第十一章 音频

听觉是 VR 重要的一部分，但直到本章才开始被正式提及。VR 系统的开发人员往往主要关注视觉部分，因为它是我们最强烈的感觉；然而，VR 的音频组件功能强大，并且存在将高保真音频体验带入 VR 的技术。在现实世界中，音频对艺术，娱乐和口头交流至关重要。正如第 2.1 节所述，音频录制和复制本身也可以被视为 VR 体验，既有穴状的版本（环绕声），也有头戴式的版本（戴着耳机）。 当其与视觉部分结合在一起时，音频有助于提供引人注目和令人舒适的 VR 体验。

本章的每一部分都是对第 4 章到第 7 章中的听觉（或音频）补充。整个过程从物理学到生理学，然后从感知到渲染。 第 11.1 节以波，传播和频率分析来解释声音的物理特性。 11.2 节描述了人耳的各个部分及其功能。这自然会引导到听觉部分，便是第 11.3 节的内容。第 11.4 节呈现了听觉渲染的过程，可以通过模型合成，产生声音或再现所采集的声音。在阅读这些章节时，务必记住每个主题的视觉部分。其中的相似之处使得它更容易被理解，差异则导致了不同寻常的工程解决方案。

11.1 声音的物理学

本节与第 4 章中的许多概念相似，后者涵盖了光的基本物理学。声波传播在许多方面与光相似，但有一些关键的差异，会导致重大的感知和工程后果。光是横波，在垂直于其传播的方向上振荡，声音是纵波，在平行于其传播的方向上振荡。图 11.1 给出了一个平行波前的例子。

声音对应于介质中的振动，通常是空气，但也可以是水或任何其他气体，液体或固体。真空中没有声音，这与光传播不同。对于声音来说，介质中的分子发生位移，引起压力变化，从极端压缩到极端稀疏解压缩。在空间的固定点，压力随时间而变化。最重要的是，这可能是人耳鼓膜上的压力变化，它会转化为感知体验。声压水平通常以分贝（简写为 dB）做单位，其定义如下

$$N_{db} = 20 * \log_{10}(p_e/p_r). \qquad (11.1)$$

上式中，pe 是峰值压缩的压力水平，pr 是参考压力水平，通常取 2×10-7 牛顿/平方米。声波通常通过振动固体材料产生，特别是当它们相互碰撞或相互作用时。一个简单的例子是敲击一口大钟，使其振动数秒钟。通过足够的空气流动，材料也可能会被强制振动，例如吹长笛的情形。人体通过使用肺来强制空气通过声带产生声音，从而使声音振动。 这使得我们会说话，唱歌，尖叫等等。

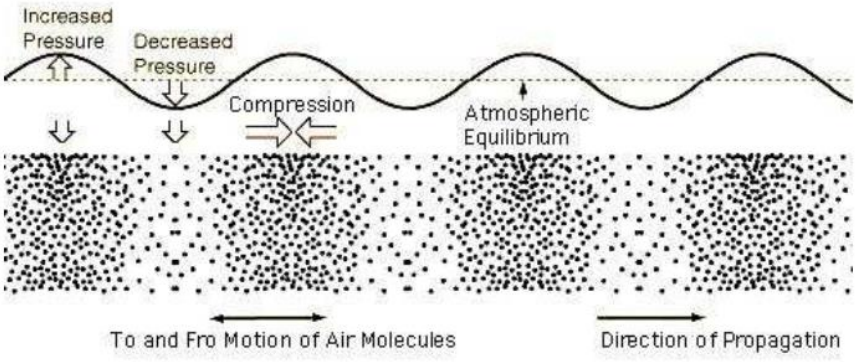

图 11.1 声音是空气分子的纵向压缩和稀疏波。 这里显示纯音的情况，这生成了正弦函数。 （图源于 Pond Science Institute）

声源和衰减

就光线而言，我们可以考虑每条声线垂直于声音传播波阵面的光线。可以定义一个点声源，它可以在各个方向产生相同功率的发射射线。这也导致功率关于离源距离的函数以二次方速率降低。这样的点源对建模很有用，但在现实世界中很难实现。平面波前可以通过振动大而平坦的平板来实现，其导致平行光的声学等效。然而，一个重要的区别是声音在媒介中传播时的衰减。由于分子振动中的能量损失，对于距离平面源的每个单位距离，声音强度以固定因子（或固定百分比）降低；这是一个指数衰减的例子。

传播速度：声波以每秒 343.2 米的速度在 20° C（68° F）的空气中传播。作为比较，光传播速度快了约 874,000 倍。 我们有飞机和汽车可以超过声速，但远不及以光速行驶。 这可能是制作 VR 系统的声音和光线之间最重要的区别。其结果是人类感官和工程传感器可轻松测量声波到达时间的差异，从而更强调时间信息。

频率和波长

如 4.1 节所述，将波分解为频率分量变得重要。对于声音，频率是每秒压缩的次数，称为音频。通常认为该范围是从 20Hz 到 20,000Hz，这是基于人类听觉的，与光的频率范围基于人类视觉的方式大致相同。超过 20,000 赫兹的振动被称为超声波，可以被某些动物听到。低于 20 Hz 的振动称为次声波。

使用 4.1 节中的（4.1）和传播速度 $s = 343.2$，还可以确定声波的波长。在 20Hz 时，波长 $\lambda = 343.2 / 20 = 17.1\text{m}$。在 20,000 赫兹时，它变为 $\lambda = 17.1\text{mm}$。 这大约是我们世界中物体的大小。所以，这会导致声音会以复杂的方式被一些物体所干扰，在尝试重现 VR 中的行为时很难进行建模。相比之下，光波的波长很小，范围从 400 纳米到 700 纳米。

多普勒效应

上述声压变化是针对固定的接收点。如果该点远离声源，则波前将以较低的频率到达。例如，如果接收器以 43.2m / s 的速度离开声源，那么波似乎仅以每秒 343.2-43.2 = 300 米的速度行进。接收到的频率由于声源和接收器之间的相对运动而发生变化。这就是所谓的多普勒效应，在接收器处测得的频率可以计算为

$$f_r = \left(\frac{s + v_r}{s + v_s} \right) f_s, \tag{11.2}$$

其中 s 是介质中的传播速度，vr 是接收器的速度，vs 是声源的速度，fs 是声源的频率。在我们的例子中，$s = 343.2$，$vr = -43.2$，$vs = 0$。结果是，频率 $fs = 1000\text{Hz}$ 的声源将被接收器感知为具有频率 $fr \approx 876.7$。这就是为什么警笛似乎在警车经过时改变声音的原因。多普勒效应也适用于光照，但在正常的 VR 环境下效果可以忽略不计（除非开发人员想要试验虚拟时间膨胀，太空旅行等等）。

