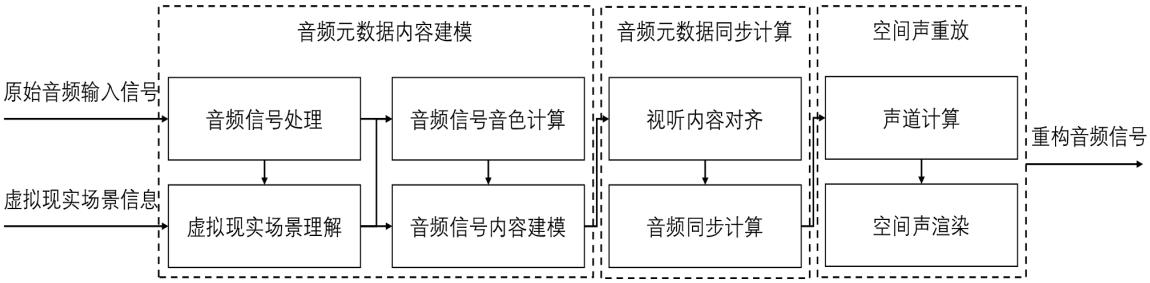

图 1 音频重构技术系统架构

5.2 音频重构技术建议

不同于现有三维声相关标准针对音频内容编码，终端解码与渲染进行技术规范，本标准对虚拟现实场景中视听关联的音频重构技术进行建议，宜遵循“交互式、沉浸式”的整体技术理念，综合考虑声源和声场两方面的交互特点，实现“所见即所听”的虚拟现实场景音频呈现，具体各个模块建议如下。

5.2.1 音频元数据内容建模

音频信号处理：对于原始音频输入信号的处理是音频重构中的关键步骤。在音频重构中，原始音频输入信号的处理流程应该包括以下步骤：

（1）预处理：这包括增益调整、去噪、预强调等，目的是确保音频信号的清晰度和适当的输入级别。

（2）分帧：为了处理的稳定性，通常需要将音频信号分帧。这意味着将从原始音频流

中截取固定帧长的音频片段进行处理。

音频信号音色计算：在进行了初步的音频信号处理之后，音色计算是一个专门调整和改善音频信号特性的过程，以匹配预设的声音标准。

（1）频谱分析与平衡：首先，使用快速傅里叶变换（FFT）将预处理后的音频信号从时域转换到频域。分析频谱成分，确定哪些频率需要增强或减弱以获得期望的音色。然后应用均衡器处理来调整这些频率的增益，以达到平衡的频谱。

（2）音色特征提取：利用如梅尔频率倒谱系数 MFCC、谱图分析等技术提取音频信号中的音色特征。这些特征包括音高、亮度、带宽、谐波结构等。

虚拟现实场景理解：进行虚拟现实场景理解的目的是为了捕捉和分析场景的特性，并将这些信息用于指导音频信号的内容建模，以便生成与虚拟环境相匹配的声音效果。

（1）场景元数据获取：收集虚拟现实环境中的元数据，包括场景的几何结构、物体属性、材料类型、光源位置等。这些信息对于理解声音在场景中的传播至关重要。

（2）视觉和听觉一致性分析：分析视觉内容与期望的听觉体验之间的关联。例如，一个开放的广场可能会产生回声，而一个小房间则可能产生更多的共鸣。确保音频建模与用户的视觉感知一致。如果场景中有动态变化的元素，如移动的物体、开关门等，这些因素也需要考虑进音色计算中，以实现实时的音频响应。

音频信号内容建模：音频信号内容建模是一个根据音色计算和虚拟现实场景理解来创造、调整和优化声音的过程，目的是确保声音与视觉内容在时间和空间上对齐，提供一个连贯且沉浸的用户体验。

（1）环境声音设计：基于场景理解的结果，设计环境声音（如背景噪声、远处的对话声等），并确定它们在 3D 空间中的分布和动态变化。

（2）音源建模：对于每个活跃的音源（如角色对话、物体互动声等），根据其在虚拟环境中的位置、属性和行为，建立相应的音频模型。这可能包括录制原始声音素材或使用合成方法生成声音。

