虚拟场景中环境声源仿真技术综述

程皓楠 张加万

（天津大学智能与计算学部天津 300350）

摘 要 环境声音作为日常生活中分布最为广泛的一类声音，是人们获取外部信息的重要来源。近十几年来，随着用户对虚拟场景真实度要求不断提升，为虚拟场景打造同步、真实的环境音效已成为构建高度沉浸式虚拟环境不可或缺的一部分。其中环境声源仿真作为打造真实感虚拟环境音效的基石，得到了研究人员的广泛关注与探索。与传统的人工声源仿真相比，通过算法程序合成具有高真实感的环境声音，不仅可以实现随用户视角变化而变化的交互式声音效果，提升游戏、动画和虚拟现实等虚拟场景的用户沉浸感，而且也避免了繁杂的调音、剪辑等工作，极大地降低了人力资源的消耗。本文对应用于虚拟场景中的环境声源仿真方法进行分类和总结，从研究背景、应用领域、基本原理、技术手段等多个角度作了比较全面的综述。首先讨论了环境声源仿真的理论研究价值和应用前景，概述了这一个多学科融合的研究领域。然后，从不同声音模型的基本原理出发，分析了物理声学模型与心理声学模型对于环境声源仿真方法设计、研究内容等方面产生的影响。其次，依据声源模型构造策略不同，从基于物理模型、基于信号模型、基于混合模型以及基于深度学习模型这四大类对现有环境声源仿真方法进行分析，分别讨论了基于每一类模型的环境声源仿真方法的发展趋势、技术创新性与局限性。最后，从算法输入参数、算法效率、仿真结果质量等方面横向比较了不同类别方法的优势与劣势，总结了现有环境声源仿真领域所面临的主要问题与挑战，并对未来研究方向进行展望。

关键词 声源仿真；环境声音；虚拟场景；视听同步；物理建模；信号处理；跨模态

中图法分类号 TP391

A survey on environmental sound synthesis for virtual scenes

CHENG Hao-Nan ZHANG Jia-Wan

（College of Intelligence and Computing, Tianjin University, Tianjin 300350）

Abstract As the most widely distributed type of sound in daily life, environmental sound is an important source for people to access external information. Over the last decade, with users' increasing requirements for the realism of virtual scenes, it has become an indispensable part of building a highly immersive virtual environment for virtual scenes to create synchronized and realistic environmental sound effects. At present, synthesizing realistic environment sound effect for virtual scene needs to go through three simulation stages: environment sound synthesis, sound propagation and sound rendering.Among them, environmental sound synthesis, as the cornerstone of creating realistic virtual environment sound effects, has received extensive attention and exploration by researchers.Compared with the traditional artificial sound synthesis, the procedural sound synthesis can not only generate the interactive sound effect, strengthen the immersion of the users in virtual scenes such as games, animation and virtual reality, but also avoid the complex manual work of tuning and editing, which greatly reduces the labor consumption.Most of the existing environmental sound source synthesis work is to explore the specific environmental sound source synthesis method for specific types of sound sources and specific virtual scenes, but the overall environmental sound source synthesis field has not been reviewed and analyzed. In view of this, this paper makes a comprehensive review of the existing environmental sound source

本课题得到国家重点研发计划(2019YFC1521200)资助.程皓楠，博士研究生，,主要研究领域为计算机图形学、声音仿真.E-mail:tjuchn@tju.edu.cn.张加万（**通信作者**），博士，教授，CCF 会员，主要研究方向为计算机图形学、可视化.E-mail: jwzhang@tju.edu.cn.

synthesis methods used in virtual scene.This paper classifies and summarizes the environmental sound synthesis methods used in virtual scenes, and makes the overview from multiple perspectives such as research background, application fields, basic principles, and technical means. Firstly, westart from the discussions on the theoretical research value and the application prospect of environmental sound synthesis, and summarize the research field of multi-disciplinary integration. Then, due to that the sound source signal perceived by the human ear is not only related to the physical acoustic principle of the sound source, but also affected by the auditory characteristics of the human ear, we analyze the influence of physical acoustic models and psychoacoustic models on the design and research content of environmental sound synthesis methods. Secondly, according to different construction strategies of sound source models, the existing environmental sound synthesis methods are analyzed from four categories: physical model based, signal model based, hybrid model based, and deep learning model based. Among them, the hybrid model based environmental sound synthesis methods combine the sound synthesis methods based on physical model and signal model, while the deep learning model based environmental sound synthesis methods simulate the sound source by constructing the mapping relationship between visual and sound, which is essentially a type of visual-audio cross-modal generation technology.Then, we discuss the development trend, technological innovation and limitations of each different model-based environmental sound synthesis methods respectively. Finally, the advantages and disadvantages of different types of methods are compared horizontally in terms of algorithm input parameters, algorithm efficiency, and simulation result quality, and the main problems and challenges faced by the current environmental sound synthesis field are summarized, and future research directions are prospected.

Key words sound synthesis; environmental sound; virtual scene; audio-visual synchronization; physical modeling; signal processing;cross-modal

1 引言

声音在人类日常生活中无处不在，其中，由周围环境产生的不具有语义信息、不遵循预定义规则的环境声音作为分布最为广泛的一类声音，是人们感受外部环境的重要信息来源[1-3]。近年来，在人类社会信息化、数字化进程不断推进的过程中，环境声音的数字化建模、仿真技术也得到了长足的发展。随着 2016 年虚拟现实产品的爆发式增长，为虚拟场景打造“声临其境”的真实感环境音效成为进一步提升沉浸式体验的关键，而其背后的环境音效仿真技术也受到了越来越多的关注。

在现实生活中，声音从产生到被人耳感知需要经历三个阶段：声源振动产生声音，空间的折射反射以及最终的人耳接收。与现实世界中声音传递过程相对应，为虚拟场景打造真实的环境音效同样需要经过三个仿真阶段（如图 1 所示）：环境声源仿真、空间传播仿真以及人耳感知仿真。其中空间传播仿真阶段主要通过声音渲染技术[4-9]来仿真声音在空间中的折射、反射、衍射效果。人耳感知仿真阶段则需要基于三维音频技术[10-16]来实现双耳渲染和声音空间化效果。由于目前已经有很多关于声音渲染技术和三维音频技术的综述性文章[17-22]，本文将不再对这两类研究工作进行进一步的论述。而环境声源仿真作为打造虚拟环境音效的第一阶段，其主要研究内容是探索声源形状、材质、运动状态等因素对环境声源发声的影响，并对声源处的声音信号进行建模和仿真。由于环境声源的真实度、准确度将直接影响后续两个阶段的仿真效果，因此环境声源仿真作为整个仿真流程的基石，对于打造真实感虚拟环境音效十分关键，具有重要的理论研究价值和广泛的应用前景。

从研究的理论价值来看，环境声源仿真是一个以计算机建模、仿真为核心，多学科融合的研究领域。在进行环境声源仿真的过程中，往往需要涉及物理声学、心理声学、信号处理等多学科内容交叉。由于环境声源不存在明显的周期性，不具有典型的谐波结构，同一物体摩擦、碰撞、敲击等行为可以产生频谱内容迥异的环境声音。因此，对环境声源的振动规律、频率分布、材质特征等声音特征进行探索，不仅有助于研究人员对环境声音的内在声学结构有更加深入的认知，同时也对图形仿真[23]、声音事件识别[24]、声音分类[25]等领域起到促进作用。

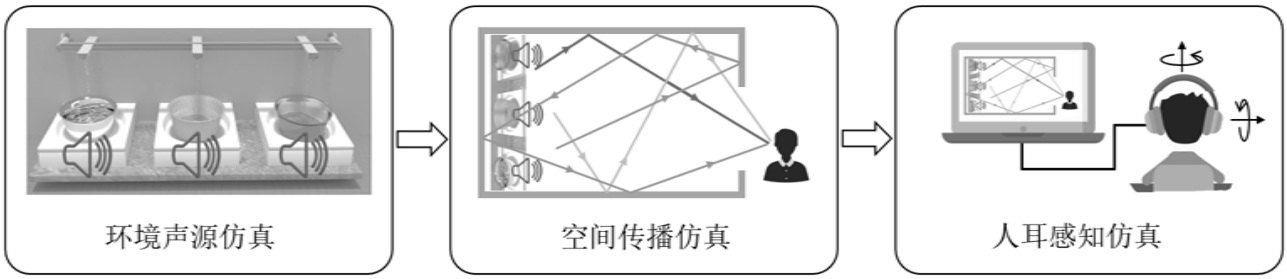


图 1 真实感环境音效的仿真流程

在应用方面，由于声音信号可以有效地补充和增强人们的视觉感知内容，因此，环境声源仿真技术被广泛地应用于动画电影配音[26,27]、计算机游戏音效制作[28,29]以及虚拟现实音效制作[30,31]等诸多领域中。在动画电影制作过程中，由于无法通过同期录音获得与动画场景相对应的环境声音，因此需要拟音师根据动画内容进行繁杂的手工拟音，而程序化的环境声源仿真则提供了更加便捷的声音合成方案。以斯坦福大学和康奈尔大学为代表的动画配音研究取得了丰硕的成果，成功为固体碎裂[32]、弹簧跳动[33]、液体倾倒[27]、火焰燃烧[34]等多种动画场景自动地合成同步并且逼真的固体碎裂声、火焰燃烧声等多种声音。在游戏、虚拟现实这类交互式虚拟场景中，随着声源形状、大小、材质、运动状态等属性变化而不断变化的音效可以使用户感受到声音所带来的虚拟场景中物体的材质感与真实感，从而提升沉浸式体验。近年来，面向交互式虚拟场景的实时声源仿真技术迅速发展，出现了挥剑声[35]、枪击声[28]、脚步声[31]、布料声[29]、海浪声[36]、划船声[30]等一系列应用于不同虚拟场景的环境声源仿真技术。其中，由微软研究院针对游戏中的枪击场景推出的实时枪击声源仿真算法已成功应用于 Xbox360 的 Crackdown II 游戏中[28]。

然而，现有环境声源仿真工作大部分是针对特定声源类别、特定虚拟场景，对具体的环境声源仿真技术进行探索，尚未对整体环境声源仿真领域进行综述与分析。鉴于此，本文对现有应用于虚拟场景的环境声源仿真技术进行较为全面的综述。具体地，本文从环境声源仿真的研究背景、应用领域、基本原理、技术手段等多个角度全方面地对这一研究领域进行了详细的综述。同时结合现有研究文献，深入讨论分析了不同类型的声源仿真技术的创新型与局限性，并从整个环境声源仿真领域的角度分析了当前存在的研究问题和难点，展望了环境声源仿真的未来发展趋势。本文第 2 节介绍了虚拟场景中的环境声源仿真的基本原理；第 3 节重点讨论了基于不同仿真模型的环境声源仿真技术的方法特点，发展趋势；第 4 节从高质量交互式声源仿真困难以及缺乏定量评价指标等方面提出现有环境声源仿真方法的局限性，并对环境声源仿真的未来方向进行展望；最后第 5 节总结全文。

2 环境声源仿真的基本原理

环境声源仿真的本质是对声源运动产生的声音信号进行建模，从而还原出不同的声源信号。而人耳感知到的声源信号，不仅与声源的物理发声原理相关，同样受到人耳的听觉特性影响。因此，物理声学原理和心理声学原理作为声源仿真的两类基本原理，在环境声源仿真过程中起到重要作用。本节将对现有环境声源仿真方法中所涉及到的物理声学与心理声学基本原理进行概括介绍，并分析不同声学原理对环境声源仿真的方法设计、研究内容等方面产生的影响。

2.1 物理声学原理

从广义看来，声学现象的实质是传声媒介（气体、液体、固体等）质点所产生的一系列力学振动传递过程的表现。因此对于固体的复杂振动发声，最简单的处理方式是将固体振动分解成一组弹簧质点振动，通过求解并叠加每一个质点振动得到最终的振动模型。具体来说，弹簧质点的阻尼运动方程为[37]：

$$m\ddot{x}+c\dot{x}+kx=F(t) \quad (1)$$

其中 x 表示质点的位移，m 是质点的质量，c 是阻尼大小，k 是弹簧的刚度，$F(t)$ 表示随着时间变化的外力。求解上述运动方程可以得到质点随时间变化的振动方程（即质点振动产生的声音）：

$$x(t)=ae^{-\beta t}\cos(2\pi ft) \quad (2)$$

其中 a 为振幅，β 是阻尼系数，f 是质点的振动频率。在早期固体声源仿真方法中[38,39]，由于受运算

能力的限制，大多声源仿真模型采用弹簧质点振动模型，通过将固体近似为大量的离散质点的集合，来仿真固体碰撞、滚动等运动产生的声音。显然，基于质点振动模型的声源仿真方法无法有效仿真声源大小、形状的改变导致的声源信号的变化。因此，质点振动原理仅适用于简单场景的声源仿真。

在现实生活中，声源的振动系统质量往往在空间中存在连续分布，并且一部分质量还包含弹性和阻尼性质，此时则无法假设声源的质量集中在一点，也不能将其视为几个离散的点，这类声源通常被称为弹性体。在构造弹性体振动方程时，不仅需要时间变量来描述其运动状态，还需要引入表示空间位置的变量。以弦的振动方程为例：

$$\frac{\partial^2 \eta}{\partial x^2} = \frac{\delta^2}{T^2}\frac{\partial^2 \eta}{\partial t^2} \quad (3)$$

其中 x 为振动位置，η 为弦离开平衡态的垂直位移，T 为张力，δ 是弦的线密度，通过对上述二阶偏微分方程求解，可以得到弦上不同位置的振动规律。在物理学中，常见的弦、棒、膜、板等形状的弹性体的振动方程已经被广泛研究，在本节中，我们将不再介绍这些模型的细节，请读者参考书籍[37,40-42]。在弹性体振动发声过程中，有很多影响弹性体振动方式的因素，例如图 2 中形状不同但材质相同的弹性体掉落到地面时，产生的声音往往差异很往大[43]。在诸多影响振动发声的因素中，弹性体的形状、材质，外力作用的位置以及外力的大小作为四个显著影响声音变化的因素[44]，成为目前环境声源仿真的主要研究内容。

尽管弹性体的振动方程为其声源仿真提供了理论基础，然而在实际声源仿真过程中，往往很难直接基于弹性体振动方程进行仿真，其主要原因在于模拟声源的可听运动比传统的图形仿真需要更高的精度。由于人耳能够听到高达 20kHz 的运动，远