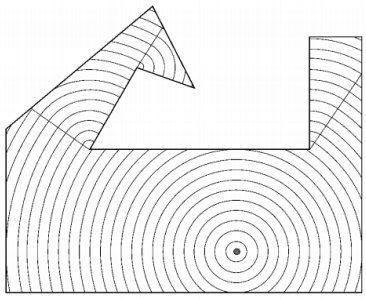

图 11.2：由于衍射，波甚至可以在角落弯曲。这里显示了房间的自顶向下视图。在三个内角处，传播的波前都在其周围扩大。

反射和传播

与光一样，波传播受到媒介传播的强烈影响。想象一下，当某人在房间里面大喊大叫的

时候，声波就会冲击内墙。由于反射，大部分声音都会被弹起来，就好像墙壁是一面镜子一样。但是，一些声能会穿透墙壁。声音通过坚固的材料传播得更快，导致其穿透时弯曲。这是折射。有些声音从墙壁的另一边穿过并通过相邻房间的空气传播，导致进一步地传播。因此，邻近房间里的人可以听到大喊。声波在撞击到墙壁之前所含的总能量会被反射和折射分开，并且由于衰减而产生额外的损失。

衍射

波前也可以在拐角处弯曲，这就是所谓的衍射；见图 11.2。这将使人们能够听到建筑物角落的声音，而不依赖任何反射或传输。更长的波长会发生更多的衍射；因此，较低音频的声音更容易在角落附近弯曲。也解释了为什么相比光衍射，我们更关心房间的声衍射，尽管后者通常对透镜来说很重要（回想第 7.3 节的菲涅耳透镜的缺点）。

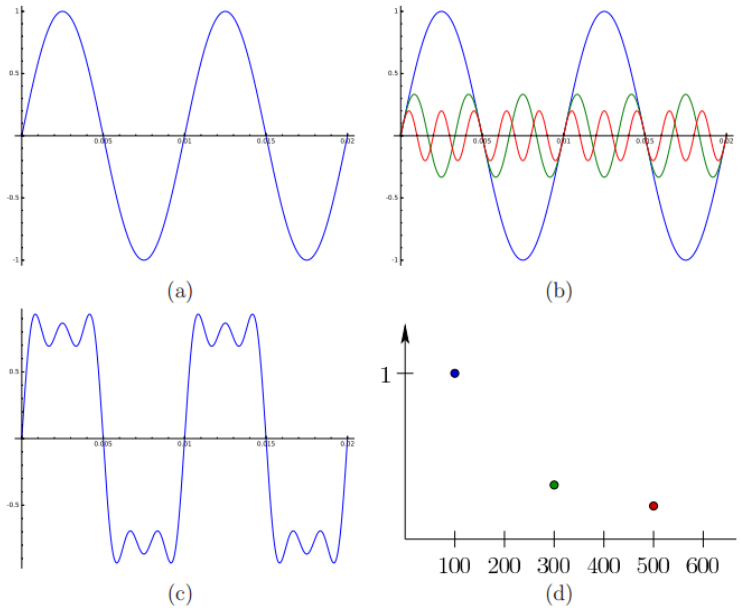

图 11.3：（a）单位幅度和频率 100Hz 的纯音（正弦曲线）。（b）三种纯音；除原始蓝色外，绿色正弦波幅度为 1/3，频率为 300 赫兹，红色波幅为 1/5，频率为 500 赫兹。（c）直接添加三个纯音，会产生接近方形的波形。（d）在频谱中，有三个非零点，每个纯音一个。

傅里叶分析

光谱分解对于表示 4.1 节中的光源和反射非常重要。在声音的情况下，它们甚至更重要。如图 11.3（a）所示，正弦波对应于纯音，其具有单个相关频率;这类似于来自光谱的颜色。更复杂的波形，例如钢琴音符的声音，可以由各种纯音的组合构成。图 11.3（b）至 11.3（d）提供了一个简单的例子。这个原理来自傅里叶分析，它可以通过简单地将它们相加来将任何周期函数分解为正弦曲线（在我们的例子中是纯色调）。每个纯音都有一个特定的频率，幅度或比例因子，以及一个可能的峰值定时，称为相位。通过简单地添加有限数量的纯音，实际上任何有用的波形都可以近似地逼近。较高频率的低幅度正弦曲线通常称为高次谐波；最大的幅度波被称为基频。作为频率函数的幅度和相位的曲线通过应用傅里叶变换获得，将在 11.4 节中简要介绍。

镜头在哪里？

在这一点上，与第 4 章相比最明显的遗漏是镜头的声学等价物。如上所述，折射发生在声音上。为什么人耳不像眼睛那样可以将声音聚焦在空间图像上呢？一个原因是与光相比较长的波长。回顾第 5.1 节，中心窝的感光密度接近可见光的波长。“耳朵凹陷”可能要跨越几米或更多，这种结构下，人类的头会非常大。另一个问题是低频声波以更复杂的方式与世

界中的物体相互作用。因此，我们的耳朵不是形成图像，而是通过执行傅里叶分析来筛选出各种频率，振幅和相位的正弦波形式的声波结构。每个耳朵更像是以每秒数万帧操作的单像素相机，而不是以较慢的帧速率捕捉较大的图像。听觉是在时间上产生重要作用，而视觉主要从空间上产生作用。尽管如此，时间和空间对听力和视觉都很重要。

11.2 人类听觉的生理学

人耳将声压波转换为神经冲动，最终导致感知的体验。人耳的解剖结构如图 11.4 所示。基于声波的流动，耳朵分为外部，中部和内部。回顾第 5.3 节眼球运动的复杂性。猫和其他动物都可以旋转它们的耳朵，但人类不能，这简化了 VR 工程的相关问题。

外耳

从头部突出的耳朵的松软部分被称为耳廓。它主要作为收集声波并将其导入漏斗状的耳道内。它具有放大 1500 至 7500Hz 频率范围内的声音的效果[362]。它还会对声音进行微妙的过滤，导致高频范围内的某些变化取决于声源的射入方向。这为声源的方向提供了强有力的线索。

沿着耳道行进后，声波使耳膜振动。耳膜是一个锥形膜，将外耳与中耳分开。它只覆盖了 55 平方毫米的面积。如果这是一台相机，那么在这一点上它将具有一个像素的分辨率，因为除了可以从膜振动推断出的信息之外不存在额外的空间信息。