（3）声学模拟：使用头相关传递函数（HRTF）和其他 3D 音频处理技术来模拟声音从音源到用户耳朵的传播过程，包括距离衰减、多普勒效应和头部追踪。使用声学仿真软件或算法来模拟声音在虚拟环境中的传播。这涉及到计算反射、吸收、衍射和散射等声学行为。

（4）交互性分析：在虚拟现实中，用户的行为可能会影响环境响应。分析这些交互行为对声音输出的潜在影响，并在音频模型中加以体现。

5.2.2 音频元数据同步计算

视听内容对齐：这个过程涉及到将音频事件与视觉事件精确匹配，并确保它们在用户的体验中协调一致。

（1）时间线匹配：创建一个共同的时间线，用于记录视觉和音频事件。确保所有的视听事件都根据相同的时间参考进行同步。

（2）空间定位一致性：使用视觉指导的音频源定位技术来确保声音的空间位置与视觉元素的位置相匹配。

（3）实时监控和反馈：在实时渲染过程中，监控系统的性能，收集用户反馈，以便快速识别并解决任何同步问题。

音频同步计算：在进行音频同步计算时，首先需要确立一个主时钟作为同步的时间基准。并在播放过程中相应调节，以维持同步状态。

在多声道的空间声处理中，每个声道都需要独立的数据流。例如，立体声需要左右两个声道的数据流，而双声道则需要两倍于单声道的数据量。因此，音频同步计算不仅要保证时间上的对齐，还要确保每个声道的数据都准确无误地提供给后续的处理过程。

5.2.3 空间声重放

声道计算：根据音频同步计算的结果，进行双声道立体声、三维声和空间声的声道计算需要遵循一系列步骤，这些步骤确保声音数据在三维空间中的准确渲染和重构。

（1）确定声音方位：为了在三维空间中创建声音，需要确定声音方位。这涉及到时间差（ITD）和声级差（ILD），这两个因素是根据声源与双耳的距离差异以及头部对声音的遮挡效应来判断的。

（2）头部追踪与动态调整：利用惯性测量单元采集头部追踪数据，并将这些数据实时反映到音频播放上。这样，当用户头部移动时，声音定位也会相应变化，增强临场感和沉浸感。

（3）音频数据处理：根据声道配置（如5.1环绕声或7.1环绕声），处理音频数据以匹配每个扬声器的位置。这涉及到对原始双声道或立体声数据的上混（upmixing）处理，以及可能的编码和解码过程。

空间声渲染：根据空间声的声道计算结果，进行空间声渲染与重构的步骤可以包括以下几个方面：

（1）声场重构：利用声道计算结果，可以通过声场重构技术如Ambisonics来创建一个三维的声场环境。Ambisonics是一种基于声波传播的数学模型，能够在不同的扬声器配置下重现一个点声源或多个点声源组成的声场。高阶Ambisonics提供了更高的空间分辨率，适合重构较大区域内的空间声场。

（2）空间声重放：在耳机等个人听觉设备上结合重构声场模拟出类似真实环境中的立体声效果，以便在用户耳边重现出具有深度和方向性的声音。

5.3 音频元数据内容建模与同步计算技术建议

5.3.1　在计算时效性方面（或者在计算速度方面），技术宜满足音频重构时间与重放时间总和小于 100 ms。

5.3.2　在物理一致性方面，音频重构内容宜满足与声源形状、材质、硬度等物理属性一致。

5.3.3　在风格多样性方面，音频重构内容宜满足支持写实、卡通、夸张等多种风格属性。

5.3.4　在情感一致性方面，音频重构内容宜满足与虚拟场景状态一致。

5.3.6　在动态同步度方面，音频重构内容宜满足与声源运动状态一致。

5.3.6　在交互同步度方面，音频重构内容宜满足与用户当前交互动作一致。

5.3.7　在编辑粒度方面，音频内容制作宜满足支持局部可编辑。

5.3.8　在采样精度方面，音频生成采样率宜满足支持 32 kHz-192 kHz。

5.4 空间声重放技术建议

5.4.1　在呈现多样性方面，技术宜满足支持 6DoF 呈现。

5.4.2　在场景一致性方面，技术宜满足虚拟现实场景中任意声源应具备方位感与指向性，与其虚拟空间位置保持一致。