Opto-Electronic Engineering
光 电 工 程

Article

2018 年，第 45 卷，第 6 期

DOI: 10.12086/oee.2018.170742

一种用于 VR 场景的声音渲染优化方法

陈天石，帖　云*，齐　林，陈恩庆

郑州大学信息工程学院，河南 郑州 450001

摘要：对于包含成百上千可移动声源的虚拟场景，由于聚类阶段所需运算代价过高，传统的空间声音渲染方案往往需要占用过多的运算资源。这已经成为 VR 音频渲染技术发展的瓶颈。本文在声音采样的过程中运用分数阶傅里叶变换这一工具，降低了模数转换阶段的量化噪声。此外，通过在聚类这一步骤中添加平均角度偏差阈值的方法提高了声音处理的运算速度，改善了整个系统的运算效率。设计并进行一项感知用户实验，证实了在可视情况下，人对不同类声源聚类产生的空间误差更加敏感这一观点。根据这一结论，本文提出了一种新的空间声音聚类方法，在可视情况下降低了不同类声源聚类为一组的可能性。

关键词：声音渲染；聚类；感知用户实验；平均角度误差

中图分类号：O436.3　　**文献标志码：**A

引用格式：陈天石，帖云，齐林，等. 一种用于 VR 场景的声音渲染优化方法[J]. 光电工程，2018，**45**(6): 170742

An improved method to render the sound of VR scene

Chen Tianshi, Tie Yun*, Qi Lin, Chen Enqing

School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China

Abstract: Based on the virtual scene containing hundreds of movable sound sources, due to the high computational cost of clustering stage, the traditional spatial sound rendering schemes often take up too much computing resources, which have become a bottleneck in the development of VR audio rendering technology. In this paper, we use fractional Fourier transform (FRFT) as a tool in sound sampling to reduce the quantization noise during the ADC conversion stage. Moreover, we improve the processing speed of sound rendering and the operation efficiency of the entire system by adding the average angle deviation threshold in the clustering step. In addition, we design and implement a perceptual user experiment, and validates the notion that people are more susceptible to spatial errors in different types of sound sources, especially if it is visible. Based on this conclusion, this paper proposes an improved method of sound clustering, which reduces the possibility of clustering different types of sound sources.

Keywords: sound rendering; clustering; perceived user experiments; average angle error

Citation: Chen T S, Tie Y, Qi L, *et al.* An improved method to render the sound of VR scene[J]. *Opto-Electronic Engineering*, 2018, **45**(6): 170742

收稿日期：2017-12-30；　收到修改稿日期：2018-03-15

作者简介：陈天石(1994-)，男，硕士研究生，主要从事虚拟现实的研究。E-mail：1243325667@qq.com

通信作者：帖云(1973-)，男，博士，教授，主要从事虚拟现实的研究。E-mail：ieytie@zzu.edu.cn

1 引 言

在 VR 领域中，空间声音渲染技术正起着越来越重要的作用。声音的空间感对提升 VR 场景中用户的沉浸感起着非常重要的作用。当前比较先进的空间声音渲染方案主要分为基于声音波形和基于射线追踪算法两种类型。目前比较重要的研究成果有以下几点：Tsingos N 提出一种基于剔除与聚类的动态空间声音渲染方法[1]，使得实时处理大规模声源成为可能。Moeck T 将视觉因素考虑在聚类算法中，降低了聚类代价[2]。Schissler C 提出的新型算法解决了聚类阶段，障碍物对聚类结果的影响[3]。陶然教授等提出了基于分数阶傅里叶变换的量化噪声抑制方法，提高了信号采样的质量[4]。然而，这些传统的声源聚类方法，通常会将声源不分类型地聚类到一个点上，这样会造成更大的空间信息误差，特别是在视觉范围内的情况，这种误差特别明显。本文提出了基于声源标签化的聚类新方法，可以有效降低这种空间感知误差。