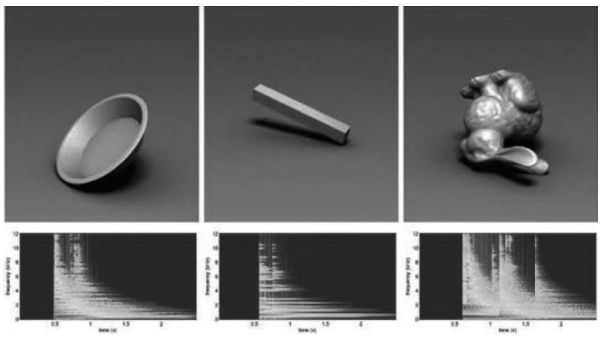


图 2 不同形状弹性体及其掉落产生的声音时频图[43]

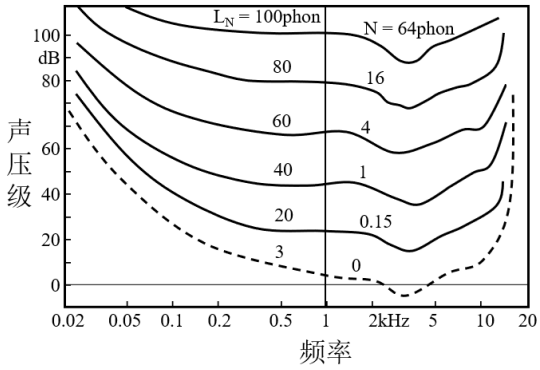


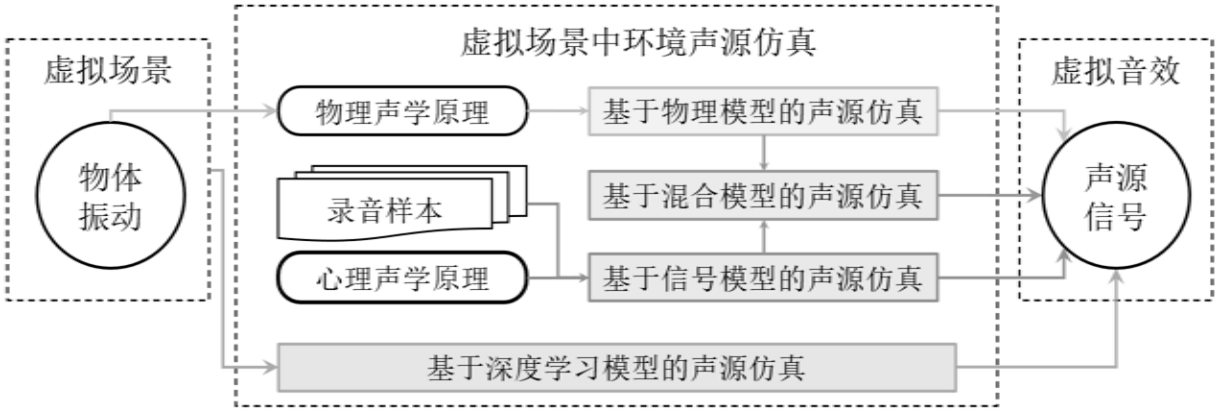
图 3 等响度曲线图，其中 N 为响度，LN 代表响度级（loudnesslevel）[46]

高于可视运动频率范围（例如可视波长范围为 3.8×10^{-7}m~7.4×10^{-7}m，可听波长范围为 0.017m~21m[45]）。因此若想在声音仿真过程中充分采样高频信息，则图形仿真积分器使用的时间步长不能大于 10^{-5} s。然而，过小的时间步长会极大地增加图形仿真时间，所以目前的图形仿真技术通常会忽略这些不可见的运动来保证高效地仿真。因此，如何平衡声学振动模型和图形仿真之间的参数不对等，成为基于物理声学原理的声源仿真方法设计过程中要解决的一个核心问题。本文将在第 3.1 节中详细讨论不同声源仿真方法的构造策略与解决方案。

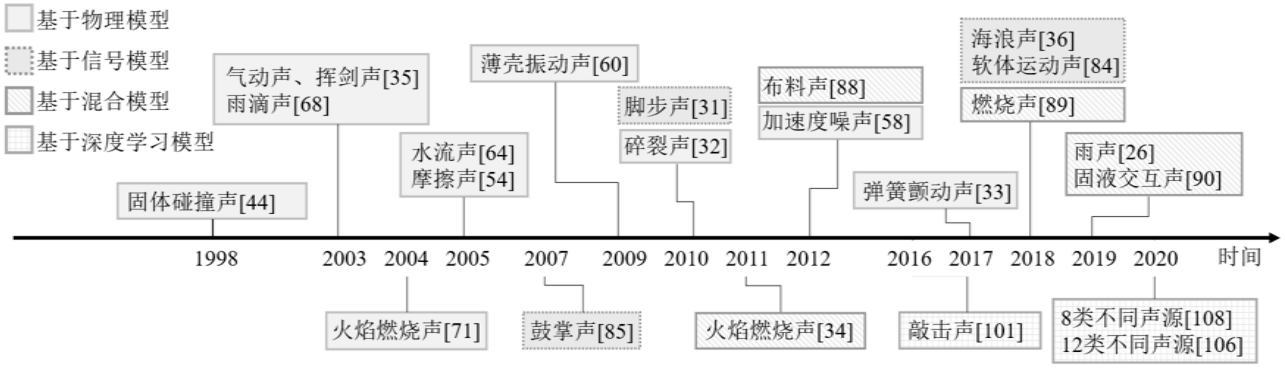
2.2 心理声学原理

心理声学（psychoacoustics）是研究声音和与之对应的人耳感知的一门学科，即探究“人脑解释声音的方式”。不同于物理声学中振幅、频率、阻尼等物理参数对声音进行客观描述，心里声学参数往往是定量地刻画人们对于声音某一特征的主观听觉感受。常见的心理声学参数包括响度（loudness）、尖锐度（sharpness）、抖动强度（fluctuationstrength）、粗糙度（roughness）以及 A 计权声压级（A-weighting sound pressure level）等[46]。以响度为例，其计量单位为 song，1song 大小的定义为声压级为 40dB 的 1kHz 纯音的响度值。这一心理声学参数反映了人耳对于声音信号强弱的感知，从图 3 中我们可以观察到，人耳对于声音强弱的感知不仅与声音真实的声压级相关，同时还受声音频率影响。如图 3 中虚线所示，对于 70dB 的 20Hz 声音与 0dB 的 2kHz 声音，尽管二者的声压级相差 70dB，但是人耳感知到这两组声音的强弱是相同的。此外，人耳对于不同频率声音的敏感度也存在差异，1933 年，贝尔实验室的 Fletcher 和 Munson[47]指出，人耳的响度感知是非线性的，在

20Hz 到 20kHz 的人耳感知范围区间



(a) 环境声源仿真方法整体框架



(b) 环境声源仿真方法时间轴

图 4 环境声源仿真方法框架与时间轴

中，人耳对声音的中频信息（3~4kHz）最敏感。在响度计算方面，目前公认的响度计算标准包括 DIN45631/A1[48]和 ISO 532-1:2017[49]。计算不同声源信号的响度与计算不同声源信号的声压级相比，前者往往可以更加准确地衡量人耳对于声源信号强弱的判断。总的来说，心理声学参数提供了更加符合人耳感知规律的定量描述，在提升人机交互精确度[50]、语音识别准确度[51]、音乐合成质量[52]等方面都逐渐展露其优越性。

而在环境声源仿真方法设计过程中，研究人员基于心理声学原理，可以直接从声源信号的感知相似性出发，通过设计具有人耳感知相似性的信号模型，从而避免了复杂的物理解算过程。总的来说，心理声学原理为声源仿真提供了另一种解决思路，有效地利用心理声学原理，不仅可以合成更加符合人耳听觉感知的声音，同时也可以对人耳听觉不敏感的声音内容进行简化运算或省略。尽管基于这一仿真思路构造的声源模型往往并不遵循准确的物理原理，但是其仿真结果在感知上合理，并且声源模型的求解过程往往更加便捷、高效。本文将在 3.2 节对这类声源仿真方法进行详细分析与讨论。

3 虚拟场景中环境声源仿真技术讨论与分析

随着物理声学、心理声学理论研究的不断深入，以及信号处理、深度学习等技术的发展，目前声源仿真模型逐渐发展为四类：基于物理声学原理来仿真声源的物理模型；基于心理声学原理来仿真声源的信号模型，将二者组合的混合模型以及近几年兴起的通过构造视听映射关系来仿真声源的深度学习模型（如图 4(a)所示）。图 4(b)按照时间顺序，列出了本节所综述的代表性方法。从图 4(b)中可以看出，早期声源仿真方法以基于物理建模的方法为主，随着人们对声音细节和声音真实度的需求不断提升，基于心理声学原理的信号模型被逐渐提出。近几年，通过组合不同模型来获得算法效率与算法精度平衡的混合模型，以及直接学习并构造视听映射关系的深度学习模型成为目前的发展趋势。本文将分别介绍基于每一类模型的声源仿真方法的发展趋势，并讨论其技术创新性与局限性。

3.1 基于物理模型的声源仿真技术

基于物理模型的声源仿真技术是声源仿真领域研究最早，也是目前使用范围最广的一类方法。这类方法基于振动声学原理，将声源信号作为物体振动的结果，通过对声源振动过程进行建模，构造出声源物体的振动模型，最终合成与振动过程相对应的声音。由于在这类方法中，声源信号可以被构造成物理表达式，因此基于物理模型的声源仿真技术具有高度灵活性，能够实现与视觉场景的细粒度同步，为固体碎裂[32]、流体运动[27]等复杂声源仿真场景实现了高度同步的声源变化效果。本节将依据声源形态的不同，从固体声源仿真、液体声源仿真以及气体声源仿真三个方面对基于物理模型的声源仿真进行介绍与分析。

3.1.1 固体声源仿真

由于固体的形状和运动相对于液体和气体更加规则，所以基于物理模型的固体声源仿真是研究人员最早展开探索的领域。尽管在物理学中，固体碰撞、摩擦、滑动等行为的物理发声机制已经得到了详尽地研究[37,40]。然而，这些振动声学模型很难直接应用于面向虚拟场景的声源仿真技术中。这主要由于高精度的物理模型往往需要复杂的物理求解过程，从而无法满足虚拟场景的实时性需求。为了实现实时仿真，Doel[44]基于质点弹性振动模型，首次为虚拟现实和计算机游戏中固体的撞击和滚动合成与之同步的声音。具体来说，对于不同形状的固体，该方法将每一个固体表面分解为小的网格面片，通过求解并叠加每一个网格节点 $X_i, i=1\cdots n$ 的弹性振动，构造出整个固体的振动方程：

$$M\ddot{X}+C\dot{X}+KX=F(t) \quad (3)$$

其中 X 是大小为 $3n\times1$ 的位移向量，对应每个节点在 xyz 三个方向上的位移变化，M，C，K 分别对应大小为 $3n\times3n$ 的质量矩阵、阻尼矩阵和刚度矩阵。通过将振动系统解耦成一组独立的一维振动方程，进而求解出整个固体的振动方程。随后，Doel 和 Pai[53]进一步探究了固体碰撞过程中形状对于声音的影响，提出了一种模拟固体碰撞声音合成的通用框架。该框架以固体的振动动力学为基础，可以计算出不同材料、形状的固体在不同碰撞点位置的碰撞声音。尽管这种将固体表面拆分成若干网格面片，并通过线性模态振动叠加的声源仿真方法可以实时地合成固体碰撞、滚动的声音，然而将固体振动简化为单一的弹性振动显然是不符合真实物理原理的。因此线性模态振动叠加方法无法准确仿真布料、纸张、金属薄壳这类可变形固体声源。

为了实现可变形固体声源仿真，O'Brien 等人[38]基于可变形固体模拟器对固体表面振动进行仿真，通过非线性有限元方法来模拟固体的运动并且分离出与可听频率相对应的表面振动参数，通过计算每个网格面的法向速度和平均速度来求解固体表面振动方程。由于该方法采用了非线性有限元方法，因此可以模拟由非线性行为(如弯曲运动)产生的声音，但是，大量的非线性计算使得声源仿真算法无法达到实时。因此，为了提升算法效率，O'Brien 等人[39]进一步提出了一种基于刚体模拟器的二阶段实时固体运动声音合成方法。由于该方法通过对固体变化过程中的形状和频率进行数值预计算，因此可以直接利用标准刚体模拟器生成的接触力数据交互地合成声音。此外，在计算求解方面，由于四面体网格中每个节点只与很少的其余节点有关联，所以弹性振动方程中的矩阵均为稀疏矩阵，通过对稀疏矩阵进行特征分解，可以高效地计算大网格的变形模态。

总的来说，线性模态振动叠加和非线性有限元这两类早期的基于物理模型的固体声源仿真技术，分别从振动模型简化、振动参数预计算两个方面进行了探索，都成功地实现为交互式虚拟场景仿真固体声源。然而，在上述对于固体声源仿真的早期探索中，实验对象通常为像球体、立方体这类几何复杂度较低的固体。而对于形状更加复杂的固体，则往往需要针对固体形状特征，建立与形状特征更加匹配的声音模型。此外，早期固体声源仿真方法所对应的场景也通常为碰撞、滚动这类较为单一的运动场景，对于复杂运动场景中特有的一些声学现象，则需要对声音模型进一步细化。因此，后续大量的研究在非线性运动声源仿真和复杂形状声源仿真两个角度，对固体声源仿真技术进行提升。

(1) 非线性运动声源仿真

固体在断裂、破碎、摩擦等更为复杂的运动场景中，其声源振动往往很难通过单一振动模型进行刻画。因此对于非线性运动场景，无论是声源振动模型的构造过程还是求解过程，都对算法效率提出了很大的挑战。