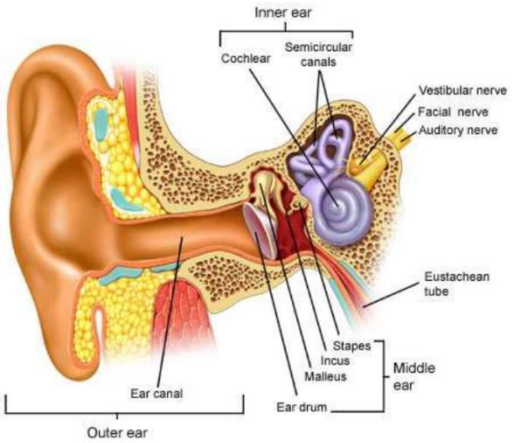

图 11.4：人类听觉系统的生理学结构

中耳

中耳的主要功能是将外耳中的振动空气分子转换成内耳中的振动液体。这是通过将耳膜连接到内耳的骨骼完成的。内耳的空气和液体具有不同的阻抗，这是抗振动的。骨骼被称为锤骨（锤），砧骨（砧）和镫骨（镫），它们通过肌肉和韧带串联连接，允许相对运动。骨骼的目的是匹配阻抗，使压力波以尽可能小的功率损失传输到内耳。这避免了阻抗较高的材料反射声音的情况。这方面的一个例子是声音在湖面上发生的是反射，而不是传播到水中。

内耳

内耳包含 8.2 节所述的前庭器官和耳蜗，它是听觉的器官。耳蜗通过机械感受器将声音能量转化为神经冲动。这是以一种完美的方式完成的，该方法在该过程中执行谱分解，以便神经脉冲对频率分量的幅度和相位进行编码。

图 11.5 说明了它的过程。如图 11.5（a）所示，耳膜振动转换为耳蜗底部卵圆窗的振荡。含有被称为外淋巴液的液体的管从椭圆形窗口延伸到另一端的圆窗。基底膜是穿过耳蜗中心的结构，其大致使含有外淋巴管的长度加倍。管的第一部分称为前庭阶，第二部分称为鼓阶。随着椭圆形窗口的振动，波沿着管道传播，导致基底膜移位。在基部附近（靠近椭圆形和圆形的窗户），膜是薄的和刚性的，并且在最远的点（称为顶点）上逐渐变软和松软；见图 11.5

（b）。这导致膜上的每个点仅在特定的窄频率范围内振动。

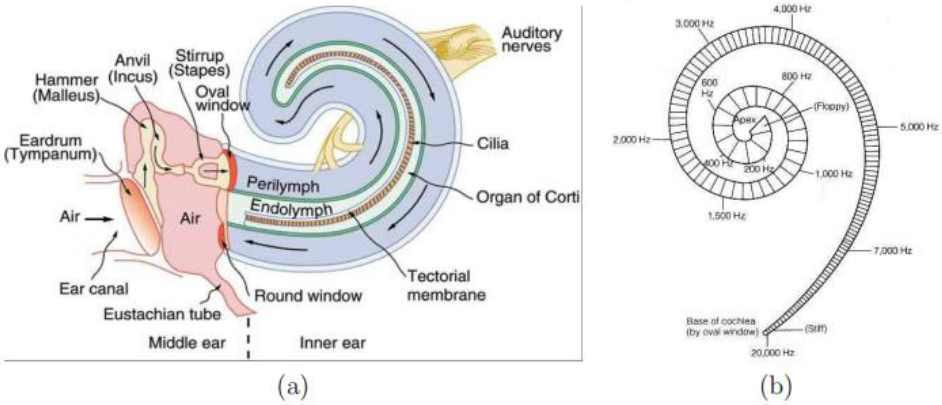

图 11.5：耳蜗的运作：（a）外淋巴传送由椭圆形窗口强迫形成的波，通过延伸耳蜗长度并返回到圆窗的管。 （b）由于厚度和硬度的变化，中心脊柱（基底膜）对特定频率的振动敏感；这触动了机械刺激感受器，并最终听觉感知，这是频率敏感的。

机械刺激感受器

基底膜被称为哥蒂氏器官的更大更复杂的结构所包围，它还含有类似于 8.2 节所示的机械刺激感受器。见图 11.6。机械刺激感受器将毛发的位移转化为神经冲动。当基底膜振动时毛发会移位，因为一些毛细血管的末端附着在盖膜上。基底膜和盖膜的相对运动引起剪毛动作，使毛发移动。每只耳朵含有约 20,000 个机械刺激感受器，这比眼睛中 1 亿个感光器少得多。

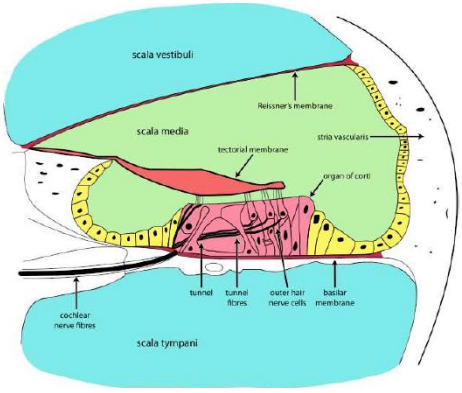

图 11.6：螺旋器器官的横截面。基底膜和盖膜相对于彼此移动，导致机械性感受器中的毛发弯曲。（来自多个维基百科用户。）

光谱分解

通过利用基底膜基于频率的灵敏度，大脑可以有效地获得入射声波的频谱分解。它与第 11.1 节讨论的傅立叶分解类似，但并不完全相同。在第 4 章的[204]提到了一些二者之间的不同之处，如果两个不同频率的单音同时出现在耳朵上，那么可以听到基底膜产生的第三种音调[149]。此外，由机械性感受器输出引起的神经冲动与频率幅度不成正比。此外，单音检测可能会导致检测到附近的音调（频率）被抑制[277]，非常类似于水平单元中的横向抑制（5.2 节）。第 11.4.1 节将阐明这些差异如何使人耳在滤波方面更加复杂。

听觉途径

神经脉冲从左耳蜗和右耳蜗传播到听觉的最高水平中枢，也就是大脑中的主要听觉皮层。像往常一样，当信号通过神经结构组合时发生分层处理，这使得可以分析多个频率和相移。

上级橄榄的早期结构从两个耳朵接收信号，以便可以处理幅度和相位的差异。这在第 11.3 节中对于确定音频源的位置将变得重要。在最高水平上，初级听觉皮层是以 tonotopically（位置基于频率）绘制出来的，与视觉皮层的表面图非常相似。

11.3 听觉感知

现在我们已经了解了听觉器官，下一部分是知道我们如何感知声音。在视觉案例中，我们看到感知体验往往令人惊讶，因为它们基于适应，缺失数据，由神经结构填充的假设以及许多其他因素。听觉体验也是如此。此外，听觉幻象与光学幻觉一样存在。第 6.4 节的 McGurk 效应是一个使用视觉诱发不正确听觉的例子。

