arXiv:2601.06621v1 [eess.AS] 10 Jan 2026

使用双耳空间自适应神经网络（BSANN）进行个人声区的立体声音频渲染

Hao Jiang，Edgar Choueiri

摘要——**本文提出一种用于个人声区（PSZ）的双耳渲染框架，使多个头部跟踪听者能够接收完全独立的立体声音频节目。当前的 PSZ 系统通常依赖单声道渲染，因此无法分别控制左右耳，这限制了空间成像的质量与精度。所提出的方法采用双耳空间自适应神经网络（BSANN），生成针对耳朵优化的扬声器滤波器，从而在多位听者的每只耳朵处重建所需声场。该框架整合了消声室测量的扬声器频率响应、解析建模的换能器指向性以及刚性球头相关传递函数（HRTF），以提高声学精度和空间渲染保真度。显式的主动串扰消除（XTC）阶段进一步改善三维空间感知。实验表明，客观性能指标显著提升，包括区间隔离度（IZI）、节目间隔离度（IPI）和串音消除（XTC）；在 100–20,000 赫兹范围内，对数频率加权值分别为 10.23/10.03 分贝（IZI）、11.11/9.16 分贝（IPI）和 10.55/11.13 分贝（XTC）。逐耳控制、精确声学建模与集成主动 XTC 相结合，形成了一种统一的渲染方法，能够在真实声学环境中实现更高的隔离性能、更强的房间不对称性鲁棒性以及更逼真的空间再现。**

索引词—**个人声区、深度学习、双耳渲染、空间音频、串音消除、声学建模**

I. 引言

PERSONAL 个人声区（PSZ） [1], [2] 旨在共享声学空间中向多位听者提供独立的音频节目，同时尽量减少听者之间的干扰。其应用包括家庭娱乐系统 [3]、汽车音频 [4], [5]、医疗保健和公共环境 [6]、以及大型户外设施 [7]。典型的 PSZ 系统会定义一个亮区（BZ），在该区域内以高保真度重放目标信号；还会定义一个暗区（DZ），在该区域内抑制不需要的信号。声学对比度控制（ACC） [8]–[9]、声压匹配（PM） [10], [11]、振幅匹配（AM） [12],和可变子空间折衷滤波（VAST） [13]–[14] 等经典滤波器设计方法已得到广泛发展，能够在静态聆听条件下实现有效控制。

尽管取得了显著进展，传统 PSZ 方法仍面临根本性限制。由于依赖一组固定测得的声学传递函数（ATF），每当环境变化或听者头部运动导致 ATF 失配时，其性能就会下降。研究表明，房间不对称性、扬声器差异性、测量噪声以及空间采样不足，都会降低可实现的隔离性能 [15]–[18]。此外，与神经网络方法不同，闭式滤波器设计公式通常需要针对不同频率和听者位置反复进行矩阵求逆，计算成本因此大幅增加 [13]。而且，近期研究表明，隔离性能很大程度上取决于听者头部位置的整体配置，包括每个头部的位置和朝向（即其姿态，也就是六自由度的位置和朝向） [19], [20]。由此形成的姿态空间维度极高，闭式方法无法预先计算或密集采样针对特定姿态的滤波器。因此，闭式滤波器计算并不适用于头部跟踪个人声音区，因为每出现一个新的头部姿态，都需要生成一组新的滤波器。因此，依赖预先设计或插值滤波器的方法，难以自然扩展到动态头部跟踪渲染，也难以实现空间精确的立体声和双耳重放。

近年来，基于深度学习的方法已成为声场控制的一种颇具前景的替代方案 [21], [22]。 [23]中提出的空间自适应神经网络（SANN）表明，个人声音区滤波器可以直接根据听者坐标生成，从而无需针对每个位置进行矩阵求逆，并提升了在模拟和真实环境中的鲁棒性。然而，作为一种单声道公式，SANN 无法保留立体声成像或双耳间空间线索，而这两者对于逼真的空间音频再现至关重要。此外，该方法依赖于点声源声学传递函数模拟，因而物理精度有限，在实际扬声器阵列中部署时隔离性能也会下降；在微调期间引入测得的声学传递函数可以提升真实感，但这种校准需要大量现场测量，因此限制了实际应用。