以摩擦场景为例，由于摩擦现象是一种强非线

表 1 代表性的基于物理模型的不同固体声源仿真方法

改进方向	代表工作	适用固体类型	适用运动类型	仿真时间
非线性运动声源仿真	Zheng 和 James[32]	刚体	断裂、破碎	接近实时
	Chadwick 等人[58]	刚体	加速度突变的撞击	非实时
	Chadwick 等人[59]	刚体	加速度突变的撞击	非实时
	Avanzini 等人[54]	简单虚拟环境中的刚体	摩擦	实时
	Ren 等人[56]	复杂虚拟环境中的刚体		实时
	Zheng 和 James[57]	刚体与可变形固体		非实时
复杂形状声源仿真	Chadwick 等人[60]	薄壳类	碰撞	接近实时
	Cirio 等人[62]	刚体	碰撞	非实时
	Cirio 等人[61]	薄片类固体	压皱	非实时
	Schweickart 等人[33]	细长形状固体	振动、抖动	实时

性耦合，所以即使在固体交互过程中只包含了很少量的接触模态，也会产生十分丰富的声音现象，因此基于传统的摩擦力物理模型很难实现实时的摩擦声源仿真，从而无法应用于交互式虚拟场景中。针对这一问题，一种解决思路是寻求更为简化的振动模型作为近似表示，如 Avanzini 等人[54]基于一种更为简单的单状态模型[55]作为摩擦力的近似表示形式，将刚体位移分解为弹性和塑性分量，并通过对振动模型离散化进行高效求解。然而该方法需要调整大量的控制参数才能获得令人信服的滑动声音效果，因此很难应用于涉及许多具有不同物理属性的复杂虚拟场景中。另外一种解决思路是将振动模型分解，通过分别构造不同的振动模型实现复杂非线性模型的简化求解。Ren 等人[56]基于这一思路，依据物体形状、可见表面凹凸度和微观粗糙度设计了三层表面表示法，建立了不同分辨率下的表面接触模型，可以实时地生成丰富、复杂的接触声音。

然而，不论是构造非线性摩擦模型的等效替代，还是将非线性模型分解后逐层构造，都会将固体在接触过程中包含的丰富的非线性接触事件，例如微碰撞、颤动等微小振动遗漏掉。尽管这些接触级微小振动在视觉上不会产生较大影响，但是却对声源仿真起着重要作用。而忽略这些接触级振动往往会使声音结果产生伪影，从而降低仿真结果的真实度。Zheng 和 James[57]针对这一问题，设计了一种基于摩擦多体接触模型的摩擦声源仿真方法来合成更高质量、高精度的接触声音。尽管这一方法通过模式自适应异步积分器优化了计算效率，却仍然无法有效解决模型精度提升所带来的大量计算问题，因而无法达到声音的实时反馈。

而其它类型非线性运动声源仿真同样面临这一问题。例如，在固体断裂、碎裂声音仿真中[32]，尽管通过预计算不同大小的椭球声音模型并构造声音数据库，使得其计算效率相比于纯物理求解，可以提升近 500 倍的运算速度。然而由于碎裂场景中碎片数量众多，因此该方法的算法效率仍然无法达到实时。同样通过预计算方案，对于固体加速撞击场景，Chadwick 等人[58,59]基于预计算的压力场来简化撞击模型的求解过程，但是由于声源数目众多，并且每个声源均需要独立求解，因此仿真效率并不理想。

从目前来看，非线性运动声源仿真最大的难点在于如何寻找到一种合适的途径来有效平衡物理模型精度与模型求解效率。至今为止，仍有很多非线性运动场景的声源仿真方法尚待探索。为非线性声音模型构造简化的等效替代模型，以及模型分解等方向，值得研究人员做进一步探索。

(2) 复杂形状声源仿真

在现实场景中，声源大多具有不规则形状，因此早期声源仿真过程中将声源视为质点的假设无法适用于形状复杂的声源仿真计算。虽然物理学中对于弦、棒、膜、板等弹性体的非线性振动进行了广泛的研究，然而无法高效求解这类复杂形状声源产生的非线性振动是虚拟场景中声源仿真的一个主要瓶颈。

常见的一种解决方案是通过对振动模态进行预计算，从而提升交互时的声源仿真效率。基于这一思路，Chadwick 等人[60]基于非线性薄壳力模型，通过预先计算数千种振动模态并构造声传递图的方案来存储不同模态之间的耦合方式。虽然这种基于声传递图的预计算缩短了交互运算时的声音计算时长，但是往往需要几十到几百兆的内存空间来

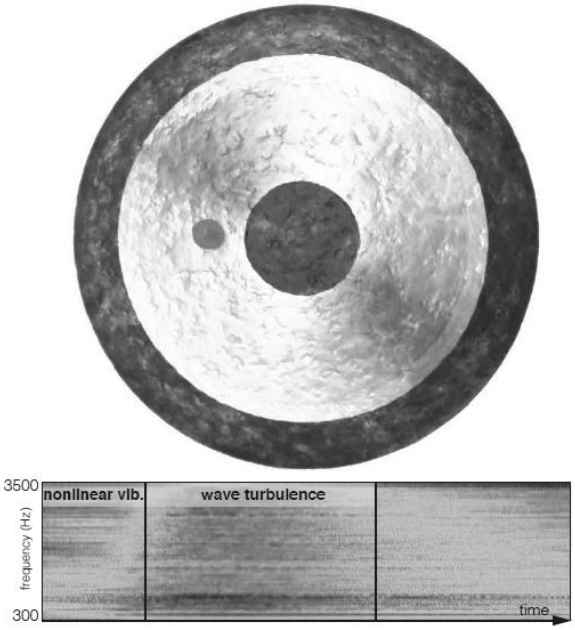


图 5 薄壳类固体声源发声的波湍流现象示意图[62]

存储声传递图，因此很难推广到音频存储空间受限的游戏应用中。

此外，与非线性运动声源仿真类似，通过对复杂形状声源的振动模型进行分解以及构造等效替代模型，同样可以提升声源仿真效率。其中，复杂形状声源的振动模型分解往往需要针对声源特点，设计更具有针对性的模型分解策略。例如对于塑料袋、锡箔纸这类薄片，其产生褶皱声音的弯曲事件在空间尺度上各不相同，并且在压皱的过程中存在的大量视觉上难以察觉但是却可以听得到的弯曲事件。基于这一声音特性，Cirio 等人[61]将声源振动按照空间尺度进行划分：视觉可感知的弯曲振动模型，以及视觉上难以察觉但是却可以听得到的弯曲振动模型。通过检测薄片曲率的突变来构造视觉上可感知的弯曲振动模型，并采用基于物理的幂律分布模型来近似仿真无法检测的弯曲事件的振动声音。然而，尽管该方法通过网格动态划分为近刚性网格块来减少重复计算，由于薄片变形时往往包含几千种模态，因此仍然无法达到实时运算。而对于垃圾桶、锣这类金属薄壳类声源，这类形状固体受到撞击时会产生特有的非线性声音现象：波湍流，即声音的能量随时间从低到高的频率扩散现象（如图 5 所示）。针对这一声音特点，Cirio 等人[62]设计了另一种基于频率信息的模型分解策略，将薄壳的非线性振动分解成两个尺度：产生低频声音的大幅度振动和产生高频声音的小幅度振动。对低频部分进行完全非线性振动仿真，而对包含模态信息更多

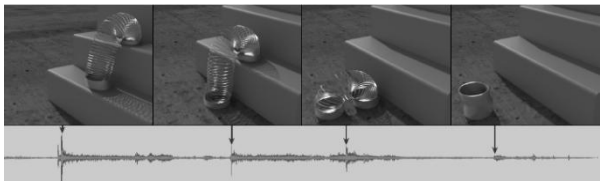


图 6 弹簧类声源仿真结果示意图[33]

的高频部分，则通过线性振动近似。然而在算法效率方面，虽然该方法相比于完全非线性求解，可以提速几十倍，但是仍然无法达到实时运算。

在构造等效替代模型方面，Schweickart 等人[33]探索了如何仿真弹簧、铁丝网围栏这类细长形状固体声源。该方法基于一种简化的棒动力学模型，有效地为高度可变形的杆状固体产生了近似的声音（如图 6 所示）。由于该方法对固体的几何形状做出特定的假设，因此极大地提升了运算速度，确保了实时运算。然而该声音合成模型无法调节碰撞时的接触刚度，因此存在部分合成声音与真实录音有较大偏差的现象。

我们按照改进方向对现有方法进行总结，如表 1 所示（其中实时仿真时间为毫秒级）。总的来说，复杂形状声源仿真同样面临无法有效平衡物理模型精度与模型求解效率的问题，尽管通过构造等效替代的棒动力学模型实现了特定场景的声源实时仿真[33]，但是对于大部分复杂形状声源，如何构造高效的替代模型仍然需要进一步研究。而基于预计算和模型分解的声源仿真技术，尽管其算法效率相比于完全非线性求解实现了较大的提升，然而仍然需要较长的运算时间，无法满足虚拟场景的实时反馈需求。

3.1.2 液体声源仿真

与固体声源相比，液体声源的形状更加不规则，同时其运动状态的随机性也更强，因此液体声源仿真难度更高，与虚拟场景的同步也更加困难。在为流体动画合成同步的真实感液体声音时，主要面临两方面问题：（1）参数信息不对等：气泡振动声音作为液体声音的主要来源之一，其中的高频率气泡声音往往需要精细的气泡几何结构信息，然而流体内部气泡振动是一种微观振动，因此，很难通过现有的流体仿真技术提取到气泡振动声音模型所需的参数；（2）算法复杂度高：流体中的气泡数量往往在百万级，数目庞大的振动模型解算往往很难实现算法效率的实时性，从而无法将液体声源仿真算法应用于交互式虚拟场景中。因此，尽管在 1933 年，Minnaert[63]就已经提出基于水中空腔气泡

表 2 基于物理模型的不同液体声音合成方法对比

方法	视听同步参数	算法效率	适用场景
Doel[64]	气泡半径和气泡到水面距离	实时	未与流体场景集成
Zheng 和 James[27]	气泡数量、气泡半径和空间位置	非实时	小型流体仿真场景
Moss 等人[65]	气泡数量、气泡几何形状	浅水场景实时	大型浅水场景和小型流体仿真场景
Langlois 等人[66]	气泡数量、气泡几何形状和空间位置	非实时	小型流体仿真场景
Zita[68]	液滴体积、液滴撞击速度以及撞击表面材料属性	非实时	单个雨滴撞击和下雨场景

振动产生声音的液体声学模型，直到 2005 年，一种基于空腔气泡振动模型的液体声源仿真技术才被 Doel[64]首次提出。为了提高算法效率，该方法基于“气泡声在气泡离水面更近的位置音调更高”这一实验观测事实，将气泡振动频率随时间的变化规律通过线性方程进行近似计算。尽管这一液体声源仿真方法通过对气泡振动频率的简化，实现了液体声源的实时仿真，但是却由于频率的过度简化，造成了仿真液体声音音色与真实液体声音的较大偏差。

针对如何提升液体声源音色真实度这一问题，后续研究从气泡的运动状态、气泡的几何形状两方面来解决该问题。Zheng 和 James[27]通过模拟气泡的夹带、平流、振动和辐射等不同运动状态实现了液体声源仿真精度的提升。同期 Moss 等人[65]则通过处理不同类型的气泡(球形和非球形)以及不同类型气泡之间的相互作用，从而提升液体声源音色真实度。然而，上述方法均对流体中真实气泡夹带过程的预测能力有限，并且依赖于特定的随机模型来估计基于粒子的气泡生成率和大小分布，因而最终的仿真液体声源音色真实度仍然存在一定的局限性。为了进一步提高算法精度以及液体声源音色真实度，Langloi 等人[66]基于复杂非球形气泡模型[67]，对单个气泡的几何形状进行估计和跟踪，并在亚毫米长度尺度上建模了气泡的夹带、合并、分裂和爆裂过程。与基于球型气泡模型和简单非球形气泡模型相比，这种基于复杂非球形气泡模型的液体声源仿真方法可以计算出更加准确的气泡振动频率值。虽然该方法通过气泡分摊求解算法进行了加速计算，但是对于气泡数目较大（十万以上）的流体场景，合成 4.5s 的声音片段仍需要 20 个小时求解气泡频率。因此，如何提升液体声源仿真速度，仍然是基于物理模型的液体声音合成方法亟待解决的问题。

液滴撞击模型作为另一类液体发声模型，主要用于雨场景中的雨声仿真。早期的雨声仿真工作[68]将单个雨滴看作独立声源，根据雨滴下落过程中动能与势能的能量转换，合成了单个雨滴撞击声音，并通过线性叠加来合成整个下雨场景声音。显然，对于通常包含几十万个雨滴的下雨场景来说，这种单雨滴叠加仿真技术很难实现雨声的高效仿真。此外，不同雨滴声音之间存在较大的音色相似性，将每个雨滴视为独立声源并分别仿真会带来大量的冗余计算，因此后续对于雨声的仿真，则不再完全基于液滴撞击模型进行仿真，而更多的是采用信号分解的方式或结合统计模型进行更加高效的计算。

总的来说，目前液体声源仿真主要基于气泡振动模型，通过求解液体中气泡振动的频率来仿真液体声音。对于液体的初始撞击模型仅进行了初步的探索，在模型准确度、模型高效求解等方面仍然具有很大的提升空间。此外，目前基于物理模型的液体声源仿真技术，更多的关注于如何在视听参数信息不对等的情况下，合成真实度、同步度更高的液体声音，在如何降低算法复杂度方面尚未展开具体的探索。表 2 对现有方法中所采用的声学模型、视听同步参数、算法效率以及适用场景进行了总结。从表中我们可以观察到，目前的液体声学模型主要适用于小型流体场景，这是由于现有方法很难在保证声音细节多样性的情况下使得算法效率达到实时。而在合成声音质量方面，基于物理模型的液体声源仿真与真实水声录音仍然存在一定的真实度差距。这是由于现有的液体声学模型仍然未包含很多影响液体声音的因素，例如液体中气泡可以相互影响产生频率耦合振动，从而改变它们的频率分布，使得单个气泡能够以多个频率辐射声波，但是这种多频气泡的计算则会进一步加大计算的开销。因此如何构造气泡频率耦合模型，如何高效地求解多频率气泡辐射，这些问题都需要进一步探索。

3.1.3 气体声源仿真

气体通过气流运动而产生声音，这种由气流运动产生的空气动力声音通常是刮风、挥剑、火焰燃

表 3 基于物理模型的不同气体声音合成方法对比

方法	声源类型	适用场景	算法效率
Dobashi 等人[35]	空气动力声音	挥剑、风吹过玻璃	实时，但需要大量的离线预计算
Dobashi 等人[71]	火焰燃烧声音	火焰燃烧	实时，但需要大量的离线预计算
Selfridge 等人[73]	空气动力声音	螺旋桨转动	实时
Selfridge 等人[72]	空气动力声音	挥剑	实时