另外，考虑到传统声源聚类算法往往对每一帧音频数据进行处理，但现实情况是声源的聚类结果通常不会变化那么频繁。即是说，在空间位置变化较小的情况下，声音仍然以每一帧为单位进行聚类运算，这无疑增加了聚类的运算代价。因此在上述改进工作的基础上，本文还提出了平均聚类角度偏差阈值作为判别是否要进行聚类的依据，减少聚类过程的重复计算。

2 经典渲染过程及其改进工作

2.1 经典的空间声音渲染方案

1) 声音数据的采集

声音资源的录制主要利用声音采集器将现场的模拟声音信号转变为数字声音信号。空间声音渲染系统就是针对这些数字声音信号进行处理的。

2) 获取声源以及收听者位置信息

声源以及听者的位置信息对计算空间声音来说非常重要，用于计算声音衰减、聚类等操作。所以整个系统首先要提取声源与听者的位置信息，以方便后面步骤的顺利进行。

3) 声压的计算以及空间化渲染

在空间声音处理系统中，声压的实时计算是极其关键的一步。根据文献[1]所述，心理声学中有一种非常重要的效应，即遮蔽效应。为了实现遮蔽效应，系统首先需要实时计算出每一帧的声压(声源最终到达左右耳朵的声压级)。其次，通过与遮蔽阈值进行比较，将人耳听不到的声源剔除掉，从而减少计算代价。最后将剔除后保留下来的声源按位置与声压强弱进行聚类，对他们的聚类代表进行空间渲染(如遮挡、混响)。

4) 空间声音的混合计算(主要直接用头部相关变换函数(head related transfer function，HRTF)，双耳效应的实现)

对经过空间渲染后的声源进行 HRTF 滤波器处理，实现声音的双耳效应，然后将声源输出到人的双耳。

2.2 改进工作

1) 在声音信号采集过程中，借鉴文献[4]中提到的基于分数阶傅里叶变换的量化噪声抑制方法去降低声音信号采样的量化噪声，以提高声音信号的质量。

2) 在经典算法的聚类阶段，将不同类声源的差异性考虑在聚类算法中，使得聚类后声音的主观体验得到了提升。

3) 另一方面，在聚类的过程中引入平均聚类角度误差阈值，简化了聚类算法的复杂度，提升了算法运行的运算效率。

最终经过改进的系统流程如图 1 所示。

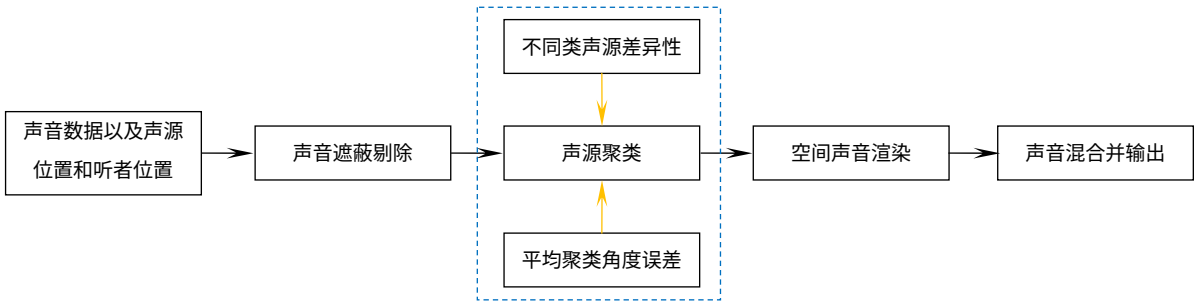


图 1 声音渲染方案的流程

Fig. 1 The process of sound rendering program

3 声音信号采集及聚类方式的优化

3.1 声音信号采集优化

声音信号的采集是空间声音渲染过程的关键。采集到的声音信号的质量往往决定最后渲染出的声音信号的好坏。然而，采集过程中的模数转换不可避免地会引入量化噪声，影响采集到的声音信号质量。为了限制量化噪声，借鉴文献[4]中提到的方法对声音采集器进行改进。由文献[4]可知，量化误差序列与样本序列不相关，量化序列的分数阶功率谱密度是量化误差序列与样本序列的分数阶功率谱密度之和。根据输入声音信号的频率特点，其分数阶傅里叶域带宽是已知的，则量化噪声可以用相应的分数阶傅里叶域数字滤波器进行抑制，从而实现声音信号质量的提升。更加详细的推导过程参考文献[4]。