为克服这些局限，我们提出了双耳空间自适应神经网络（BSANN），这是一种用于 PSZ 中双耳音频渲染的框架。BSANN 将原有的单声道架构扩展为：针对多位头部跟踪听者的每只耳朵，分别优化声场。这种逐耳控制通过明确调节所有听者的耳朵，减轻 PSZ 系统对房间不对称性的敏感度，从而在实际声学环境中稳定隔离性能。为提高对真实扬声器—听者系统的保真度，该框架整合了消声室测量的扬声器频率响应、解析建模的换能器指向性 [24], [25], 和刚性球体头相关传递函数 [19], [26], [27]。这些组件共同带来更准确的训练数据和更强的泛化能力

本研究得到 Sound United Corporation 研究基金的资助。（通讯作者：Hao Jiang。）作者来自美国新泽西州普林斯顿 08544 的普林斯顿大学 3D音频与应用声学（3D3A）实验室（电子邮箱：hj3737@princeton.edu；choueiri@princeton.edu）。

以适应真实声学条件。此外，尽管使用测量得到的扬声器频率响应会使该方法依赖于特定的扬声器阵列设计，但除了网络训练期间已设定的条件外，并不会对听音环境施加任何限制。由于测量是在消声环境中进行的，因此这些响应能够很好地泛化到不同房间。

该框架进一步引入了明确的主动串扰消除（XTC）阶段 [28]–[30] ，以实现空间保真度更高的基于扬声器的双耳渲染 [20], [31]。这种集成天然支持头部跟踪双耳渲染，并与逐耳 PSZ 目标相辅相成，形成统一的立体声与双耳架构，从而能够更忠实地传输空间线索，包括双耳时间差（ITDs）和双耳声级差（ILDs）。

本文的贡献主要有三点：

1）我们提出了 BSANN，将单声道 SANN 泛化为逐耳立体声架构。通过对听者的所有耳朵进行显式控制，BSANN 提高了对房间不对称性的鲁棒性，同时实现了具有更高且更一致隔离性能的立体声与双耳 PSZ 渲染。

2）我们通过结合仿真 ATF、消声室测量的扬声器响应、解析建模的换能器指向性以及刚性球头相关传递函数，提高了训练数据的声学真实性。这种物理信息建模提升了学习得到的滤波器的物理准确性，并在实际声学条件下带来更高且更一致的区间隔离度（IZI）和节目间隔离度（IPI），以及更出色的串音消除效果。

3) 我们集成了一个显式的主动 XTC 阶段，以实现精确的空间渲染和强大的串音消除，从而为头部跟踪听者提供高保真的三维成像。

本文其余部分安排如下。第II节介绍所提出的 BSANN 框架，包括按耳朵分别建模的公式、网络架构、物理信息声学建模，以及主动 XTC 阶段的集成。第III节描述扬声器配置、听者几何布局、训练数据集和客观评估指标。第IV节报告测得的客观评估结果，并将所提方法与现有方法进行比较。最后，第V节总结研究发现，讨论立体声和双耳空间区域化框架的意义，并简要补充方法论说明；详细分析留待未来工作。

II. 方法

A. 双耳空间自适应神经网络（*BSANN*）

所提出的 BSANN 框架扩展了 [23] 中介绍的单声道空间自适应神经网络模型，以便为多个头部跟踪听者实现逐耳控制。在这项工作中，我们重点研究两位听者的情况，同时指出该方法可轻松扩展至更多听者。设 $X_\ell(\omega)$ 为扬声器 ℓ 的驱动信号，并设 $H_{e,\,m,\,\ell}(\omega)$ 表示从扬声器 ℓ 到控制点 $m = 1$、.....、M周围耳朵 $e \in \{\mathrm{L}_1, \mathrm{R}_1, \mathrm{L}_2, \mathrm{R}_2\}$。耳朵 e 的控制点 m 处重现的声压为

$$Y_{e,m}(\omega) = \sum_{\ell=1}^{L} H_{e,m,\ell}(\omega)\, X_\ell(\omega). \tag{1}$$