烧等场景声音的主要来源[69,70]。尽管在物理学中，通过求解可压缩的 Navier-Stokes 方程来对气流的微小波动进行数值分析，可以合成与之对应的空气动力声音。然而，这种数值分析往往需要高昂的计算开销，这是由于与大气压力相比，气流波动的数量级大小是10^{-5}，所以往往需要密集的网格模型和非常小的时间步长来捕捉这种微小的气流波动，因此很难基于高精度的物理模型直接求解空气动力声音。

为了实现高效的气体声源仿真，一种解决方案是通过预计算声音纹理来提升交互计算时的算法效率。基于这一思想，Dobashi 等人分别为空气动力声音[35]和火焰燃烧声音[71]设计了实时的声源仿真算法。具体来说，这类方法首先在预处理过程中为空气动力声音和旋涡声创建声音纹理。随后在用户交互的过程中，根据气体运动状态，对预先计算好的声音纹理进行重采样，最终通过叠加各个小声源实现空气动力声音和火焰燃烧声音的实时仿真。然而，这类方法存在两个主要问题：（1）需要针对不同场景进行大量的离线预计算过程，从而很难应用于场景多变的游戏和虚拟现实应用中；（2）叠加声音纹理的本质是对真实物理现象的近似求解，因此最终生成的声音缺少高频细节，与真实录音存在一定差异。

另一类高效声源仿真方案则通过对声源模型分解并分别构造简化的气体声源模型来实现实时仿真。这一方向的代表性工作是 Selfridge 等人[72]设计的实时挥剑声音仿真方法。在该方法中，有八个紧凑的声源被用来模拟挥剑场景中产生的声音（如图 7 所示），通过为每个声源建立并求解风动声音模型得到最终的挥剑声。基于同样的思路，Selfridge 等人[73]将螺旋桨叶片转动声源分为两部分：负载噪声和旋涡声音，通过对每一类声源分别建模，实现了实时的螺旋桨转动声源仿真。然而上述方法在声源仿真过程中进行了大量的数值近似，从而导致了合成声音细节的缺失。此外，空气动力声音往往受多方因素影响，以螺旋桨转动声音为例，真实的螺

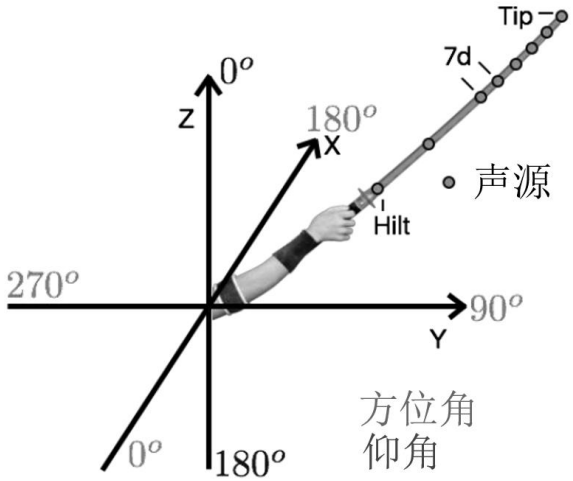


图 7 挥剑声源分布示意图[72]

旋桨飞机产生的声音往往受飞机的电机型号、叶片设计等因素影响，因此基于物理模型仿真生成的螺旋桨声音与真实录音相比，仍存在一定的差异。

表 3 从合成声源类型、适用场景以及算法效率三个方面对现有基于物理建模的气体声音合成方法进行了汇总。从目前来看，虽然现有的基于物理模型的气体声音合成方法通过不同的简化算法实现了声音的实时仿真，然而生成的气体声音往往包含音色细节较少，声音真实度较低。其主要原因是由于气体运动的随机性更强，声源模型复杂度更高，因此基于简化的气体声源模型进行仿真对声音真实度的影响更大。此外，对气体声源的仿真往往还需要考虑到气流中物体的运动、物体对气流的遮挡作用等因素，因此现有工作中完全基于物理模型与其他模型混合使用，从而实现更高精度的气体声源仿真。

3.2 基于信号模型的声源仿真技术

基于信号模型的声源仿真技术不再关注声音产生的物理机制，而是将声源信号视为由多种信号组成的混合信号，通过将不同的信号叠加组合，从而重构出具有相同感知效果的声源信号。在这类方法中，声源建模与场景建模是相互独立的（如图 8 所示），因此，为了实现声音与虚拟场景中物体运动

表 4 基于信号处理的环境声音合成方法对比

方法		适用场景	仿真时间
基于经验公式的声源仿真	Verron 等人[74]	简单的固体碰撞、液体流动、气体运动场景	实时
	Verron 和 Drettakis[77]	火焰燃烧、刮风和下雨场景	实时
	Conan 等人[78]	固体滚动	实时
	Lejemble 等人[79]	纸张撕裂	实时
	Nordahl 等人[31]	脚步声	实时
基于录音样本的声源仿真	Cardle 等人[80]	无明显时序特征的场景	实时
	Picard 等人[82]	固体运动	实时
	Schreck 等人[83]	纸张运动	实时
	Su 和 Joslin[29]	软体运动	非实时
	Su 和 Joslin[84]		实时
	Peltola 等人[85]	鼓掌	实时
	Wang 和 Liu[63]	海浪	实时
	Cheng 和 Liu[59]	固液交互	实时

的同步，往往需要对组成声源的信号进行参数化处理，以确保视听映射模型的设计。

目前获取组成声源的信号主要通过两种方式：构造经验公式以及录音采样。其中，基于经验公式的声源仿真方法生成的声源信号具有丰富的参数信息，但是构造准确的经验公式十分困难。与经验公式相比，录音样本往往更加方便获取，但是其可以提供的参数有限，视听同步过程具有更大的挑战。本节接下来将详细分析与讨论基于这两类信号的声源仿真技术特点和适用范围。

3.2.1 基于经验公式的声源仿真

基于经验公式的声源仿真方法，其本质是通过构造数学经验方程作为物理声学模型的等效替代，从而避免了复杂的物理建模与求解。这类方法在乐器声音合成领域有着广泛的应用[75,76]，这是由于音符信号的频率分布更加规则，从而更容易通过信号叠加、滤波等方式近似合成。而环境声音的频率分布往往不具有明显特征，因此很难设计出理想经验公式作为真实物理模型的等效替代。早期基于经验公式的声源仿真主要面向具有一定音色相似性的自然界声音。例如下雨、刮风这类虚拟场景，Verron 等人[74,77]构造了三种短脉冲信号、两种连续噪声信号，将虚拟场景中的高级描述符（例如水流速度、风速）作为声源控制参数，来变换、组合、叠加这五类基本信号，实现了实时的同步声源仿真。然而，由于这五类基本信号的结构较为简单，因此合成的声音缺乏音色细节，与真实场景声音仍有一定的差异性。此外，对五种基本信号简单的叠加组合与真实的物理组合存在较大差异，因此合成的声音往往存在明显的分离感。

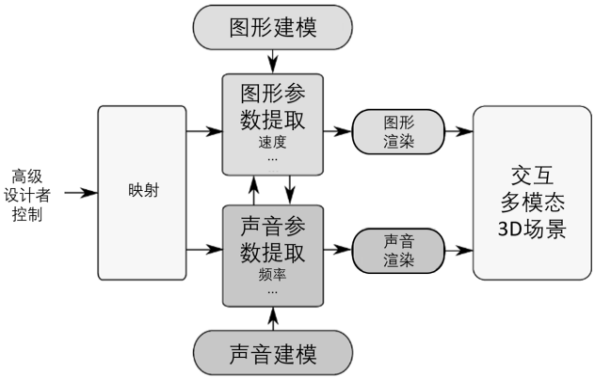


图 8 经典的基于信号模型的声源仿真流程[74]

为了提高仿真声音的真实度，后续的研究方法在构造经验公式的过程中，大多采用物理启发式策略，即从物理发声原理或物理现象出发，通过构造不同的模态谐振器和激励信号，从而实现可以描绘对应物理发声原理或物理现象的经验公式。这种基于物理启发的经验公式，成功仿真了固体滚动声音[78]、纸张撕裂声音[79]以及脚步声[31]，并且由于这类物理启发式经验公式的模型复杂度较低，因此可以实现高效求解计算，满足虚拟现实交互式应用的实时性要求。然而，这类方法中信号样本的构造往往基于大量的实验经验和参数调整，并且部分参数需要用户手动预设，因此限制了其在虚拟场景中的应用范围。

总的来说，目前基于经验公式的声源仿真方法主要存在三方面局限性：（1）简化的声音模型导致了声音细节信息缺失；（2）经验公式的有效性过度

依赖参数的选择与调整；（3）经验公式构造难度大，往往基于研究人员大量的实验观察和经验总结。因此，尽管基于经验公式的声源仿真存在明显的参数化优势，然而由于上述局限性，目前基于经验公式的声源仿真研究较少，而更多的研究倾向于基于更易于获得的录音样本进行声源仿真。

3.2.2 基于录音样本的声源仿真

录音样本作为最常见的音频信号，被广泛地应用于动画配音、游戏制作等领域。与经验公式相比，录音样本具有真实度高、容易获取等优势。然而，由于录音样本相当于封装好的信号组合，因此如何建立声音与视觉场景的映射关系并实现细粒度的试听同步成为这类声源仿真方法的主要挑战。此外，基于录音样本的声音仿真技术，其结果的多样性往往依赖于录音样本库的大小。而在游戏、虚拟现实这类交互式应用中，分配给音频的存储空间往往十分有限，例如在游戏 Crackdown II 中全部音频预算仅为 25MB[28]。因此如何基于有限的录音样本合成多种多样的声源信号，是这类方法面临的另一挑战。

在视听同步方面，早期的一种解决方案通过计算视觉画面的相关性来重组声音信号[80]。具体来说，这类方法基于传统的音频颗粒合成方法[81]，将原始音频样本拆分成一组声音颗粒，即持续时间很短的声音片段，然后将声音颗粒重新组合并得到新的、与视觉画面同步的声音。为了构造重组规则，用户需要提供有声视频样本作为源视频画面和源声音，通过计算源视频画面与待配音画面的相似度，从而得到每个源视频画面对应声音颗粒的概率分布。选择概率最高的声音颗粒进行重排列，最终生成与源声音具有音色相似性的新声音并且实现与待配音画面的同步。这类方法可以实现音色差异性较大的多种场景声源仿真，具有很强的通用性和可扩展性。但是，这种首先计算视觉相关性，进而再构造视听映射的声源仿真方法，不仅需要用户提供额外的源视频画面，并且其声音结果仿真的准确度往往依赖于视觉画面相关性计算的准确度。因此后续的方法大多直接基于视觉画面与录音样本的相关性构造视听映射关系。

对录音样本进行分类并手动添加音频类别标签是目前最常采用的一种视听相关性构造策略。例如 Picard 等人[82]根据接触事件的类型，将录音样本分为脉冲事件的声音和连续事件的声音，实现了固体运动声音仿真。Schreck 等人[83]将纸张产生的声音分为摩擦声和弯曲声，实现了纸张运动声音的仿真。显然，对于这类基于音频类别标签的声源仿真方法，其声源同步准确度与音频标签的细化程度密不可分。针对软体这类形变更复杂，声音变化更加丰富的声源，Su 和 Joslin[29,84]为不同的软体对象建立了类别标签更加细化的声音数据库，例如分别录制了布料摩擦声音、布料褶皱声音、绳索摆动声音以及绳索拉伸声音，并根据布料的运动速度、接触面积对声音数据库进一步细分，从而实现从声音数据库中提取到更加精确的录音样本。然而，这种对录音样本的逐层细分往往需要大量的前期准备工作。不同声源的录音样本划分、音频标签的细化程度往往不尽相同，因此这类方法的通用性较差，很难基于同一仿真框架合成不同种类声源。此外，在录音样本划分的过程中，人为干预过多导致了音频类别往往存在一定的主观性。

另一类构造录音样本与视觉画面相关性的策略则降低了人为干预，通过从录音样本中提取声音特征来建立录音样本与视觉画面的关联性。以鼓掌声为例，Peltola 等人[85]通过实验观察，建立了鼓掌时手掌的空腔构型和鼓掌声音频谱之间的映射关系，合成了多种类型的鼓掌声音。而对于更加复杂的流体声源，Wang 和 Liu[36]提出了一种实时的基于样本声音响度特征的海浪声音仿真方法，通过将海浪划分为不同的海浪块，构造了样本声音的响度与海浪块速度的映射函数，将声音片段串联得到最终的海浪声音。然而，仅仅通过单一声音特征声音构造视觉画面与声源的映射关系往往很难实现高精度的同步效果。频谱质心、频谱能量等声音特征与视频画面的关系值得进一步探索。

现有基于录音样本的声源仿真方法在视听同步方面进行了一系列探索。然而，对于如何基于有限的录音样本合成多种多样的声源信号这一问题，大多方法则仍采用在时域空间中对录音样本进行简单的振幅调整，因此仿真结果的多样性仍然依赖于录音样本库的大小。在最新的一项针对固液耦合场景的声源仿真技术中，Cheng 和 Liu[30]首次设计了一种基于频域空间信号重构的声源仿真方法，该方法在提取了合适的录音样本后，将录音样本转换到频域空间中，并通过在频域上重建相位实现了录音样本的多样性调整（如图 9 所示）。与时域空间的信号调整相比，这种在频域空间的信号重合成策略提供更多的声音调整选项，为平衡声音多样性与音