优先效应

更常见的听觉错觉是优先效应，如果两个几乎相同的声音在稍微不同的时间到达，则只有一个声音被感知，见图 11.7。声音经常从表面反射，引起混响，这是由于从反射，透射和衍射获取的不同传播路径而引起的声音延迟“副本”到达耳朵造成的。但人们通常听到的仍是一个声音，而不是听到混乱的声音。这通常由具有最大振幅声波决定。如果定时差异大于回声阈值（在一项研究中，其范围为 3 至 61ms [358]），则会感知到回声。其他听觉幻觉会产生不正确的定位（Franssen 效应和 Glissando 幻觉[60]），幻觉的连续性[339]以及永久增加的音调（Shepard tone illusion [285]）。

图 11.7：由于优先效应，如果头部放置在立体声扬声器之间，则会出现听觉错觉，因此一个比另一个靠得更近。如果它们同时输出相同的声音，则人会感知到来自靠近扬声器的声音，而不是感知回声。

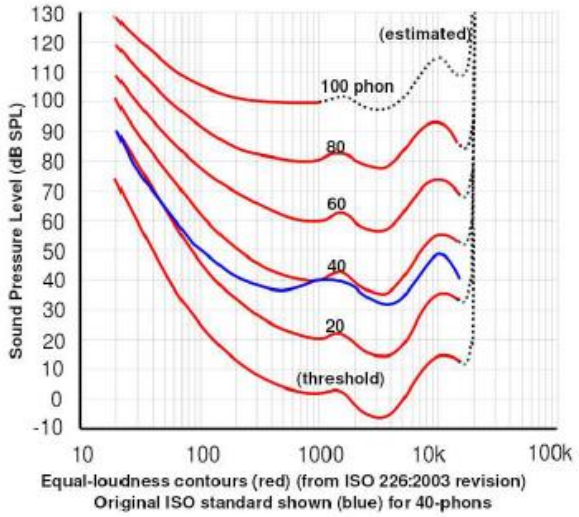

图 11.8：作为频率函数的等响度感知等高线。

心理声学和响度感知

第 2.3 节介绍的心理物理学领域专门用于听觉感知情况下的心理声学。史蒂文斯的感知刺激强度定律和韦伯定律的明显差异（JNDs）是重要的理论基础。例如，史蒂文斯定律的指数（回忆（2.1）），对于 3000Hz 纯音的感知响度 $x = 0.67$ [311]。这大致意味着，如果声音增加到更高的压力水平，我们认为它只是更响亮一点。来自心理声学的一个更复杂的例子如图 11.8 所示，它对应于作为频率函数的等响度感知的等高线。换句话说，随着频率的变化，声音被认为有相同的响度时是在什么水平？这需要仔细设计人体对象的实验，这也是 VR 开发过程中常见的问题，见第 12.4 节。

音调感知

当考虑感知时，声波的频率被称为音调。知觉心理学家已经研究了人们检测目标音调的能力，尽管其他波长和相位的声音都有所混淆。一个基本的观察结果是，听觉感知系统执行临界带遮蔽以有效地阻挡具有特定感兴趣范围外的频率的波。另一个深入研究的问题是对音高（或频率）差异的认识。例如，对于 1000 赫兹的单音，有人可以将其与 1010 赫兹的音调区分开来吗？这是 JND 的一个例子。事实证明，对于低于 1000 赫兹的频率，人类可以检测到小于 1 赫兹的频率变化。随着频率的增加，鉴别能力下降。在 10,000 赫兹时，JND 约为 100 赫兹。就百分比而言，这意味着音调感知在低频率方面有优于 0.1%的差异，但对于较高频率则增加到 1.0%。

另外关于音高感知，当基频从复杂波形中移除时，会出现令人惊讶的听觉错觉。回想一下图 11.3，方波可近似表示为增加幅度越来越小但幅度更高的正弦曲线。事实证明，人们可以感觉到基本频率的音调，即使它被移除，只有高次谐波仍然存在，在[204]中的第 5 章总结了几种理论。

定位

心理声学的主要领域之一是定位，这意味着通过听到声音来估计声源的位置。这对许多 VR 体验至关重要。例如，如果人们正在社交，那么他们的声音应该来自相应虚拟形象的嘴巴才会更真实。换句话说，听觉和视觉线索应该匹配。任何类型的声音效果，如汽车经过，都应该有相应的线索。

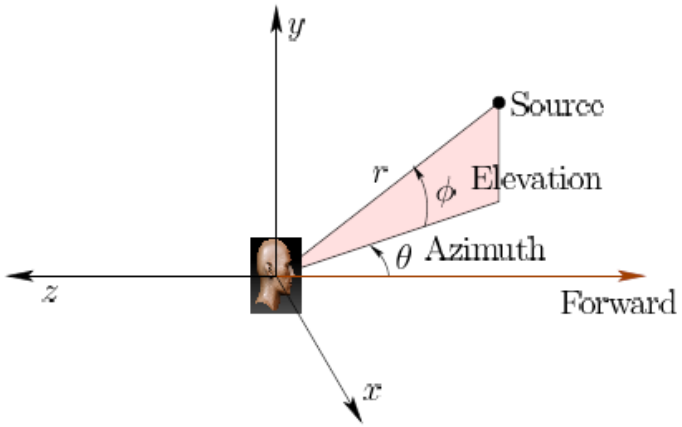

图 11.9：球形坐标用于听觉定位中的源点。假设头部以原点为中心并朝向-z 方向。方位角 θ 是在将源投影到 xz 平面之后相对于正向的角度。仰角 φ 是由一个垂直三角形构成的内角，该垂直三角形将原点连接到光源并将光源投影到平面中。半径 r 是从原点到源的距离。

将 JND 概念应用于定位以获得最小可听角度（MAA），其是可以由人类听觉检测到的角度变化的最小量。通常使用球坐标系进行定位，其中听者头部位于原点，见图 11.9。在正向和源之间的水平面内的角度被称为方位角，其从-180 延伸到 180 度。角度对应于源与水平面的偏差称为高度，从-90 度延伸至 90 度。第三个坐标是从原点（头部中心）到源的半径或距离。MAA 取决于频率和源的方向。图 11.10 显示了 MAA 作为频率函数的曲线图，几个方位角值。变化量令人惊讶。在某些频率和地点，MAA 降至 1 度。但是，在其他组合中，定位非常糟糕。