3.2 聚类方式的优化

在计算大规模移动声源的情况下，当前主流声音聚类算法(Tsingos 等学者提出)在执行的过程中需要实时对每一音频帧进行聚类计算。虽然聚类结果往往与声源相对于听者的空间位置有关，随空间位置的变化而变化。但只有在空间位置变化量增大到某个范围时，聚类结果才会被改变。这就导致了重复的运算，带来了极高的运算代价。

为了减少不必要的聚类运算，引入了平均聚类角度偏差 θ 这一参数作为阈值，用来判断是否需要进行聚类运算，θ 的计算式：

$$\theta_t = \sum_{i=1}^{k} \frac{\theta_i^t}{k} \quad , \tag{1}$$

其中：θ_i^t 表示第 i 个声源在第 t 帧聚类后的位置与原来位置相对于听者位置的角度偏差。θ_t 表示第 t 帧整体聚类的平均角度偏差。在实验过程中，运用自适应算法测出判别阈值为 30° 。如果 $\theta_t \le 30°$ ，则聚类结果与上一帧结果相同；如果 $\theta_t > 30°$ ，则聚类结果依据原来的聚类算法进行实时更新。不同聚类方式的实验效果如表 1 所示。

由实验结果表明，经过改进的声源聚类方法能有效降低声源聚类的可能性，大概是传统方法的一半。另外，从 CPU 占有率看，新的方法可以大幅降低声音处理系统对 CPU 的占有率，这就使得上千数量级别的声源数量同时进行高质量的空间音频渲染成为可能。虽然，新的算法会增加聚类时的平均角度误差，但从实验结果上看，与传统方法相比，增加的空间角度误差只有 5°左右，影响不大。

4 基于声源标签化的聚类算法

4.1 主观实验分析

选择以下实验以提供一些关于是否应该将不同类声源分开聚类的依据。主观测试场景由十个静态物体构成，每个物体分别发出不同的声音(主要分为环境、语音、音乐三类声源)。针对每一类声源，我们准备了十个音频用于实验。

在此设置了两个主要的对比项，同类声源聚类与不同类声源聚类。参与实验的人员有十个人(三女七男，年龄在 23 岁到 45 岁之间，听觉与视觉都正常)。他们之前都没有进行听觉测试相关的经验。在开始实验之前，要求他们先熟悉单个声源音效以及该声源相对应的视觉代表(如汽车和引擎声)。实验要求测试者头戴 HTC vive 头盔，在动作捕捉摄像头工作的区域内进行测试。

头戴式耳机可以用于声音输出，系统运用听觉 HRTF 数据组进行双耳化渲染。测试者在网上测出最佳的 HRTF 数据组并分别应用于实验中，这样可以使测试者获得更加准确的空间信息。安排五个测试者进行同类声源场景测试，另外五个进行不同类声源测试。

表 1 传统方法与新提出的方法的比较

Table 1 Comparison of two methods

聚类数量	声源聚类的可能性/%		平均角度误差/(°)		CPU 占有率/%	
	传统方法	新提出的方法	传统方法	新提出的方法	传统方法	新提出的方法
200	64	20	17.3	27.8	18	12
400	56	23	21.1	25.4	36	20
600	60	19	16.5	28.6	68	33
800	52	29	22.4	27.9	87	46

为了使效果更加明显，将发声物体放置于以测试者为中心的圆形范围内。按声源类别配置为六种场景：1(all ambient)，2(all musics)，3(all speech)，4(ambient+music)，5(ambient+speech)，6(music+speech)。其中包含两类声源的按每种声源选取五个配置实验。每种场景都配合随机变化地声源位置进行实验，并且通过位置的变化重复测试 20 次。