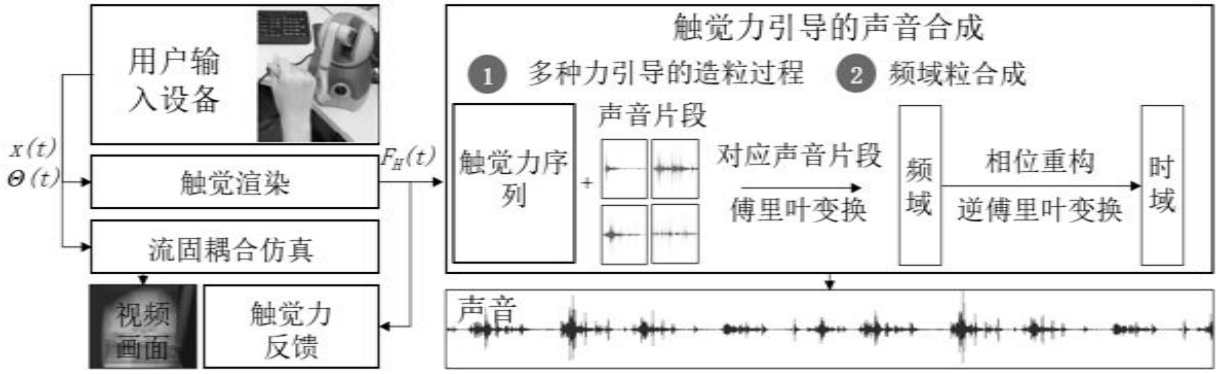


图 9 典型的在频域空间重构声源信号流程图[30]

频内存预算提供了可能。

表 4 列出了采用上述两种获取组成声源的信号方式的代表性方法，以及其视听映射策略、适用场景和算法效率。总的来说，基于录音样本的声源仿真方法通常具有较高的算法效率，因此十分适用于交互式应用系统中，在虚拟现实、游戏等应用中具有广阔的前景。但同时，这类方法所面临的同步挑战和内存挑战仍然需要进一步探索其解决方案。现有方法中所采用的同步方案还比较单一，将多种声音特征组合使用，或将类别标签与声音特征进行组合是否会提升同步准确度值得研究人员进一步探索。而对于内存挑战，频域声源重构提供了一个可能的解决方案，然而这方面的工作刚刚起步，仍有大量的开放问题需要进一步研究。

3.3 基于混合模型的声源仿真技术

随着基于物理模型和基于信号模型的声源仿真技术不断发展，这两类方法各自的优势逐渐凸显，基于物理模型的声源仿真方法具有明显的同步性优势，而基于信号处理的声源仿真方法则往往算法效率较高，合成声音包含更加丰富的声音细节。因此，将这两类模型结合的混合模型逐渐成为近几年的研究趋势。现有方法主要通过参数估计和细节增强这两类策略实现混合模型的构造，本节将按照上述两类模型构造策略分别展开讨论。

3.3.1 参数估计策略

参数估计策略是一种早期的混合模型构造策略，其核心思想是通过对录音样本进行信号分析，从真实录音中提取构造物理声学模型的关键参数，从而避免了复杂的物理建模与参数解算。然而，由于录音信号相当于封装好的声音模型，因此从录音信号中逆向推导得到物理模型参数往往十分困难。

早期的声源仿真对象大多是一些只涉及简单的物理声学模型的固体运动场景。以刚体声源仿真为例，Doel 等人[86]设计了一种通过提取录音中声音的频率分布参数来驱动刚体撞击声、摩擦声和滚动声重合成的实时声源仿真方法。该方法不再通过求

(a)真实场景(b)虚拟场景

图 10 声源参数估计场景示意图[86]

解固体振动的动力学方程获得振动方程 $x(t)=ae^{-\beta t}\cos(2\pi ft)$ 所需的振幅 a、阻尼 β 和频率 f 等参数，而是首先对真实场景（如图 10(a)所示）中的固体运动进行录音，通过对真实录音进行频谱分析，提取振动方程所需的参数（振幅 a、阻尼 β 和频率 f），最终合成虚拟场景（如图 10(b)所示）中固体任意运动状态所对应的声音。尽管这类方法可以实现声源的实时仿真，然而仿真结果的真实度极大地依赖于样本声音录制的准确度。此外，固体的复杂运动往往不再遵循这种简单的振动模型。因此，对于复杂的虚拟场景，例如滑动表面的微碰撞，则很难通过这类合成理想的声音效果。

为了提升模型参数估计准确度，后续研究将心理声学模型与物理声学模型结合。以影响声源材质属性的阻尼参数为例，Ren 等人[43]设计了一种基于声音材质感知相似性的阻尼参数估计方法。该方法通过最小化材质感知相似度来确定声源物理模型的最优阻尼参数。其中，材质感知相似度是基于心理声学实验得到的材质感知不变量[87]。虽然该方法基于估算出的阻尼参数可以合成材质相同但形状各异的物体碰撞声音，然而采用估算的阻尼参数合成的固体撞击声与真实录音样本仍然有一定的差异性，这是由于影响材质声音的因素并不唯一，材质的密度分布等特征同样会对材质声音产生影响。

表 5 基于细节增强的环境声音合成方法对比

文献	声源类型	同步方案	算法效率
An 等人[88]	粗糙布料声音	人为手动匹配	非实时
	高真实度布料声音		
Chadwick 和 James[34]	低频火焰燃烧声音	频率带宽扩展	非实时
	高频火焰燃烧声音		
Yin 和 Liu[89]	燃烧噪声	声音能量匹配	非实时
	旋涡声音		
	爆破声		
Cheng 和 Liu[90]	气泡声音	基于法向力和网格面积的关键帧融合	实时
	撞击声音		
Liu 等人[26]	基本雨声	声源响应算法	非实时
	材质声		

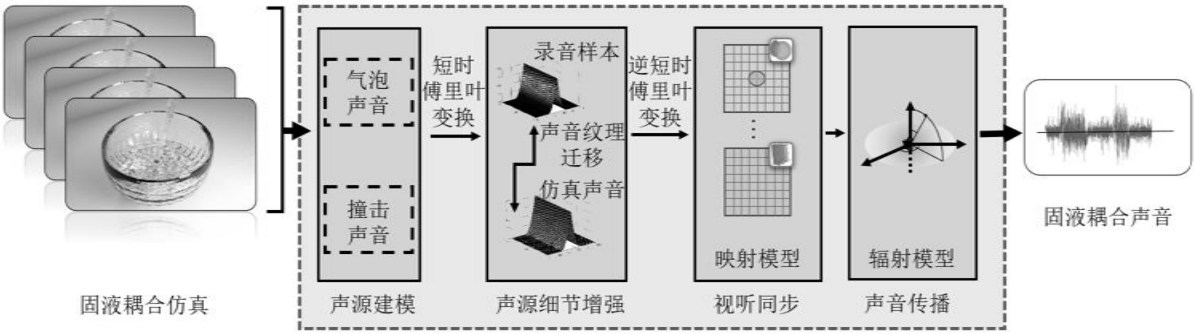


图 11 基于声音纹理迁移的声源仿真流程[90]

总的来说，参数估计策略可以在一定程度降低物理模型构造求解的复杂度，从而缩短声源仿真的时间开销。然而，目前基于参数估计的混合模型适用范围有限，仅适用于简单运动场景中的刚体声源仿真。这是由于刚体复杂运动、软体、液体和气体的声学模型往往是非线性模型，因此很难基于录音样本逆向还原其声学模型参数。此外，基于录音样本估算得到的模型参数，虽然在一定程度上简化了物理模型的构造过程，但是估算得到的参数往往受限于录音样本的选择，采用不同的录音样本估算得到的参数往往会有偏差。总之，基于参数估计的混合模型构造难度较大，尽管心理声学模型的引入提供了一种解决思路，但是仿真结果仍然不是很理想，需要进一步探索潜在的解决方案。

3.3.2 细节增强策略

基于细节增强策略构造的混合模型大多用于复杂虚拟场景的声源仿真。这主要由于复杂虚拟场景声源具有两个特点：（1）在很多复杂的虚拟场景中，人们感知到的声音并不是来自单一声源，而这类多声源场景很难通过构造准确的声学模型来仿真声音；（2）对于软体、气体和液体这些非线性变化的物体，其运动产生的声音中往往包含大量的非线性声音细节，仅基于物理模型得到的仿真结果往往音色不够丰富，而仅基于信号模型又很难实现细粒度视听同步。因此，针对这些复杂虚拟场景的声源仿真，研究人员采取的常见流程为：首先基于物理模型对声源的主体进行仿真，然后基于信号模型来丰富声源主体的音色细节，最后通过设计不同的同步、耦合策略，将物理模型与信号模型相结合，从而仿真出富含声音细节的声源信号。

在混合模型构造过程中，有效的耦合同步策略成为影响仿真结果好坏的决定性因素之一。一种最直接的解决思路是用户手动选择来获得最匹配的声音片段[88]。首先基于物理模型合成低精度同步声音，然后通过用户手动选择，从预先录制的声音数据库中提取与低精度声音变化相似的录音样本。将录音样本拼接得到与低精度声音最匹配的高真实度声音。显然，这种基于用户手动选择的方法，更适用于动画音效制作，而很难推广应用于交互式虚拟系统中。因此，在后续的工作中，研究人员更关注于完全自动化的耦合同步策略。

提取声音特征作为不同声源仿真模型的关联参数是另一种常见的耦合同步策略。以火焰燃烧场景声源仿真为例，作为一类典型的多声源场景，火

焰燃烧声音不仅包含燃烧导致的气体旋涡声音，同时包含复杂的化学反应产生的燃烧噪声以及固体可燃物的爆破声。然而，不论是热声学效应、还是固体可燃物爆破，都很难通过直接构造物理声学模型合成对应的声音。针对这一多声源仿真场景，Chadwick 和 James[34]基于物理模型仿真得到低频火焰燃烧声音，然后基于信号模型合成具有燃烧声音细节的高频火焰燃烧声音。为了实现这两部分的耦合同步，该方法基于声音的频率特征，设计了两种频率带宽扩展方案来插入高频火焰燃烧声音，从而增强低频火焰燃烧声音细节。在近期的研究中，Yin 和 Liu[89]则基于声音能量特征，通过计算基于不同模型仿真得到的火焰燃烧声音的声音能量值，叠加声音能量相似度最高的声音片段实现了可以区分固体可燃物的燃烧声音仿真。然而，在这类细节增强策略中，关联参数的选择大多基于实验观察，不同场景的声源仿真需要构造特定的关联参数，因此关联参数的选择往往会影响最终仿真结果真实性。

在最新的工作中，Cheng 和 Liu[90]首次针对固液耦合场景，提出一种基于“声音纹理迁移”的细节增强策略，这一策略源于图像处理领域的图像纹理迁移技术，其核心思想是不再构造基于不同模型仿真得到的声音之间的关联参数，而是将具有丰富音色细节的录音样本作为纹理，迁移到同步度高但缺失细节的仿真声源中（如图 11 所示）。而在下雨场景这类典型的室外大型固液交互场景中，研究人员则进一步优化了“声音纹理迁移”策略，设计了一种基于声源着色器的方案，实现了雨滴落在不同表面的声源仿真[26]。这类方法为声源细节增强提供了一种全新的思路，将传统的声音同步问题转化为特征迁移问题。然而，由于声音纹理特征与图像纹理特征相比，其特征分布往往更具有随机性，特征也并不明显，因此如何有效识别、提取声音纹理特征仍需进一步探索。

总的来说，与基于参数估计的环境声源仿真方法相比，基于细节增强的环境声源仿真方法具有更大的适用范围，可以为软体、气体以及液体等多种复杂的虚拟场景合成音色更加丰富的声音。然而，由于这类方法不仅需要构造并解算声学模型，而且需要建立样本声音库来丰富声音细节，因此导致声音的合成过程不仅很难达到实时，同时还需要额外的音频存储空间。表 5 从声源类型、同步方案、算法效率等方面详细比较了基于细节增强策略的环境声源仿真方法。

总之，不论是基于参数估计策略还是基于细节增强策略的混合声源仿真模型，都极大地丰富了合成声音的音色细节，并在一定程度上提升了环境声音合成方法的计算效率。然而，目前对于复杂声学场景的声源仿真仍处于探索阶段，对于多声源的虚拟场景，其场景中的声源划分大多依赖于人为主观判断，因此合成的环境声音与真实场景声音仍然具有一定的差距。

3.4 基于深度学习模型的声源仿真技术

随着深度学习技术的快速发展，基于深度学习模型的声源仿真技术在近几年逐渐兴起。这类方法通过构造视觉和声音的映射关系来仿真声源，本质上是一种视觉-听觉跨模态生成技术。尽管已经有很多工作在视听跨模态感知领域展开研究[91-94]，然而这些方法只需要对音频信号进行检索与分类，并不需要对声音信号的细节进行建模。

基于深度学习模型的声源仿真技术则需要通过对不同模态数据的学习得到复杂的生成函数，并且能够有效地编码和解码包含在不同模态数据中的信息，因此这类跨模态生成问题往往十分具有挑战性。在跨模态生成领域中，文本到语音(TTS)的跨模态生成受到研究人员的广泛关注[95-98]。工业界中，科大讯飞[99]、百度[100]等机构已经取得了一系列的应用成果。然而，与 TTS 任务中文本和语音具有显著的信息对应关系相比，在环境声源的仿真任务中，视觉和声音信息则没有严格的对应关系，因此生成模型的构造往往更加困难。现有方法通常将这一视觉-听觉跨模态生成问题描述为一个回归问题，其目标是构造视频帧序列和音频特征序列之间的映射关系。本节将从网络模型框架设计以及音频数据集构造两方面分别展开讨论。

3.4.1 网络模型框架设计

跨模态生成成功的一个关键方面是能够有效地编码和解码包含在不同模式中的信息。其中卷积神经网络(ConvolutionalNeuralNetwork, CNN)已经在面向图像的多种任务中表现出良好的信息编码性能。而循环神经网络(Recurrent Neural Network, RNN)则在对序列的非线性特征进行学习时具有一定优势。因此 Owens 等人[101]设计了一种由 CNN 与长短期记忆网络（Long Short-Term Memory, LSTM）组成的网络结构，训练得到了视频序列和音频特征的映射关系（如图 12 所示）。为了完成训练，该方法收集录制了 Greatest Hits 数据库，并通过训练好的模型合成了木棍敲打在不同材质物体上发出的

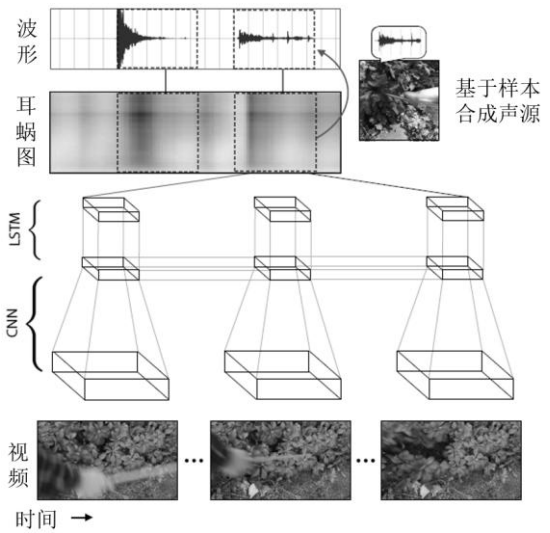


图 12 典型的基于深度学习模型的声源仿真流程[101]