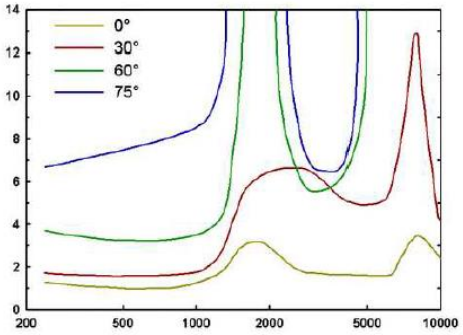

图 11.10：作为频率函数的最小可听角（MAA）图。每个绘图对应于不同的方位角。

单声道提示

听觉定位类似于视觉的深度和尺度感知，6.1 节已对此进行了介绍。由于人类有一对耳朵，因此可以将定位分为使用单耳和使用双耳的两种提示。这类似于单眼和双目视觉提示。单声道提示依靠到达单耳的声音来约束可能的声源集合。几个单声道提示[363]：

1.耳廓的形状是不对称的，所以传入的声音会以取决于它到来的方向，特别是仰角的方式变形。尽管人们没有意识到这种扭曲，但听觉系统将其用于定位。

2.声音的振幅随着距离的变化而呈二次曲线下降。如果它是一种熟悉的声音，那么它的距离可以从感知的幅度来估计。熟悉度影响这种线索的方式与熟悉物体使得深度和尺度感知被分开的方式相同。

3.对于远处的声音，会出现频谱失真，因为高频分量比低频分量衰减更快。例如，遥远的雷声被认为是深沉的隆隆声，但是附近的雷声包含了更高的爆音。

4.最后，随着声音反弹，进入耳朵的混响提供了强大的单声道提示，这在室内体现的尤其强大。尽管优先效应阻止我们感知这些混响，但大脑仍然使用这些信息进行定位。这种提示单独被称为回声定位，这是由一些动物，包括蝙蝠所使用的。有些人可以通过发出咔嗒声或其他尖锐的声音来执行此操作；这使得盲人可以进行声学寻路。

双耳定位

声源定位则使用双耳听到的线索作为结果。最简单的情况是耳间水平差异（ILD），这是每个耳朵听到的声音幅度的差异。例如，一只耳朵可能面对声源，而另一只耳朵则位于声影中（由声源前方的物体造成的阴影与来自光源的阴影相似）。较近的耳朵会比另一耳朵更强烈的振动。

另一种双耳线索是双耳时间差（ITD），与第 9.3 节所述的 TDOA 感知方法密切相关。两个耳朵之间的距离大约为 21.5 厘米，这会导致来自声源的声音到达时间不同。请注意，声音在大约 0.6ms 内传播 21.5 厘米，这意味着定位使用了令人惊讶的小差异。

假设大脑将到达时间的差异测量为 0.3ms。源可能产生的地方是什么？这可以通过设置代数方程来解决，这会形成锥面被称为双曲面。如果不知道哪个声音首先出现，那么这组可能的地方是两个不相交的双曲面。由于大脑知道哪一个先出现，这可能区域被缩小成一个双曲面片，这被称为混淆锥，见图 11.11（在大多数情况下，即使它是双曲面，它看起来大致像一个圆锥体）。然而，通过使用耳廓的变形，可以部分地解决该锥体内的不确定性。

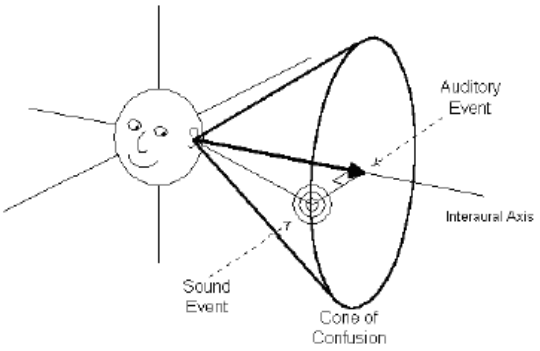

图 11.11：混淆锥是使用 ITD 双耳线索后点源可能存在的位置集合。它在实现上是一个双曲面，但大致看起来像一个锥体。

运动的力量

更重要的是，人类通过简单地移动头部就可以解决很多模糊问题。正如头部运动导致视差带来的强大视觉深度提示一样，它也提供了更好的听觉定位。事实上，听觉视差甚至提供了另一个定位提示，因为附近的音频源会更快地改变它们的方位角和仰角。关于 ITD，想象在短时间内，每个头部姿势都有一个不同的混淆锥。通过整合其他感官，可以估计相对的头部姿势，这便可以大致形成多个混淆锥的交集，直到声源精确定位为止。最后请记住，源相对于接收器的运动会引起多普勒效应。相比于视觉而言，基于听觉的感知结果与物体的实际运动相关联。这对媒介过程是有帮助的（回顾第 8.2 节）。

11.4 听觉渲染

我们现在开始讨论为虚拟世界制作声音并将它们发送到听觉显示器（扬声器）的问题，以便用户感觉它们是为 VR 体验设计的。它们应该与视觉线索和现实世界中的听觉体验一致。无论是录制的声音，合成声音还是其组合，虚拟压力波及对演讲者的渲染，这些虚拟声音都应能欺骗到用户的大脑。

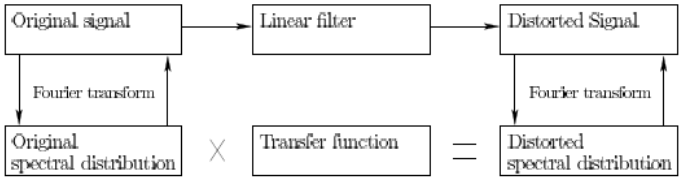

图 11.12：线性滤波器的概述及其与傅里叶分析的关系。顶部块对应于时域，而最下行是频率（或频谱）域。

11.4.1 基本信号处理

到目前为止，声波中频率分量的重要性应该是清楚的。这仍然适用于 VR 合成声音的工程问题，该问题属于信号处理领域。这里给出一个简要的概述，见[11，189]以供进一步阅读。由于这个主题的核心是对信号进行转换或扭曲的滤波器的表征或设计。在我们的例子中，信号是可以完全合成的，如使用麦克风捕获的声波或某种组合。（回想一下，合成模型和捕获模型也存在于视觉案例中。）

图 11.12 显示了整个方案，将在本节中介绍。原始信号出现在左上角。首先，按照从左到右的路径。信号进入一个标有线性滤波器的黑盒并变为失真状态，如右图所示。什么是线性滤波器？在回答这个问题之前，需要一些背景概念。