使用文献[5]中提到的实验方法进行实验设置：每种场景都有 A 和 B 两个版本以及为进行聚类的参考版本 R。AB 两个版本中总有一个与参考版本 R 完全一样，另外一个与六种配置中的一个有关。针对每一种场景，测试者可以通过游戏手柄将场景在 A、B、R 三个版本之间进行转换。测试者需要将 AB 两个版本与 R 进行比较，给出仿真质量的评估分数。在测试的时候，要求测试者尤其关注声源的空间位置来进行评判，最后的评判有四个等级：无差别(90 分~120 分)，轻微的差别(60 分~90 分)，有差别(30 分~60 分)，差别较大(0~30 分)。测试结果如图 2 所示。

实验结果分析如下：所有数据都是每个场景经过 20 次试验后得出分数的平均值，分数越高表示与参考版本越接近，满分是 120 分。前三个场景都是同类声源的聚类效果，后三个场景是不同类声源混合的聚类效果，可以明显看出同类声源聚类后的主观听觉效果更好。实验结果表明，在视觉范围内，人对不同类声源聚类的空间误差更加敏感，需要在聚类的过程中将声源的差异性考虑在内，减少不同类声源进行聚类的可能性。

4.2 聚类算法的改进

4.2.1 经典聚类算法

在文献[1]中，为了能够处理大量声音资源，Tsingos 等学者提出了一种行之有效的分层聚类方法。输入原始的声音数据后，该方法会根据声音资源与听者之间的相对位置与相对角度进行聚类。由于每个声源的感知重要性是不同的，他们使用 L 作为权重因子加以区分。声音资源依据聚类规则分组。在这方面，他们将声源 S_k 与聚类代表 C_n 之间的空间信息偏差作为聚类规则。这些空间偏差由两个参数构成，即距离偏差和角度偏差。通过式(2)可以计算出空间偏差。

$$d(C_n,S_k)=L_t^k\left(\beta\log_{10}\left(\frac{\|C_n\|}{\|S_k\|}\right)+\gamma\frac{1}{2}(1-C_n\cdot S_k)\right)\text{。}\quad(2)$$

接着运用文献[6]中提出的聚类方法对每个声源进行处理，产生聚类集合。这种方法在处理大量音源时是有效的。但是，该方法并没有将声音的种类这一因素考虑在算法中。

4.2.2 新提出的聚类算法

1) 预处理

由于所有输入的音频数据类型是已知的，设定一个标签参数 e，可以将音源大体分为环境声，音乐以及语音这三类并分别标注标签参数 e 为 1，2，3。通过将不同声源的 e 进行比较，可以计算出它们之间的相似度。计算相似度的标准化公式：

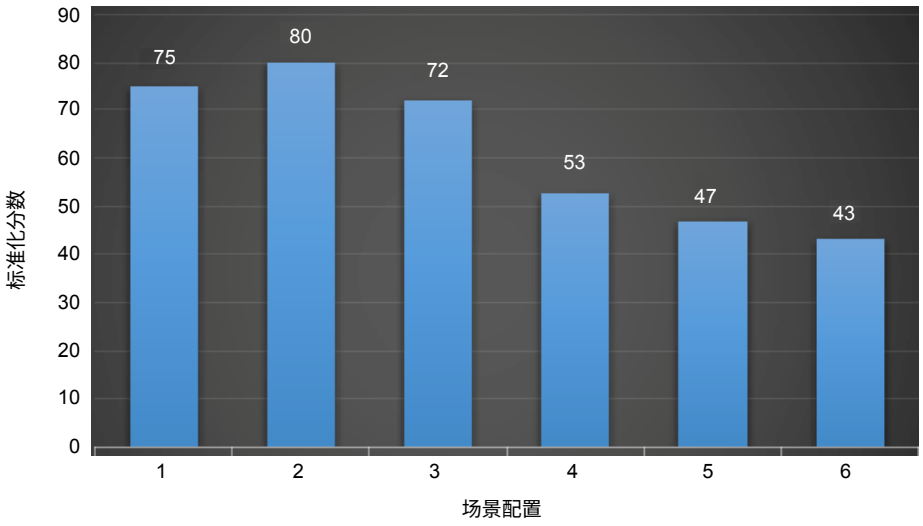


图 2 两种条件对声音渲染的影响

Fig. 2 The influence of two conditions on sound rendering

$$x(S_n,S_k)=\begin{cases}a, & \text{if } e(S_n)=e(S_k)\\ b, & \text{else}\end{cases}, \quad (3)$$