声音。然而这一方法并没有直接构造视频画面与音频信号的映射关系，而是构造了视频帧与耳蜗图（cochleagrams）的映射模型，因此需要后声源匹配算法来生成最终的音频信号。

为了解决这种需要二次转换的局限性，Zhou 等人[102]将 CNN 与层次化循环神经网络（Sample Recurrent Neural Network, SampleRNN）相结合，首次实现了为无声视频合成原始声音信号（raw audiosignals）。具体来说，Zhou 等人[102]为视频编码器设计了三类不同的视觉编码方法：帧到帧的编码，序列到序列的编码以及基于光流图的编码。其中基于光流图的编码需要将原视频帧和视频帧的光流图同时作为视频编码器的输入。光流图的引入使得视频编码器可以显式地捕获运动信号，从而提高其视听同步的准确率。然而上述方法存在两方面明显的局限性：（1）需要针对不同类别的声源，分别训练得到对应的视听映射模型；（2）同步效果仍然不理想，这主要是由于 RNN 对视觉内容的学习存在一定的局限性[101,103]。针对第一点局限性，在近期的工作中，Chen 等人[104]在原有网络模型中加入声音分类网络，并通过一个预先训练好的 SoundNet[105]来计算感知损失，从而提升合成声音的准确度。Ghose 等人[106]则通过将一个鲁棒的多尺度 RNN（FSLSTM）与 CNN 相结合，从而更好地理解复杂的视频-音频随时间变化的关联性。

在优化仿真声音视听同步性方面，生成对抗网络（GenerativeAdversarial Networks, GANs）以其较强的学习能力，吸引了越来越多研究人员的关注。Cheng 等人[107]将条件生成对抗网络（Conditional Generative Adversarial Network, CGAN）和 SampleRNN 相结合来提升声音同步的准确度。具体来说，该方法通过基于 CGAN 的时序同步网络模型来实现声音与视频画面的同步，并通过基于 Sample RNN 的音色增强网络模型来进一步提升声音的真实感。随后，Chen 等人[108]将视听对齐问题进一步分为时间对齐和内容对齐两方面，并提出了 REGNET 框架。具体来说，该框架包含了一个基于 GAN 的音频转换正则化器，从而防止模型学习视频帧与画面外物体发出的声音之间的不正确映射，有效实现了高精度的视频-音频同步效果。

3.4.2 音频数据集构造

基于深度学习模型的声源仿真作为一类数据驱动的方法，训练数据的好坏往往决定着训练结果的质量。而现有的音频数据集由于音视频匹配关系不明显，音频清晰度不高、音频标签过于简化等原因，大多无法直接用于基于深度学习模型的声源仿真，例如用于音频事件识别的大规模数据集 AudioSet[109]，视频中的声音通常是多个声源和背景噪声的混合，并且只为音频事件提供类别标签，而没有提供标签关于声源属性（材质、形状等）的标签。因此，现有基于深度学习模型的声源仿真方法，往往需要为模型训练构造高质量的视听同步数据集。表 6 汇总了现有基于深度学习模型的声源仿真方法中所采用的训练数据集，其中最具代表性的是 Zhang 等人[110]构造的一个可以区分声源形状、材质的训练数据库：Sound-20K。该数据库不再通过录音和人为添加标签来标注声音，而是通过基于物理模型的声源仿真技术来直接合成声源信号。其构造流程如图 13 所示，对于给定物体形状属性、材质属性和场景参数，首先通过物理引擎模拟物体碰撞和运动。然后基于物理引擎提供的碰撞数据和预先计算的物体声音辐射场，通过音频引擎来仿真物体声音。最后通过图形引擎渲染得到物体运动呈现的视频画面。与搜集有声视频片段组成的数据库相比，这种合成的数据库具有更强的可扩展性，而且声音质量更高，不存在背景噪声等干扰声音，此外，合成的声音具有充分的注释，可以提供声源信号详细的物理参数属性。然而，正如在 3.2 节所讨论的，目前基于物理模型的环境声源仿真仍存在算法时间复杂度较高、音色细节缺失等局限性，这些都限制了合成数据库的进一步发展。

表 6 基于深度学习模型的声源仿真数据库

数据库	视频数目	视频类别数目	视听映射标签	声源场景
Greatest Hits[101]	977	按照动作类别划分：2 类 按照材质类别划分：18 类	动作标签 材质标签	木棍敲击、刮擦声源
VEGAS[102]	28109	10 类不同场景	场景类别标签	人类/动物声源以及环境声源
VIG[104]	16,024	15 类不同场景	场景类别标签	声音与画面高度同步的声源
AFD[106]	1000	12 类不同场景	场景类别标签	常见的拟音声源
Sub-VEGAS + Sub-AudioSet[108]	13008	8 类不同场景	场景类别标签	声音与画面高度同步的声源
Sound-20K[110]	20378	按照场景类别划分：22 类 按照形状类别划分：39 类 按照材质类别划分：7 类	物体的形状、材质属性标签	具有明显材质、形状属性的室内固体声源

图 13 合成数据库的构造流程[110]

总的来说，目前基于深度学习的声源仿真技术为虚拟场景中的环境声音仿真提供了新的思路。然而，这类方法尚处于初期探索阶段，目前仅针对视频与音频映射关系进行了研究。与视频画面不同的是，动画、游戏等虚拟场景中的物体具有空间信息，相比于二维的图像内容理解，三维图形数据具有更高的复杂性与多样性，因此学习三维虚拟场景的视觉内容与声音内容之间的映射关系往往具有更大的挑战。尽管如此，跨模态的深度学习方法不再对每一类物体的发声机制与映射关系分别建模，为通用的环境声音合成提供了可能性。

4 存在问题与研究展望

由于虚拟场景种类繁多，不同场景之间差异较大，因此，与之相对应的环境声源仿真方法也十分的多样化。基于不同模型的声源仿真方法的特点与局限性已经在第三节中分别讨论过，接下来，本节将从整个虚拟场景中环境声源仿真领域的角度概括分析目前亟待解决的问题，并对接下来的研究趋势进行展望。

4.1 存在的问题与挑战

4.1.1 高真实感交互式仿真仍然困难

虽然现有环境声源仿真方法众多，然而每一类方法都存在一定的局限性，为虚拟场景实现高真实感的交互式环境声源仿真仍然十分困难。表 7 列出了目前基于不同模型的环境声源仿真方法在算法输入、算法效率以及仿真结果质量等不同角度的对比结果。具体来说，在算法所需输入参数方面，由于基于物理模型的声源仿真方法通过构造声源的物理声学公式来仿真，因此只需要输入简单的物理参数即可实现视听同步的声源仿真，对于输入参数要求最低。而基于信号模型的声源仿真方法则需要额外构造视听映射关系，因此往往需要根据仿真场

表 7 基于不同模型的声源仿真方法对比

模型类别	算法输入	算法效率	仿真结果质量
物理模型	物理参数	(1)一般很难实时仿真 (2)部分简单场景可以达到实时仿真	(1)音色真实度较低 (2)视听同步度很高
信号模型	(1)图形仿真参数 (2)部分方法需要录音样本声音库	算法效率很高，大部分方法都可以达到实时仿真	(1)音色真实度很高 (2)视听同步度较低
混合模型	(1)物理参数 (2)部分方法需要录音样本声音库	(1)只有少量算法实现实时仿真 (2)大部分仍无法达到实时	(1)音色真实度较高 (2)视听同步度较高
深度学习模型	大量画面与声音同步的视频数据库	需要大量的模型训练时间	(1)音色真实度较低 (2)视听同步度较低

景提取不同的图形仿真参数，对于输入参数要求较高。混合模型则结合了上述两类模型的优势，因此往往只需要物理参数即可完成同步。但是基于混合模型和基于信号模型的声源仿真方法往往需要基于录音样本来丰富声音细节，因此需要构造额外的录音样本声音库作为输入。基于深度学习模型的声源仿真方法则不仅需要音频数据库，同时需要输入与音频同步的视觉画面，因此这类方法对于输入数据的要求最高，其数据库的构造也最困难。

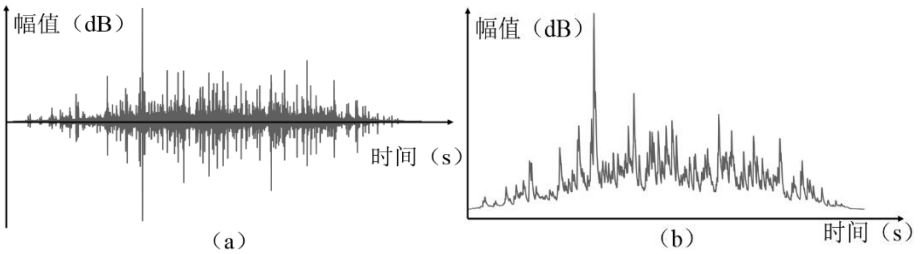
然而在算法效率方面，基于物理模型的环境声源仿真求解过程往往十分复杂，因此这类声源仿真方法大多无法达到实时仿真。而信号模型则由于往往只追求感知上的合理性，而不严格遵循物理原理，因此这类声源模型的求解过程十分高效，大多可以实现实时仿真。基于混合模型的环境声源仿真方法目前大多用于实现复杂虚拟场景中的声源仿真，虽然可以实现仿真难度更高的非线性声源变换，也因此往往受限于物理模型的复杂求解过程，从而导致算法效率无法达到实时。而基于深度学习模型的声源仿真方法分为两个阶段，其训练过程往往需要耗费大量的时间，训练得到模型之后则可以实时生成与输入无声视频同步的音频信号。

在仿真质量方面，声源仿真结果的真实感取决于两方面：音色真实度和同步准确度，音色真实度指的是仿真声音与真实录音音色的相似度，而同步准确度则表示仿真声音与虚拟场景变化的一致度。由于基于信号模型以及基于混合模型的环境声源仿真方法往往需要额外输入真实的录音样本声音，因此其仿真结果的音色与真实录音更加接近，音色真实度往往较高。而基于物理模型的环境声源仿真方法则很难构造真实录音中的音色细微变化与音色细节特征，因此，其仿真结果的音色往往更加单一，与真实录音中丰富的音色存在一定的差异。

在同步准确度方面，由于基于物理模型的声源方法是完全基于虚拟场景中物体运动变化来构造与之对应的声音模型，因此仿真声音与物体运动高度同步。而在基于信号模型的声源仿真方法中，无论所采用的是经验公式还是录音样本，均不包含与物体运动直接对应的声音参数，因此需要依赖于人为经验构造视听同步映射函数，从而很难实现高精度的视听同步。基于深度学习模型的声源仿真方法的结果质量则往往依赖于训练样本的质量，而现有用于环境声源仿真训练的数据库，其声音与视觉画面的匹配度仍然是一种粗精度的匹配。因此，很难基于现有的训练数据库实现可以区分不同形状、材质、外力作用位置以及外力大小的声源仿真。因此无论在音色真实度还是同步准确度方面，都存在较大的提升空间。总的来说，目前环境声源仿真仍然缺乏有效的平衡声音质量和算法效率的方案，从而导致现有方法很难为交互式虚拟场景生成高真实感的声源信号。

4.1.2 缺乏定量的评价指标

现有环境声源仿真方法对于合成声音质量评价主要通过两种方案，基于声音图像表示的客观评价以及基于用户调查的主观评价。常见的声音图像表示包括合成声音的波形图（图 14（a））、包络图（图 14（b））、频谱图（图 14（c））和声谱图（图 14（d））。其中波形图和包络图是声音在时域上的图像表示，声音的波形图反映了声音响度随时间的变化。而声音波形的轮廓线则为声音的包络，因此声音的包络图可以更加直观地反映出声音响度随时间的变化。声音的频谱图则是通过对声波进行傅里叶变换（Fourier transform）得到的，是声音在频域上的图像表示，反映了声音的频率分布与频率大小。而声谱图（又称为语谱图）则同时包含了声音的时域和频域信息，通过对声波进行短时傅里叶变



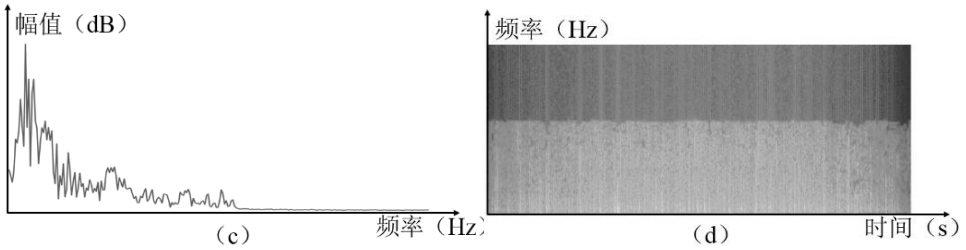


图 14 常见声音的图像表示，图（a）-（d）分别为同一段雨声的不同图像表示

表 8 不同声音的图像表示对比

分析角度	名称	横坐标	纵坐标	适用场景
时域	波形（waveform）图	时间	幅值	分析声音幅度随时间变化的规律
	包络（envelope）图	时间	幅值	分析声音幅度、峰值随时间变化的规律
频域	频谱图（spectrum）	频率	幅值	分析声音中的频率内容及分布规律
时域-频域	声谱图（spectrogram）	时间	频率	分析声音中频率随时间的变化规律

换（short-time Fourier transform）得到声音随时间变化的频率分布，声谱图中颜色的深浅则表示频率的强弱，越深的颜色对应的频率强度越弱。表 8 汇总了不同声音的图像表示方法。

尽管将声音信号转化为图像可以更加直观地观察到声音的变化与特征分布，然而这种基于图像的定性评价存在很大的局限性。对于两段音色听起来十分相似的环境声音，其图像表示往往仍然会存在较大的差异。这是由于环境声音的频率分布与语音、音乐相比更不具有规则性，所以仅通过比较声谱图、频谱图很难进行准确的评价。此外，仿真声音与虚拟场景变化的同步准确度同样是评价仿真结果质量的重要内容，而通过声音的图像表示往往无法比较声音与虚拟场景的同步准确度。因此，现有环境声源仿真方法，仍然以用户调查这类主观评价方法为主要评价手段。然而，由于用户对于声音质量的评价往往存在个体差异，因此主观评价通常采取单因素方差分析等方法来尽量减小个体差异导致的结果误差。总之，如何设计准确度更高的仿真声音定量评价指标，是环境声源仿真领域亟待解决的问题。