采样率

信号处理公式既适用于连续时间，它可以提供很好的公式和数学证明，也可以用于离散时间，直接对应于计算机处理信号的方式。由于其实用价值，我们主要关注离散时间案例。

将信号作为时间的函数，其值表示为 x（t）。使用数字信号处理，它将定期进行采样。

设 t 是采样间隔。采样率或（采样频率）大约为 1/Δt Hz。例如，采样频率为 1000Hz 时，Δt 为 1 毫秒。根据奈奎斯特 - 香农采样定理，采样率应该至少是信号中最高频率分量的两倍。由于音频的最高频率分量为 20,000 Hz，这表明采样率应至少为 40,000 Hz。并非巧合，CD 和 DVD 的采样率分别为 44,100 Hz 和 48,000 Hz。

通过对信号进行采样，可以生成一个数组。在 1000Hz 条件下，数组每秒钟将包含一千个数值。使用索引变量 k，我们可以将第 k 个样本称为 x [k]，它对应于 x（kΔt）。例如，第一个样本是 x [0] = x(0)。

线性滤波器

在信号处理的背景下，滤波器的作用是将一个信号映射到另一个信号。每个信号都是时间的函数，滤波器就像一个黑盒子，接收一个信号作为输入，并产生另一个信号作为输出。如果 x 代表一整个信号（全程），那么可以令 F（x）代表通过滤波器运行后得到的信号。

线性滤波器是一种特殊的滤波器，它满足两个代数性质。第一个代数属性是加法性质，这意味着如果两个信号被添加并通过滤波器发送，则结果应该与它们各自独立地通过滤波器并加和后的信号相同。即对于任何两个信号 x 和 x'，有 F（x + x'）= F（x）+ F（x'）。例如，如果两个不同的声音被发送到滤波器中，无论它们在过滤之前还是之后进行组合，结果都应该是相同的。随着多个正弦波通过滤波器，这个概念将变得很有用。

第二个代数属性是同质性的，这意味着如果信号在经过滤波器发送之前通过一个常数因子进行缩放，则结果将与之后由相同因子缩放的结果相同。使用符号表示，对于每个常数 c 和信号 x，有 cF（x）= F（cx）。这意味着如果我们将声音幅度加倍，那么来自滤波器的输出声音也将其幅度加倍。

线性滤波器通常采用这种形式

$$y[k] = c_0x[k] + c_1x[k-1] + c_2x[k-2] + c_3x[k-3] + \cdots + c_nx[k-n], \quad (11.3)$$

其中每个 ci 都是一个常数，并且 n + 1 是与滤波器相关的样本数量。可以考虑 n 趋于无穷的情况，但这里不会。毫不意外，（11.3）是一个线性方程。这种特殊的形式也导致了这种滤波器是因果滤波器，因为等式右边的样本从时间上不会晚于样本 y [k]。非因果性滤波器需要依赖未来的样本，这对记录的信号是合理的，但不适用于现场采样（未来是不可预测的！）。

以下是线性滤波器的一些例子（（11.3）的特例）。它们取最后三个样本的移动平均值：

$$y[k] = \frac{1}{3}x[k] + \frac{1}{3}x[k-1] + \frac{1}{3}x[k-2]. \quad (11.4)$$

这是指数平滑（也称为指数加权移动平均）的示例：

$$y[k] = \frac{1}{2}x[k] + \frac{1}{4}x[k-1] + \frac{1}{8}x[k-2] + \frac{1}{16}x[k-3]. \quad (11.5)$$

有限脉冲响应

一个重要且有用的结果是，线性滤波器的行为可以用其有限脉冲响应（FIR）来充分表征。（11.3）中的滤波器通常被称为 FIR 滤波器。有限脉冲是所有 k> 0 时 x [0] = 1 且 x [k] = 0 的信号。任何其他信号可以表示为时移有限脉冲的线性组合。如果有限冲动发生偏移，例如 x [2] = 1，对于所有 k≠2 且 x [k] = 0 的情况下，线性滤波器会产生相同的结果，但它仅延迟两步。由于滤波器的线性，有限脉冲可以重新调整，输出只是重新调整比例。通过滤波器发送缩放和移位脉冲的结果也可以由线性直接获得。

非线性滤波器

不遵循式（11.3）的任何（因果）滤波器称为非线性滤波器。回忆第 11.2 节，人类听觉系统的操作几乎就是一个线性滤波器，但展现出非线性滤波器的特性。线性滤波器被优先选择，因为它们与信号的频谱分析或频率分量紧密连接。即使人类听觉系统包含一些非线性行为，但基于线性滤波器的分析仍然是有价值的。

返回到傅里叶分析

现在考虑图 11.12 的底层。线性滤波器的操作在频域中易于理解和计算。这是通过对信号进行傅立叶变换而获得的函数，该信号为每个频率和相位的组合提供一个幅度。这个转换在 11.1 节简要介绍过，并在图 11.3 中说明。形式上，它被定义为离散时间系统

$$X(f) = \sum_{k=-\infty}^{\infty} x[k]e^{-i2\pi fk}, \tag{11.6}$$

其中 X（f）是所得到的频谱分布，是频率 f 的函数。指数包含 i = √-1，并通过欧拉公式与正弦曲线相关：

$$e^{-i2\pi fk} = \cos(-2\pi fk) + i\sin(-2\pi fk). \tag{11.7}$$

单位复数被用作表示相位。傅里叶逆变换的形式类似，并将频谱分布转换回时域。这些计算在实践中通过使用快速傅立叶变换（FFT）可以快速执行[11,189]。

传递函数

在一些情况下，需要通过修改谱分布来设计线性滤波器。它可以放大一些频率，同时抑制其他频率。在这种情况下，滤波器根据传递函数来定义，其应用如下：1）使用傅立叶变换来变换原始信号，2）将结果与传递函数相乘以获得失真的频谱分布，最后 3）应用傅里叶逆变换以获得作为时间函数的结果。通过将离散拉普拉斯变换（称为 z 变换）应用于有限脉冲响应，可以从线性滤波器计算传递函数[11,189]。