其中：$x(S_n,S_k)$ 用于第 n 个声源与第 k 个声源音频相似度的标准化计算。参数 a 和 b 可以由具体的环境测试得出，仅代表声源相似度对音源聚类的影响权重。$e(S_n)$ 表示第 n 个声源的标签参数数值，$e(S_k)$ 表示第 k 个声源的标签参数数值。

2) 改进聚类公式

$$d(C_n,S_k)=L_t^k\left(\beta\log_{10}\left(\frac{\|C_n\|}{\|S_k\|}\right)+\gamma\frac{1}{2}(1-C_n\cdot S_k)+\mu(1-x(S_n,S_k))\right), \quad (4)$$

其中：μ 表示权重因子，可以通过测试得出。C_n 是以 S_n 为聚类代表的集群。这种新的评价参数可以有效减少感知空间误差，本次实验中的聚类效果如图 3 所示。

图 3 显示的是不同聚类方法的结果，颜色相同的目标被聚为一类。图 3(a)将声源不分类别地加以聚类，使得声源的空间信息更加混乱。图 3(b)采用新型的聚类方法，将不同类别的声源进行分割处理，如人的声音与汽车的噪声进行分离聚类。

5 实验与分析

设置两种实验条件来评估新型空间声音渲染方案：1) 包含大量点声源的外部场景(街道和高速公路场景)；2) 包含声音反射的内部场景(办公室)。所有实验的运行平台为一台包含 Nvidia GTX 970 显卡和 i7 处理器的计算机。另外，声源的采样频率设置为 44.1 kHz，每帧音频的采样点数为 1200 个。表 2 总结了新型空间声音渲染方案在三个场景中的实验表现。

街道场景总计拥有 769 个声源，其中大部分是移动声源。这些声源可分类为行人的脚步声、汽车货车的汽笛声等。基于该场景的实验可以有效评估新型渲染方案在复杂虚拟环境中的表现。高速公路场景包含 893 个声音资源点，由火车声、汽车汽笛声和车载音响声构成。办公室场景中的声音由人的语音、脚步声以及广播声音构成，包含 578 个声音资源。总体而言，第三节提出的算法改进减少了声音引擎 30%~40%的 CPU 占有率。空余出来的 CPU 可以用于其他的任务处理，如视觉渲染、游戏控制等。此外，应用第四节提出的改进算法，通过对参考场景与仿真场景直接进行切换，能清晰地感受到新型渲染方案应用于这三个场景对声音仿真效果的提升。实验数据如表 2 所示。

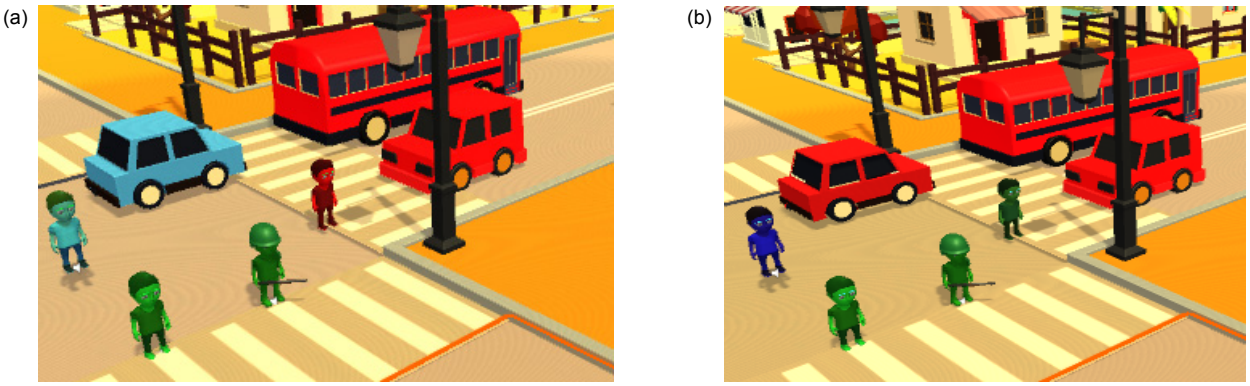


图 3 两种聚类算法的计算结果。(a) 传统聚类算法；(b) 新型聚类算法

Fig. 3 The results of two clustering algorithms. (a) Traditional algorithm; (b) Improved algorithm