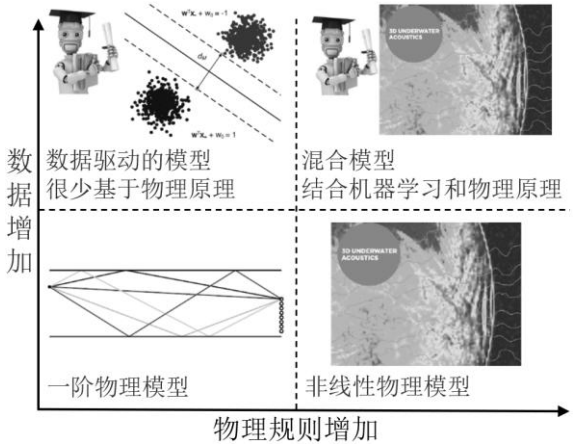


图 15 物理规则与数据共同驱动的声学发展示意图[111]

4.2 未来发展方向

4.2.1 实时声源仿真

随着虚拟现实技术和互联网技术的发展，游戏和虚拟现实场景呈现出电影化趋势[3]，与视觉场景同步的动态声场势必成为提高用户沉浸感的关键要素之一，而实时反馈的动态音效对于声源仿真的时间复杂度提出了更高的要求。尽管现有方法通过预计算、对声源模型进行分解、构造等效替代声源模型，结合录音样本等方案，对声源高效求解进行了一系列探索，然而其在仿真声源类别、声源复杂

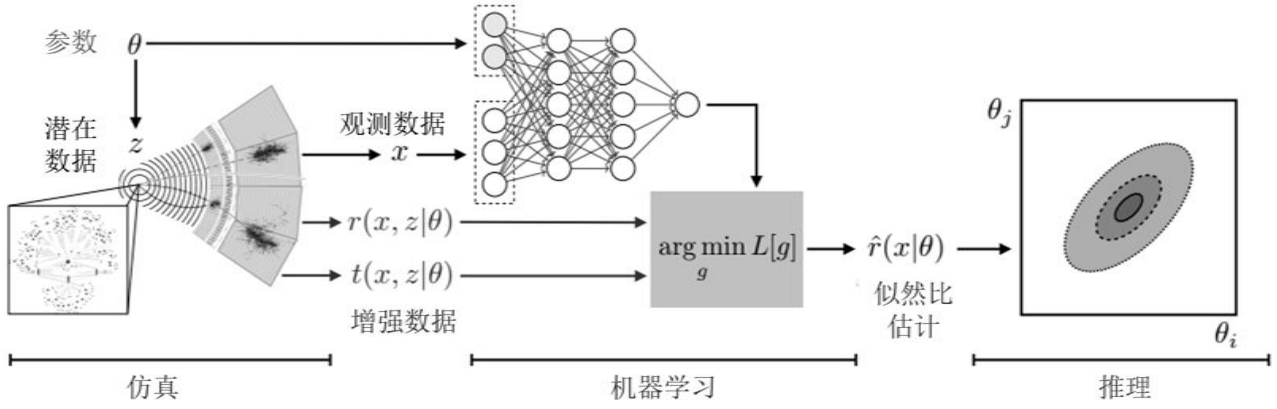


图 16 结合机器学习的似然比估计流程[112]

度方面还远远不能满足虚拟现实系统的要求。尤其是对于涉及多种形态声源耦合发声（液固耦合、气固耦合等）的研究尚处于初期探索阶段，由于其声源模型具有高度非线性，声源仿真复杂度较高，因此目前面向这类声源的仿真方法很难实现实时的声音反馈。基于当前研究背景，实时声源仿真的未来发展趋势可以分为如下几个方面：（1）寻求物理模型的高效求解方案，通过并行计算、GPU 加速等方案提高对物理模型的解算速度；（2）提升信号模型的可解释性，从而实现耦合度、同步度更高的视听映射关系。（3）寻求基于不同原理的混合模型构造，通过样本数据来简化物理模型中复杂的模型求解。总的来说，与虚拟场景高度同步的实时动态声源仿真具有广阔的发展前景，是打造强互动性、高沉浸感的虚拟现实系统的突破点之一。

4.2.2 大数据背景下的模型构造

近年来，随着大数据时代的到来以及深度学习技术的发展，基于深度学习的声源仿真作为跨模态问题，成为目前深度学习领域的热点问题之一。传统的声学仿真模型聚焦于构造高阶、非线性的复杂物理模型来实现声源的高精度还原(如图 15 所示)。而随着数据的增多，基于数据驱动的深度学习方法在一些简单声源仿真方面取得了巨大的成功。因此，在大数据背景下的混合模型构造，势必成为未来声源仿真的重要发展方向，具体可概括为：（1）构造高精度、多标签的视听训练数据库，训练得到可以区分声源形状、材质、外力作用位置以及外力大小等信息的深度学习模型；（2）将物理模型与深度学习模型相结合，物理模型可以提升深度学习模型的可解释性，而深度学习模型通过从数据中归纳现象而降低了物理模型描述现象的复杂度。在最新的物理学研究中[112]，已经将物理求解过程中昂贵的数值积分替换为机器学习进行似然比估计（如图 16 所示）。总之，基于混合模型的环境声源仿真方法已成为未来的发展趋势，具有广阔的发展前景。

5 结论

声音作为增强虚拟场景真实感、沉浸感的重要组成部分，为虚拟场景自动地仿真真实、同步的声源信号具有十分重要的意义。本文综述了环境声源仿真方法的研究背景、应用领域、基本原理、研究现状、存在的问题以及未来发展趋势。首先，从物理声学和心理声学两方面，介绍了环境声源仿真的基本原理。然后依据声源仿真所基于的模型不同，从基于物理模型、基于信号模型、基于混合模型以及基于深度学习模型四大类，分别针对每一类环境声源仿真方法详细介绍并分析了各自的方法设计策略、创新性与局限性。随后，讨论了现有环境声源仿真方法在交互式虚拟场景应用、定量评价指标等方面存在的问题，并基于现有环境声源仿真研究中的局限性对未来研究趋势进行展望。

致 谢 感谢编辑部老师和各位审稿人对本文的宝贵建议！

参 考 文 献

[1] Gygi B, Kidd G R, Watson C S. Similarity and categorization of environmental sounds. Perception and Psychophysics, 2007, 69 (6): 839–855

[2] Gaver W W. What in the world do we hear? An ecological approach to auditory event perception. Ecological Psychology, 1993, 5(1):1-29

[3] Kenwright B. There's more to sound than meets the ear: sound in interactive environments. IEEE Computer

Graphics and Applications, 2020, 40(4): 62-70

[4] Raghuvanshi N, Snyder J M. Parametric directional coding for precomputed sound propagation. ACM Transactions on Graphics, 2018, 37(4): 108:1-108:14

[5] Zhang Z, Raghuvanshi N, Snyder J M, *et al*. Ambient sound propagation. ACM Transactions on Graphics, 2018, 37(6): 184:1-184:10

[6] Zhang Z, Raghuvanshi N, Snyder J M, *et al*. Acoustic texture rendering for extended sources in complex scenes. ACM Transactions on Graphics, 2019, 38(6): 222:1-222:9

[7] Tang Z, Manocha D. Scene-aware sound rendering in virtual and real worlds. //Proceedings of the IEEE VR Workshops, GA, USA, 2020: 535-536

[8] Cao C, Ren Z, Schissler C, *et al*. Interactive sound propagation with bidirectional path tracing. ACM Transactions on Graphics, 2016, 35(6): 180:1-180:11

[9] Chaitanya C R A, Raghuvanshi N, Godin K W, *et al*. Directional sources and listeners in interactive sound propagation using reciprocal wave field coding. ACM Transactions on Graphics, 2020, 39(4): 44:1-44:14

[10] Hai N D, Chaudhary N K, Peksi S, *et al*. Fast HRFT measurement system with unconstrained head movements for 3D audio in virtual and augmented reality applications. //Proceedings of the ICASSP, LA, USA, 2017: 6576-6577

[11] Singhani A, Morrow A. Real-time spatial 3D audio synthesis on FPGAs for blind sailing. //Proceedings of the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, CA, USA, 2020: 104-110

[12] Song Y, Wang X, Yang C, *et al*. Frame-independent and parallel method for 3D audio real-time rendering on mobile devices. //Proceedings of the MMM. Reykjavik, Iceland, 2017: 221-232

[13] Tashev I J. Capture, representation, and rendering of 3D audio for virtual and augmented reality. International Journal on Information Technologies and Security, 2019, 11(SP2): 49-62

[14] Zhang A, Shi J, Pan Z. Realistic spatial sound rendering in virtual environment. Journal of Software, 1996, 7(Supplement): 120-126 (in Chinese)
(张爱东，石教英，潘志庚. 虚拟环境中真实感空间声音合成. 软件学报, 1996, 7(增刊):120-126)

[15] Gao J, Gong B, Yang C. A fast 3D acoustic rendering method. Journal of Image and Graphic, 2003, 8(A):869-874 (in Chinese)
(高剑，龚斌，杨承磊. 快速三维声音渲染算法. 中国图象图形学报, 2003, 8(A):869-874)

[16] Luo F, Wang X, Peng X. Sound rendering technology and its application in virtual environment. Journal of System Simulation, 1999, 11(5):364-367 (in Chinese)
(罗福元，王行仁. 声音渲染技术及其在虚拟环境中的应用. 系统仿真学报, 1999, 11(5):364-367)

[17] Ding R, Liu J, Liu S. An overview of techniques on sound propagation simulation. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(8): 1267-1277 (in Chinese)
(丁锐，刘锦，刘世光. 声音传播模拟技术综述.计算机辅助设计与图形学学报, 2019, 31(8): 1267-1277)

[18] Zeng X. Forty years of development of room acoustic computer simulation. Audio Engineerin, 2008, S1:12-17+23 (in Chinese)
(曾向阳. 室内声场计算机模拟发展40年(1968～2008). 电声技术, 2008, S1: 12-17+23)

[19] Hu R, Wang X, Zhang M, *et al*. Review on three-dimension audio technology. Journal of data acquisition and processing. 2014, 29(5):661-676 (in Chinese)
(胡瑞敏，王晓晨，张茂胜，等. 三维音频技术综述. 数据采集与处理, 2014, 29(5):661-676)

[20] Yin F, Wan L, Chen Z. Review on 3D audio technology. Journal on Communications, 2011, 32(2):130-138 (in Chinese)
(殷福亮，汪林，陈喆. 三维音频技术综述. 通信学报, 2011, 32(2):130-138)

[21] Zhang Y, Zhao J, Wang J, *et al*. Present situation and development of 3D audio technology in virtual reality. Audio Engineering, 2017, 41(6):56-62 (in Chinese)
(张阳，赵俊哲，王进，等. 虚拟现实中三维音频关键技术现状及发展. 电声技术, 2017, 41(6):56-62)

[22] Beig M, Kapralos B, Collins K, *et al*. An introduction to spatial sound rendering in virtual environments and games. Computer Games Journal, 2019, 8(3-4): 199-214

[23] Langlois T R., James D L. Inverse-foley animation: synchronizing rigid-body motions to sound. ACM Transactions on Graphics, 2014, 33(4): 41:1-41:11

[24] Chandrakala S, Jayalakshmi S L. Generative model

driven representation learning in a hybrid framework for environmental audio scene and sound event recognition. IEEE Transactions on Multimedia, 2020, 22(1): 3-14

[25] Park H, Yoo C D. CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification. IEEE Signal Processing Letters, 2020, 27: 411-415

[26] Liu S, Cheng H, Tong Y. Physically-based statistical simulation of rain sound. ACM Transactions on Graphics, 2019, 38 (4): 123:1-123:14

[27] Zheng C, James D L. Harmonic fluids. ACM Transactions on Graphics, 2009, 28 (3): 37:1-37:12

[28] Lloyed D B, Raghuvanshi N, Govindaraju N K. Sound synthesis for impact sounds in video games. //Proceedings of the Symposium on Interactive 3D Graphics and Games, San Francisco, USA, 2011: 55-62

[29] Su F, Joslin C. Procedural sound generation for soft bodies in video games. //Proceedings of the ACM SIGGRAPH Conference on Motion in Games, Newcastle, UK, 2019: 17:1-17:12

[30] Cheng H, Liu S. Haptic force guided sound synthesis in multisensory virtual reality (VR) simulation for rigid-fluid interaction. //Proceedings of the IEEE Virtual Reality, Osaka, Japan,2019, 111-119

[31] Nordahl R, Serafin S, Turchet L. Sound synthesis and evaluation of interactive footsteps for virtual reality applications. //Proceedings of the IEEE Virtual Reality, Waltham, USA, 2010: 147-153

[32] Zheng C, James D L. Rigid-body fracture sound with precomputed soundbanks. ACM Transactions on Graphics, 2010, 29 (4): 69:1-69:13

[33] Schweickart E, James D L, Marschner S. Animating elastic rods with sound. ACM Transactions on Graphics, 2017, 36 (4): 115:1-115:10

[34] Chadwick J N, James D L. Animating fire with sound. ACM Transactions on Graphics, 2011, 30 (4): 84:1-84:8

[35] Dobashi Y, Yamamoto T, Nishita T. Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. ACM Transactions on Graphics, 2003, 22 (3): 732-740

[36] Wang K, Liu S. Example-based synthesis for sound of ocean waves caused by bubble dynamics. Computer Animation and Virtual Worlds, 2018, 29(3–4):e1835

[37] DuG, Zhu Z, Gong X. Fundamentals of Acoustics. Nanjing, China: Nanjing University Press, 2012(in Chinese)
(杜功焕, 朱哲民, 龚秀芬. 声学基础. 南京,中国: 南京大学出版社, 2012)

[38] O'Brien J F, Cook P R, Essl G. Synthesizing sounds from physically based motion. //Proceedings of the ACM SIGGRAPH, Los Angeles, USA, 2001: 529-536