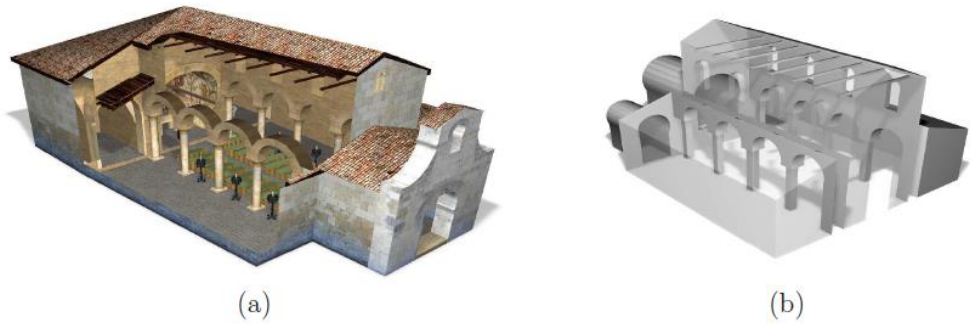

图 11.13：音频模型要简单得多。（来自 Pelzer，Aspock，Schroder 和 Vorlander，2014，[248]）

11.4.2 声学建模

除了视觉方面之外，第 3.1 节中的几何建模概念适用于 VR 的听觉方面。事实上，两种都可以使用相同的模型。在虚拟世界中反射光的墙也反射声波。因此，两者都可以用相同的三角形网格表示。理论上这没问题，但细节的精细程度或空间分辨率对音频无关紧要。由于视觉敏锐度高，为视觉渲染设计的几何模型可能具有高度的细节。回想 5.4 节，人类可以根据不同的视角区分 30 条或更多的条纹。在声波的情况下，小的结构基本上不可感知。一个建议是，声学模型需要的空间分辨率只有 0.5m [334]。图 11.13 显示了一个例子。因此，任何小波纹，门把手或其他精细结构都可以被简化。自动转换为视觉渲染而设计的 3D 模型到

为针对听觉渲染优化的 3D 模型仍然是一个挑战。

现在考虑虚拟环境中的声源。例如，这可能是发出声波或振动平面表面的点。白光的等价物称为白噪声，理论上它包含可听频谱中所有频率的相等权重。来自模拟电视或收音机的纯静电就是一个近似的例子。在实际环境中，感兴趣的声音在特定频率中集中度很高，而不是均匀分布。

声音如何与表面相互作用？这与 7.1 节中的阴影问题类似。对于光线情况，漫反射和镜面反射依赖颜色发生。而在声音的情况下，同样存在两种可能性，同样依赖于波长（或者等同于频率）。对于大而平滑的表面，会发生声波的镜面反射，出射角度等于入射角度。反射的声音通常具有不同的幅度和相位。由于部分声音会被吸收到材料中，所以振幅可能会减小一个常量。该因数通常取决于波长（或频率）。 [334]包含许多常见材料的吸收系数。

对于较小的物体或具有重复结构的表面（如砖块或波纹），声波可能会以难以表征的方式散射。这与光的漫反射相似，但声音的散射模式可能难以模拟和计算。一个不幸的问题是散射行为取决于波长。如果波长比结构（整个物体或波纹）的尺寸小得多或比其大的多，那么表面主要反射声波。如果波长接近结构尺寸，则可能发生显著且复杂的散射。

在建模重任的极端情况下，可以构建双向散射分布函数（BSDF）。BSDF 可以通过放置在不同位置的扬声器和麦克风阵列的组合来估计真实世界中的等效材料，以测量特定方向上的散射。对于波长较大的平面材料来说，这种等效的结果可能很好，但它仍不能处理可能出现在表面上的各种复杂结构和图案。

捕捉声音

声音也可以使用麦克风在现实世界中被捕捉，然后将其带入物理世界。例如，匹配区域可能包含麦克风，这些麦克风会成为现实世界中等效的扬声器。就视频采集而言，制作一个完全捕捉声场的系统是具有挑战性的。在[256]中提出了基于多个麦克风捕获的声音插值的简单而有效的技术。

11.4.3 听觉化

虚拟世界中的声音传播与视觉渲染一样，处理波的传播主要有两种方式。最昂贵的方法是尽可能准确地模拟物理世界，包括计算精确模拟波传播的偏微分方程的数值解。廉价的方法是拍摄可见光线并表征声源，表面和耳朵之间的主要相互作用。两种方法之间的选择还取决于特定的设置；一些系统便涉及这两种计算[208,334]。如果波与环境中的物体相关程度较大，换句话说，频率低并且几何模型具有高度的细节时，则优选数值方法。。在高频或更大，更简单的模型中，基于可见性的方法将被优选。

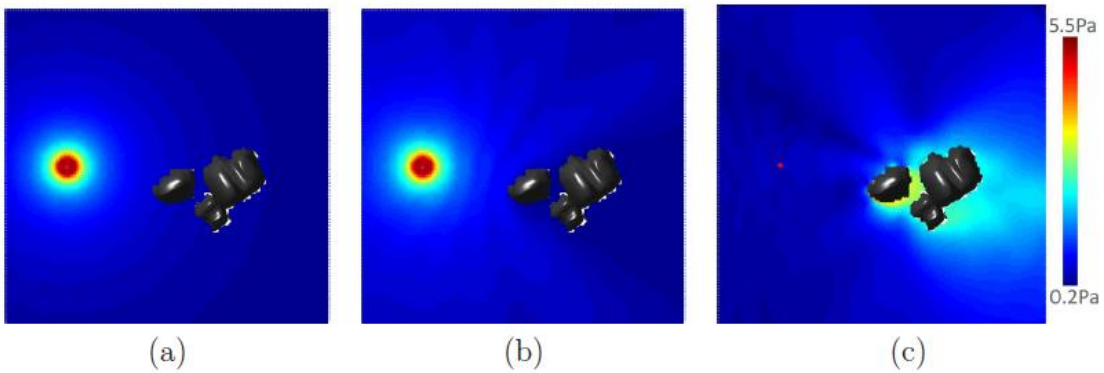

图 11.14：通过数值求解 Helmholtz 波动方程计算声传播结果（取自[208]）：
（a）考虑障碍物相互作用前的压力大小。 （b）考虑到散射后的压力。
（c）散射分量，即（b）的压力减去（a）的压力。

数值波传播

Helmholtz 波动方程根据压力函数的偏导数来表示 R3 中每个点的约束。它基于频率的形式是

$$\nabla^2 p + \frac{\omega^2}{s^2} p = 0, \tag{11.8}$$

其中 ρ 是声压，▽2 是微积分的拉普拉斯算子， ω 与频率 f 有关， ω =2 π f。

（11.8）封闭形式的解决方案不存在，除了微不足道的情况。因此，数值计算是通过迭代更新空间上的值来执行的；[208]中对听觉渲染的方法进行了简要的调查。波动方程是在虚拟世界的无障碍部分上定义的。这个空间的边缘变得复杂，导致边界条件。边界的一个或多个部分与声源相对应，声源可被视为振动物体或将能量带到世界的障碍物。在这些位置，（11.8）中的 0 被替换为强制函数。在其他边界，波可能经历吸收，反射，散射和衍射的某些组合。这些情况极难建模；详情请参见[264]。在一些渲染应用中，这些边界的相互作用可以用简单的 Dirichlet 边界条件和 Neumann 边界条件进行简化和处理[361]。如果虚拟世界是无界的，那么需要额外的 Sommerfield 辐射条件。有关各种设置中声音传播的详细模型和方程式，请参见[264]。图 11.14 显示了一个计算声场的例子。