表 2 三种场景下算法的性能表现

Table 2 The performance of algorithm in three scenes

环境	声源数量/个	剔除/%	聚类数	聚类时间/ms	帧数/(f·s^{-1})
街道	769	56	25	0.78	30
办公室	578	64	25	0.62	34
高速公路	893	43	25	0.83	27

实验表明，将经过多种方法改进后的空间声音渲染方案应用于实验时，三种场景的实时每秒传输帧数(frames per second, FPS)均较高，满足 VR 系统对声音处理的延时要求。除此之外，数据显示声音渲染过程中聚类这一步骤的平均耗时不足一毫秒，表明该算法具有很高的运算效率。

6 结 论

本文先是引用文献[4]中基于分数阶傅里叶变换的方法抑制了声音采样过程中的量化噪声。其次，在传统空间声音渲染的基础上，提出了一种新型的声音聚类方法，用于解决实时渲染大规模声源的问题。该方法先是从聚类的具体过程入手，将所有聚类结果的平均角度误差作为阈值，用于判断是否应该继续进行聚类运算，降低整体的运算代价。其次，通过主观实验验证了对不同类声源进行聚类会造成更大的空间信息误差这一结果。运用这一结果，本文提出了基于声源标签的方法，解决了不同类声源聚类这一问题。最后通过三个场景实验，验证了新型方法的可行性。

参考文献

[1] Tsingos N, Gallo E, Drettakis G. Perceptual audio rendering of complex virtual environments[J]. *ACM Transactions on Graphics*, 2003, **23**(3): 249–258.

[2] Moeck T, Bonneel N, Tsingos N, *et al*. Progressive perceptual audio rendering of complex scenes[C]//*Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*, 2007: 189–196.

[3] Schissler C, Manocha D. Interactive sound propagation and rendering for large multi-source scenes[J]. *ACM Transactions on Graphics*, 2016, **36**(4): 121–139

[4] Lu M F, Ni G Q, Bai T Z, *et al*. A novel method for suppressing the quantization noise based on fractional Fourier transform[J]. *Transactions of Beijing Institute of Technology*, 2015, **35**(12): 1285–1290. 鲁溟峰, 倪国强, 白廷柱, 等. 基于分数阶傅里叶变换的量化噪声抑制方法[J]. 北京理工大学学报, 2015, **35**(12): 1285–1290.

[5] ITU. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems: ITU-R BS.1116-1[R]. Geneva: ITU, 1994: 1128–1136.

[6] Hochbaum D S, Shmoys D B. *A best possible heuristic for the k-center problem*[J]. *Mathematics of Operations Research*, 1985, **10**(2): 180–184.

[7] Schissler C, Nicholls A, Mehra R. Efficient HRTF-based spatial audio for area and volumetric sources[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2016, **22**(4): 1356–1366.

[8] Takala T, Hahn J. Sound rendering[J]. *ACM SIGGRAPH Computer Graphics*, 1992, **26**(2): 211–220.

[9] Schissler C, Loftin C, Manocha D. Acoustic classification and optimization for multi-modal rendering of real-world scenes[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2017, **24**(3): 1246–1259.

[10] Taylor M T, Chandak A, Antani L, *et al*. RESound: interactive sound rendering for dynamic virtual environments[C]//*Proceedings of the 17th ACM International Conference on Multimedia*, 2009: 271–280.

[11] Raghuvanshi N, Snyder J, Mehra R, *et al*. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes[J]. *ACM Transactions on Graphics (TOG)*, 2010, **29**(4): 142–149.