[39] O'Brien, J F, Shen C, Gatchalian C. M. Synthesizing sounds from rigid-body simulations. //Proceedings of the ACM SIGGRAPH/Eurographics symposium on Computer animation, San Antonio, USA 2002: 175-181

[40] Morse P M, Ingard K U. Theoretical Acoustics. New Jersey, USA: Princeton University Press, 1987

[41] Leighton T G. The acoustic bubble. Salt Lake City, USA: Academic Press, 1994

[42] Howe M S. Theory of vortex sound. Cambridge, UK: Cambridge University Press, 2003

[43] Ren Z, Yeh H, Lin M C. Example-guided physically based modal sound synthesis. ACM Transactions on Graphics, 2013, 32 (1): 1:1-1:16

[44] Doel K V D. Sound synthesis for virtual reality and computer games. The University of British Columbia, Vancouver, Canada, 1998

[45] Zheng C. Physics-based sound rendering for computer animation. Cornell University, State of New York, USA, 2012

[46] Fastl H, Zwicker E. Psychoacoustics: facts and models. Heidelberg, Germany: Springer, 2007

[47] Fletcher H, Munson W A. Loudness, its definition, measurement and calculation. The Bell System Technical Journal, 1933, 12(4): 377-430

[48] Calculation of Loudness Level and Loudness From the Sound Spectrum-Zwicker Method-Amendment 1: Calculation of the Loudness of Time-Variant Sound, German Nat. Standard DIN45631/A1:2010, Mar. 2010

[49] Acoustics-Methods for Calculating Loudness-Part 1: Zwicker Method, ISO Standard 532–1:2017(E), Mar. 2017

[50] Ziemer T, Nuchprayoon N, Schultheis H. Psychoacoustic sonification as user interface for human-machine interaction. 2019, arXiv:1912.08609v1

[51] Schönherr L, Kohls K, Zeiler S, *et al*. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. //Proceedings of the 26th Annual

Network and Distributed System Security Symposium, San Diego, USA,2019: 1-15

[52] Ziemer T. Perceptual sound field synthesis concept for music presentation. //Proceedings of the meetings on acoustics Acoustical Society of America, Boston, USA, 2017: 1-13

[53] Doel K V D, Pai D K. The sounds of physical shapes. Presence, 1998, 7 (4): 382-395

[54] Avanzini F, Serafin S, Rocchesso D. Interactive simulation of rigid body interaction with friction-induced sound generation. IEEE Transactions on Speech and Audio Processing, 2005, 13 (5): 1073-1081

[55] Dupont P, Hayward V, Armstrong B, *et al*. Single state elastoplastic friction models. IEEE Transactions on Automatic Control, 2002, 47 (5): 787-792

[56] Ren Z, Yeh H, Lin M. Synthesizing contact sounds between textured models. //Proceedings of the IEEE Virtual Reality, Waltham, USA,2010: 139-146

[57] Zheng C, James D L. Toward high-quality modal contact sound. ACM Transactions on Graphics, 2011, 30 (4): 38:1-38:11

[58] Chadwick J N, Zheng C, James D L. Precomputed acceleration noise for improved rigid-body sound. ACM Transactions on Graphics, 2012, 31 (4): 103:1-103:9

[59] Chadwick J N, Zheng C, James D L. Faster acceleration noise for multibody animations using precomputed soundbanks. //Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Lausanne, Switzerland,2012: 265-273

[60] Chadwick J N, An S S, James D L. Harmonic shells: a practical nonlinear sound model for near-rigid thin shells . ACM Transactions on Graphics, 2009, 28 (5): 119:1-119:10

[61] Cirio G, Li D, Grinspun E, *et al*. Crumpling sound synthesis. ACM Transactions on Graphics, 2016, 35 (6): 181:1-181:11

[62] Cirio G, Qu A, Drettakis G, *et al*. Multi-scale simulation of nonlinear thin-shell sound with wave turbulence. ACM Transactions on Graphics, 2018, 37 (4): 110:1-110:14

[63] Minnaert M. On musical air-bubbles and the sounds of running water. Philosophical Magazine, 1933, 16 (104): 235-248

[64] Doel K V D. Physically-based models for liquid sounds. ACM Transactions on Applied Perception, 2005, 2 (4): 534-546

[65] Moss W, Yeh H, Hong J M, *et al*. Sounding liquids: Automatic sound synthesis from fluid simulation. ACM Transactions on Graphics, 2010, 29 (3): 21:1-21:13

[66] Langlois T R, Zheng C, James D L. Toward animating water with complex acoustic bubbles. ACM Transactions on Graphics, 2016, 35 (4): 95:1-95:13

[67] Strasberg M. The pulsation frequency of nonspherical gas bubbles in liquids. Journal of the Acoustical Society of America, 1953, 25 (3): 536-537

[68] Zita A. Computational real-time sound synthesis of rain. Linköping University, Linköping,Sweden, 2003

[69] Powell A. Theory of vortex sound. Journal of the Acoustical Society of America, 1964, 36 (1): 177-195

[70] Howe M S. Theory of vortex sound. Cambridge, UK: Cambridge University Press, 2002

[71] Dobashi Y, Yamamoto T, Nishita T. Synthesizing sound from turbulent field using sound textures for interactive fluid simulation. Computer Graphics Forum, 2004, 23 (3): 539-546

[72] Selfridge R, Moffat D, Reiss J D. Real-time physical model for synthesis of sword swing sound. //Proceedings of the Sound and Music Computing, Espoo, Finland, 2017: 299-305

[73] Selfridge C, Moffat D, Reiss J D. Physically derived sound synthesis model of a propeller. //Proceedings of the Audio Mostly Conference, London, UK,2017: 16:1-16:8

[74] Verron C, Drettakis G. Procedural audio modeling for particle-based environmental effects. //Proceedings of the Audio Engineering Society Convention 133, San Francisco, USA, 2012: 1-11

[75] Schwarz D. Corpus-based concatenative synthesis. IEEE Signal Processing Magazine, 2007, 24(2): 92-104

[76] Cao S, Wu Y, Cheng W. Research on piano sound simulation spectrum model. Journal of electronic measurement and instrumentation, 2017, 31(1):125-131 (in Chinese)
(曹莎莎，吴永忠，程文娟．钢琴乐声仿真频谱模型研究．电子测量与仪器学报, 2017, 31(1):125-131)

[77] Verron C, Grégory P, Aramaki M, *et al*. Controlling a spatialized environmental sound synthesizer. //Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA,2009: 321-324

[78] Conan S, Derrien O, Aramaki M, *et al*. A synthesis model with intuitive control capabilities for rolling sounds. IEEE/ACM Transactions on Audio Speech and Language Processing, 2014, 22(8):1260-1273

[79] Lejemble T, Fondevilla A, Durin N, *et al*. Interactive procedural simulation of paper tearing with sound. //Proceedings of the ACM SIGGRAPH Conference on Motion in Games, Paris, France,2015: 143-149

[80] Cardle M, Brooks S, Bar-Joseph Z, *et al*. Sound-by-numbers: motion-driven sound synthesis. //Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, San Diego, USA,2003: 349-356

[81] Bar-Joseph Z, Lischinski D, Werman M, *et al*. Granular synthesis of sound textures using statistical learning. //Proceedings of the International Computer Music Conference, Beijing, China,1999: 178–181

[82] Picard C, Tsingos N, Faure F. Retargetting example sounds to interactive physics-driven animations. //Proceedings of the AES 35th International Conference on Audio for Games, London, UK,2009:1-8

[83] Schreck C, Rohmer D, James D L, *et al*. Real-time sound synthesis for paper material based on geometric analysis. //Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Zurich, Switzerland,2016: 211-220

[84] Su F, Joslin C. Procedurally-generated audio for soft-body animations. //Proceedings of the Audio Mostly Conference, Wrexham, UK,2018: 16:1-16:8

[85] Peltola L, Erkut C, Cook P R, *et al*. Synthesis of hand clapping sounds. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15 (3): 1021-1029

[86] Doel K V D, Kry P G, Pai D K. FoleyAutomatic: physically-based sound effects for interactive simulation and animation. //Proceedings of the ACM SIGGRAPH, Los Angeles, USA,2001: 537-544

[87] Ren Z, Yeh H, Klatzky R L, *et al*. Auditory perception of geometry-invariant material properties. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(4): 557-566

[88] An S S, James D L, Marschner S. Motion-driven concatenative synthesis of cloth sounds. ACM Transactions on Graphics, 2012, 31 (4): 102:1-102:10

[89] Yin Q, Liu S. Sounding solid combustibles: non-premixed flame sound synthesis for different solid combustibles. IEEE Transactions on Visualization and Computer Graphics, 2018, 24 (2): 1179-1189

[90] Cheng H, Liu S. Liquid-solid interaction sound synthesis. Graphical Models, 2019, 103(101028): 1-11

[91] Gebru I D, Ba S O, Li X, *et al*. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1086-1099

[92] Lathuilière S, Massé B, Mesejo P, *et al*. Neural network based reinforcement learning for audio-visual gaze control in human-robot interaction. Pattern Recognition Letters, 2019, 118: 61-71

[93] Xuan H, Zhang Z, Chen S, *et al*. Cross-modal attention network for temporal inconsistent audio-visual event localization. //Proceedings of the AAAI, New York, USA,2020: 279-286

[94] Gebru I D, Alameda-Pineda X, Forbes F, *et al*. EM algorithms for weighted-data clustering with application to audio-visual scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(12): 2402-2415

[95] Ma D, Su Z, Wang W, *et al*. FPETS: fully parallel end-to-end text-to-speech system. //Proceedings of the AAAI,New York, USA, 2020: 8457-8463

[96] Aggarwal V, Cotescu M, Prateek N, *et al*. Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech. //Proceedings of the ICASSP, Barcelona, Spain,2020: 6179-6183

[97] Li N, Liu Y, Wu Y, *et al*. A robust transformer-based text-to-speech model. //Proceedings of the AAAI, New York, USA, 2020: 8228-8235

[98] Fu R, Tao J, Wen Z, *et al*. Focusing on attention: prosody transfer and adaptative optimization strategy for multi-speaker end-to-end speech synthesis. //Proceedings of the ICASSP, Barcelona, Spain, 2020: 6709-6713

[99] Ai Y, Ling Z. A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2020, 28: 839-851

[100] Ping W, Peng K, Gibiansky A, *et al*. Deep voice 3: scaling text-to-speech with convolutional sequence learning. //Proceedings of the ICLR, Vancouver,

Canada,2018: 1-16

[101] Owens A, Isola P, Mcdermott J, *et al*. Visually indicated sounds. //Proceedings of the IEEE CVPR, Las Vegas, USA,2016: 2405–2413

[102] Zhou Y, Wang Z, Fang C, *et al*. Visual to sound: generating natural sound for videos in the wild. //Proceedings of the IEEE CVPR, Salt Lake City, USA,2018: 3550-3558

[103] Donahue C, McAuley J, Puckette M. Synthesizing audio with generative adversarial networks. 2018, CoRR abs/1802.04208

[104] Chen K, Zhang C, Fang C, *et al*. Visually indicated sound generation by perceptually optimized classification. //Proceedings of the ECCV Workshops, Munich, Germany, 2018: 560-574

[105] Aytar Y, Vondrick C, Torralba A. SoundNet: learning sound representations from unlabeled video. //Proceedings of the NIPS, Barcelona, Spain,2016: 892-900

[106] Ghose S, Prevost J J. AutoFoley: artificial synthesis of synchronized sound tracks for silent videos with deep learning. 2020, arXiv:2002.10981

[107] Cheng H, Li S, Liu S. Deep cross-modal synthesis of environmental sound. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(12): 2047-2055 (in Chinese)
(程皓楠, 李思佳, 刘世光. 深度跨模态环境声音合成. 计算机辅助设计与图形学学报, 2019, 31(12): 2047-2055)

[108] Chen P, Zhang Y, Tan M, *et al*. Generating visually aligned sound from videos. IEEE Transactions on Image Processing, 2020, 29:8292-8302

[109] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al*. Audioset: an ontology and human-labeled dataset for audio events. //Proceedings of the ICASSP, New Orleans, USA, 2017: 776-780

[110] Zhang Z, Wu J, Li Q, *et al*. Generative modeling of audible shapes for object perception. //Proceedings of the ICCV, Venice, Italy,2017: 1260-1269

[111] Bianco M J, Gerstoft P, Traer J, *et al*. Machine learning in acoustics: Theory and applications. The Journal of the Acoustical Society of America, 2019, 146(5): 3590-3628

[112] Johann B, Kyle C, Gilles L, *et al*. Constraining effective field theories with machine learning. Physical Review Letters, 2018, 121(11): 111801

CHENG Hao-Nan, Ph.D. student. Her research interests including computer graphics and sound synthesis.

ZHANG Jia-Wan, Ph.D., professor. His research interests including computer graphics and visualization.

Background

Sound synthesis is an important issue in computer graphics. One of the central goals for the field of computer graphics is the simulation of realistic natural phenomena, while generating convincing scenes requires not only the visual aspects of the scene, but its audio components (environmental sound) as well. Therefore, in recent years, several sound synthesis methods were proposed for different virtual objects, such as rigid body, fire, water, etc. Some of these methods have been widely utilized in many areas such as animation, video games and virtual reality. However, although there have been a large number of researches on sound synthesis, most of the existing work is aimed at synthesizing the sound of a specific category. There does not yet exist a comprehensive survey of sound synthesis at present.

In this paper, we classify and summarize the existing environmental sound synthesis methods which were utilized in virtual scene, and divides them into four categories: physical model based, signal model based, hybrid model based and deep learning model based environmental sound synthesis methods. According to the different shapes of objects, we further subdivide the physical model into solid acoustic model, liquid acoustic model and aeroacoustic model. According to the different signal types, the signal processing based sound synthesis method can be further divided into two categories: recording based and signal based. According to the different mixing strategies of the model, the hybrid model based sound synthesis method is further divided into two types: parameter estimation method and detail enhancement method.For each subcategory, we discuss the research progress, advantages and disadvantages of each method in detail. Based on the discussion of existing methods, we summarize the main problems and challenges in the current research on environmental sound synthesis, and also discuss and look forward to some future research subjects on this topic.

Our group has researched on sound synthesis of liquid-solid interaction and proposed some signal processing

based and hybrid model based sound synthesis methods.This work was supported by National Key Research and Development Program of China under Grant No.2019YFC1521200.