基于可见性的波传播

数值计算的替代方法是逐步将压力数传播到空间中，这是基于可见性的方法，它考虑从源发出并在障碍物之间反弹的声线路径。这些方法涉及确定射线与几何模型基元的交点，这类似于 7.1 节的射线追踪操作。

了解虚拟世界中声源的脉冲响应是非常具有远见性的。如果环境被认为是线性滤波器，那么脉冲响应为任何其他声音信号提供一个完整的表征[209,248,258]。图 11.15 显示了矩形房间内反射脉冲响应的简单情况。基于可见性的方法在模拟混响方面特别好，这对于感知原因重现非常重要。更一般地说，基于可见性的方法可以考虑与所有反射，吸收，散射和衍射情况相对应的射线。由于表征所有射线的高计算成本，随机射线追踪通过对射线及其与材料的相互作用进行随机采样提供了一种实用的替代方案[334]。这属于蒙特卡洛方法，例如，这些方法用于近似高维集成和优化问题的解决方案。

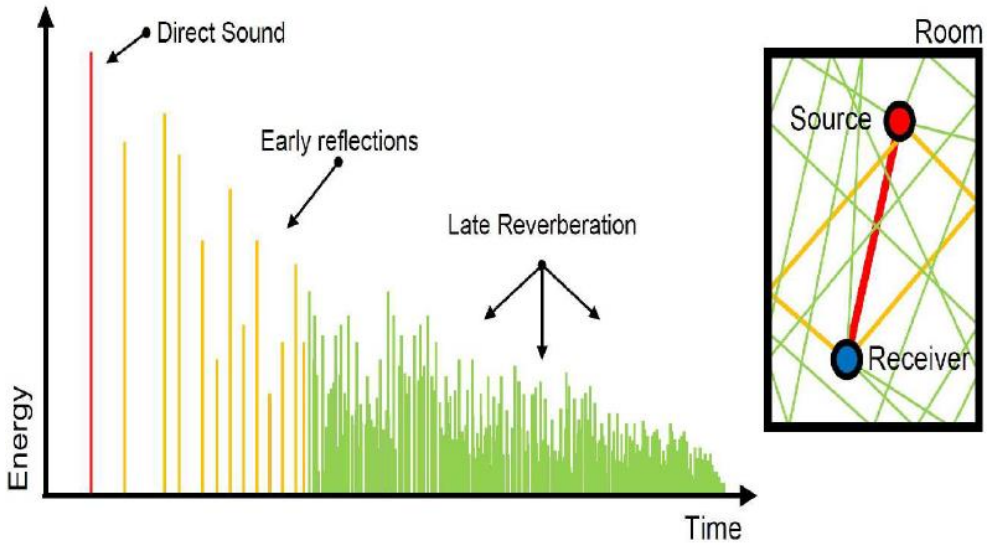

图 11.15：混响。 （来自 Pelzer，Aspock，Schroder 和 Vorlander，2014，[248]）

耳朵输入

在虚拟世界中产生的声音必须传输到物理世界中的每个耳朵。仿佛置于虚拟世界中的虚拟麦克风捕捉模拟声波。然后通过位于耳前的扬声器将其转换为音频输出。回想第 11.3 节，

人类能够从听觉线索本地化声源。如果所有的声音都来自固定扬声器，这在 VR 上会如何发生？实际上，可以通过确保每只耳朵接收到合适的声音幅度和相位来提供 ILD 和 ITD 提示，以确定幅度和时间的差异是否正确。这意味着物理头部必须在虚拟世界的某个细节层次上进行复制，以便正确计算这些差异。例如，耳朵之间的距离和头部的大小可能变得尤为重要。

HRTF

这种解决方案仍然不足以解决混淆的问题。回想第 11.3 节，耳廓的形状会以一种与方向有关的方式扭曲声音。为了充分考虑可能会使输入声音失真的耳廓和头部的其他部分，解决方案是开发一种头部相关传递函数（HRTF）。这个想法是把这种失真看作一个线性滤波器，它可以用传递函数来表征（回忆图 11.12）。这是通过将人体放入消声室并将声源置于头部周围空间的不同位置来完成的。在每个位置，在扬声器上产生脉冲，并且利用放置在人或人体耳道内的小型麦克风记录脉冲响应。通过递增地改变距离，方位角和仰角来选择位置；回忆图 11.10 中的定位坐标。在很多情况下，远场近似是合理的，在这种情况下，距离值较大且固定。这导致 HRTF 仅取决于方位角和仰角。

当然，为每个用户构建一个 HRTF 是不切实际的。使用具有代表性的单个 HRTF 是比较合理的作法；然而，困难在于它在某些应用中可能不够用，因为它不是为个人用户设计的（参见[334]的第 6.3.2 节）。一种妥协方式可能是向用户提供一小部分 HRTF，以解释人群之间的差异，但他们可能无法挑选出最适合其特定的耳廓和头部模型。另一个问题是传递函数可能取决于经常变化的因素，例如戴帽子，穿上带帽或大衣领的外套，或理发。回想一下，适应性几乎发生于整个人类感知和 VR 的所有方面。如果人们适应现实世界中他们头部附近的频繁变化的状况，那么他们也许也会适应并不完美的 HRTF。这个领域仍然存在重要的研究问题。

跟踪问题

最后的挑战是确保物理和虚拟耳朵在区域中相匹配。如果用户转动头部，则应该相应地调整声音。如果声音来自固定源，则在转动头部时应该将其视为固定。这是稳定感的另一个例子。因此，需要跟踪耳朵姿势（位置和方向）以确定适当的“视点”。这相当于用右耳和左耳简单的位置和方向偏移进行头部跟踪。可以跟踪头部朝向，每个耳朵的完整姿态由头部模型确定（回忆图 9.8）。或者，可以跟踪完整的头部姿势，通过偏移变换直接提供每个耳朵的姿态。为了优化性能，用户特定参数可以提供完美匹配：沿着 z 轴从眼睛到耳朵的距离以及耳朵之间的距离。后者类似于 IPD。

进一步阅读

有关声学的数学和计算基础，请参见[264，321]。[204]的[218,362]和第 4 章和第 5 章介绍了生理学和心理声学。本地化在[23]中有详细的介绍。[290]中讨论了混淆的锥体。回声阈值涵盖于[268，358]。

一些基本的信号处理文本是[11,189]。有关听觉显示的概述，请参阅[335]。[256]提供了从心理物理角度简单地放置音频声源的方法。听觉渲染在书[334]中有详细介绍。一些关于听觉渲染的重要文章包括[87,209,248,257,258]