# 视觉驱动三维声 超高清视听的“声音革命”

在超高清视频成为行业主流的今天，4K/8K 画质早已走进千家万户，但你是否有过这样的体验：画面清晰到纤毫毕现，声音却总感觉“跟不上”--或定位不准，或杂音干扰，沉浸感大打折扣？“基于视觉理解的三维声智能化重合成技术”由中国传媒大学与广东南方新媒体股份有限公司联合研发，以深度学习为核心，实现了视觉与听觉的精准协同，为超高清视听内容生产提供了高效解决方案。

## 一、技术背景

### 超高清时代的“声音刚需”

当前，超高清产业发展与国家文化数字化战略高度契合，4K/8K 超高清技术广泛应用于多地频道、博物馆展厅、端游等场景。作为超高清六维技术的核心组成，声音直接影响用户的沉浸体验。数据显示，目前全球三维声市场需求持续暴涨，预计 2030 年市场规模将增长至 164.6 亿美元，复合年增长率高达 12.81%。然而，传统三维声合成依赖人工塑声，不仅成本高、制作周期长，效果也因混音师主观经验而异，难以满足大规模标准化生产需求。更突出的问题是“视听两张皮”：一方面，沉浸式三维声重合成技术相对视觉超高清技术发展缓慢。另一方面，大多数三维声重合成方法只侧重音频，忽略了与视觉信息的协同互补，使沉浸式体验大打折扣。还有一些使用视觉作为空间信息重构指导的方法则普遍存在视听同步性不足、空间定位精度低、背景噪声干扰等问题，无法实现声音与画面动作、场景空间的动态适配，成为制约超高清产业发展的关键瓶颈。因此，亟需一套“视听协同”的全新解决方案。

## 二、三大核心技术

### 重构三维声生产逻辑

“基于视觉理解的三维声重合成技术”的核心突破是用“视听计算”替代“手工录制”，构建了一套“体验量化指导—算法自动补采—视听动态映射”的三维声重合成智能系统，并提出了“评测端定位、采集端降本、重构端提质”的三维优化思路，实现了从人为主观感知到算法客观评价，从专用设备依赖到通用音频计算，从视听静态匹配到动态同步的跨越。

### 1、脑电技术赋能，让“听感”客观且可量化

当前视听质量评估多采用人工多级评分方法，但该方式受个体审美偏好、主观判断差异的影响显著。面对海量视听内容，迫切需要构建一套统一、稳定的质量评估标准，并形成客观化的视听感知量化方法。为此，团队创新提出“群体客观性度量”的解决方案，以脑电特征为客观表征载体，实现跨个体的感知质量统一度量，突破传统主观评价的局限性。通过采集不同年龄段被试者的脑电时域、空域、频域特征，将视听感知质量拆解为清晰度、一致性、沉浸度三个核心维度，成功将“主观经验判断”转变为“客观数据度量”，为三维声合成提供了精准的优化依据。

具体技术实施流程可概括为：首先选取多组音视频片段作为刺激材料，生成标准化的音视频刺激序列，随后开展脑电实验，采集被试接收刺激时的脑电响应。之后，基于采集的脑电信号，提取其时域、频域及空域特征，进而构建脑电感知评分预测模型，最终形成“基于脑电响应

的质量评价指标”，完成视听感知质量的客观量化。总的来说，这一技术建立了脑电特征与视听质量之间的定量度量关系，为后续三维声合成的自主驱动优化，提供了客观、可量化的感知质量依据。

**2、智能去噪，留住纯净原声**

由于录制设备和录制环境的随机性，设备间的电路噪声以及录制环境的背景噪声会直接影响未空间化的音频听感。而现有的音频去噪方案对于不同类型、不同频段的含噪声音频采用无差别处理模式，这导致噪声残留，或损坏非噪声谐波结构。因此，亟需建模音频谐波结构，实现自适应去噪。针对这一问题，团队创新提出自适应高效去噪模型，包括两个模块：第一个是基于高效通道注意力机制的特征学习模块，通过高效通道注意力机制捕捉通道间局部依赖，分区挖掘局部细节特征，结合监督注意力子模块强化目标音频特征，针对性建模音频谐波结构，在数据驱动模式下区分噪声与音频的有效成分。第二个模块则基于细粒度特征实现自适应降噪，避免无差别处理带来的听感损失。该技术能平衡噪声滤除效果和音频表达完整度，实现54.6%的噪声滤除度，显著优于传统去噪方法，为三维声合成提供纯净输入源。

**3、视听时空动态同步，声随画动**

在空间音频重合成的子领域，即立体声音频重合成方法中，往往采用视听内容整体分析策略，无法捕捉声源动态变化，导致位置错误、发声状态误判等问题。本团队提出“分离-混合”两步法，首先通过时空动态分析算法，将复杂场景拆分为多个独立视听区域，之后并行完成各区域视听特征提取与融合。技术上，创新采用基于声源区域的视听特征融

合编码方法，通过 ResNet-18 网络提取视频帧浅深层特征，精准定位潜在发声区域。结合关联金字塔网络实现跨模态特征融合。最后将各独立视听区域的三维声音频按通道混合，实现声源位置与画面动态的实时匹配。该方案显著提升了合成精度与效率，其中视听一致度高达 64.3%，计算效率方面英伟达 RTX A5000 单卡处理 10 秒音频仅需 0.491 秒。

## 三、10 年深耕结硕果

### 技术落地多场景惠及千万用户

经过近 10 年深耕，“基于视觉理解的三维声重合成技术”的研发团队构建了“理论研究-技术突破-平台开发-产业应用”的完整创新链。不仅在 IEEE TPAMI、ACM MM 等顶级期刊和会议发表论文 10 余篇，还申请/授权发明专利 5 项，发布行业标准 4 项，软件著作权 1 项。扎实的技术积累，最终转化为实实在在的应用价值。目前，这项技术已在超高清视频制作、老电影修复、文博展览等多个领域规模化落地，成效显著。

## 四、未来可期

### 智能+定制化三维声在路上

“基于视觉理解的三维声智能化重合成技术”通过打通视觉与听觉的跨模态协同壁垒，不仅破解了超高清内容生产中制作效率低、视听不同步等行业痛点，还为影视制作、网络直播、VR/AR 等领域提供了低成本、高质量的三维声解决方案。未来，团队还将进一步引入语音识别、文字识别等多模态信息，强化复杂场景适配能力，并结合用户个性化需求，实现定制化三维声合成。相信在各项技术的共同推动下，超高清视听产业将持续向智能化、标准化、规模化方向发展。

来源：科技视听