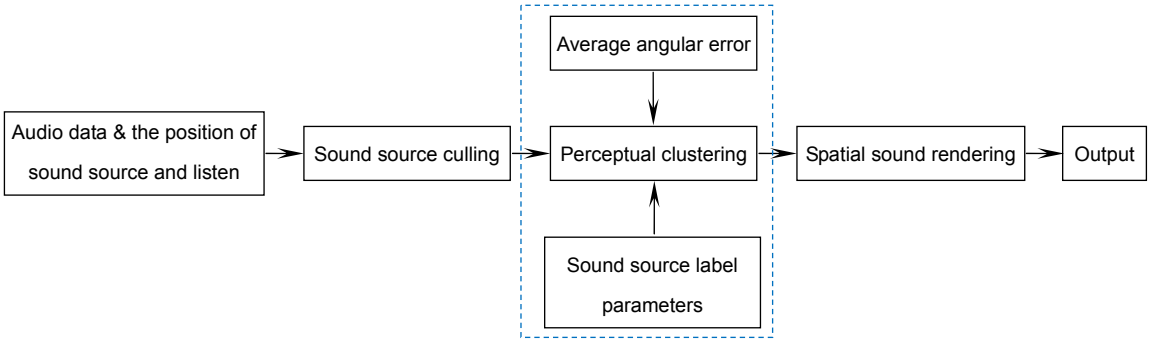
[12] Grelaud D, Bonneel N, Wimmer M, *et al*. Efficient and practical audio-visual rendering for games using crossmodal perception[C]//*Proceedings of the 2009 Symposium on Interactive 3D Graphics and Games*, 2009: 177–182.

[13] Innami S, Kasai H. On-demand soundscape generation using spatial audio mixing[C]//*Proceedings of 2011 IEEE International Conference on Consumer Electronics*, 2011: 29–30.

An improved method to render the sound of VR scene

Chen Tianshi, Tie Yun*, Qi Lin, Chen Enqing

School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China



The process of sound rendering program

Overview: In the field of VR, spatial sound rendering technology plays an increasingly important role. The spatial sense of sound plays a very important role in enhancing the immersive sense of the user in the VR scene. At present, the advanced space sound rendering scheme is mainly divided into two types: the sound waveform and the ray-based tracking algorithm. Recently, important research includes the following points: Tsingos proposed a dynamic spatial sound rendering method based on culling and clustering, which made it possible to process large-scale sound sources in real time. Moeck considered the visual factors during clustering, which reduced the clustering cost. The new algorithm proposed by Schissler eliminated the impact of obstacles on the clustering results. Tao et al. proposed quantization noise suppression method based on fractional Fourier transform to improve the quality of the audio signal sampling. Based on the virtual scene containing hundreds of movable sound sources, due to the high computational cost of clustering stage, the traditional spatial sound rendering scheme often takes up too much computing resources, which has become a bottleneck in the development of VR audio rendering technology. In this paper, we improve the processing speed of sound rendering and the operation efficiency of the entire system by adding the average angle deviation threshold in the clustering step. In addition, we design and implement a perceptual user experiment that validates the notion that people are more susceptible to spatial errors in different types of sound sources, especially if it is visible. Based on this conclusion, this paper proposes an improved method of sound clustering, which reduces the possibility of different types of sound sources clustering. Summarized as follows: focusing on rendering of complex virtual auditory scenes comprising hundreds of moving sound sources using spatial audio mixing, we propose a new clustering algorithm considering average angle error. We presented an effectiveness of our algorithm over specific condition to reduce computational costs caused by frequently clustering. In addition, the result of subjective experiments expresses that the clustering of different types of sound sources will cause more spatial information errors. Using this result, this paper proposes a method based on sound source label parameters, which solves the problem of clustering different kinds of sound sources. In the end, three scene experiments verified the feasibility of the new method.

Citation: Chen T S, Tie Y, Qi L, *et al*. An improved method to render the sound of VR scene[J]. *Opto-Electronic Engineering*, 2018, **45**(6): 170742

* E-mail: ieytie@zzu.edu.cn