对于立体声 PSZ 渲染，每位听者的左、右节目通道均作为独立输入，BSANN 会根据声区目标调节这些通道对耳朵所有控制点的贡献。

BSANN 根据两位听者的位置预测复值扬声器滤波器组

$$\mathbf{g}(\omega) = [g_1(\omega), \ldots, g_L(\omega)]^{\mathsf{T}}, \tag{2}$$

。令$\mathbf{S}(\omega)$表示四个输入节目通道，分别对应为听者1和听者2提供的左、右节目音频信号：

$$\mathbf{S}(\omega) = \begin{bmatrix} S_{1,\mathrm{L}}(\omega) \\ S_{1,\mathrm{R}}(\omega) \\ S_{2,\mathrm{L}}(\omega) \\ S_{2,\mathrm{R}}(\omega) \end{bmatrix}, \qquad p \in \{1, 2, 3, 4\}.$$

对于每个扬声器 ℓ，BSANN 会预测复数权重$g_{\ell,\,p}(\omega)$，以确定每个节目通道的贡献。因此，

$$X_\ell(\omega) = \sum_{p=1}^{4} g_{\ell,p}(\omega)\, S_p(\omega), \tag{3}$$

或者，等价地，

$$X_\ell(\omega) = \mathbf{g}_\ell^{\mathsf{T}}(\omega)\, \mathbf{S}(\omega), \tag{4}$$

其中$\mathbf{g}_\ell(\omega) = [g_{\ell,1}(\omega) \ldots, g_{\ell,4}(\omega)]^{\mathsf{T}}$。

图1所示的映射由傅里叶位置编码、一个 N 层共享多层感知机（MLP） [32], [33]以及一个最终线性层组成，该线性层在所有离散频率频点上输出$\mathbf{g}(\omega_n)$的实部和虚部。非线性变换表示为

$$\mathbf{g}(\omega_n) = f_\theta(\mathbf{s}), \tag{5}$$

其中$\mathbf{s}$ 表示听者头部位置（耳朵位置通过几何方式获得），θ 表示可训练网络参数。

为支持逐耳立体声重放，BSANN采用一种损失函数，将单声道空间自适应神经网络的目标函数推广到四耳立体声设置。对于离散频率频点$\{\omega_n\}_{n=1}^{N}$，损失项定义如下。

1）亮区准确度：设

$$\tilde{\mathbf{p}}_{\mathrm{T},B}(\omega_n) = \begin{bmatrix} \mathbf{p}_{\mathrm{T,BL}}(\omega_n) \\ \mathbf{p}_{\mathrm{T,BR}}(\omega_n) \end{bmatrix}, \; \tilde{\mathbf{H}}_{\mathrm{B}}(\omega_n) = \begin{bmatrix} \mathbf{H}_{\mathrm{BL}}(\omega_n) \\ \mathbf{H}_{\mathrm{BR}}(\omega_n) \end{bmatrix}, \tag{6}$$

其中$\mathbf{p}_{\mathrm{T},\,\mathrm{BL}}$ 和 $\mathbf{p}_{\mathrm{T},\,\mathrm{BR}}$ 分别表示左、右亮区耳朵的目标声压，$\mathbf{H}_{\mathrm{BL}}$、$\mathbf{H}_{\mathrm{BR}}$ 则是相应的耳朵专属 ATF 子矩阵。

亮区损失定义为

$$\mathcal{L}_1 = \frac{1}{N M_{\mathrm{B}}} \sum_{n=1}^{N} \Big\| \, |\tilde{\mathbf{p}}_{\mathrm{T},B}(\omega_n)| - |\tilde{\mathbf{H}}_{\mathrm{B}}(\omega_n)\mathbf{g}(\omega_n)| \, \Big\|_2^2. \tag{7}$$

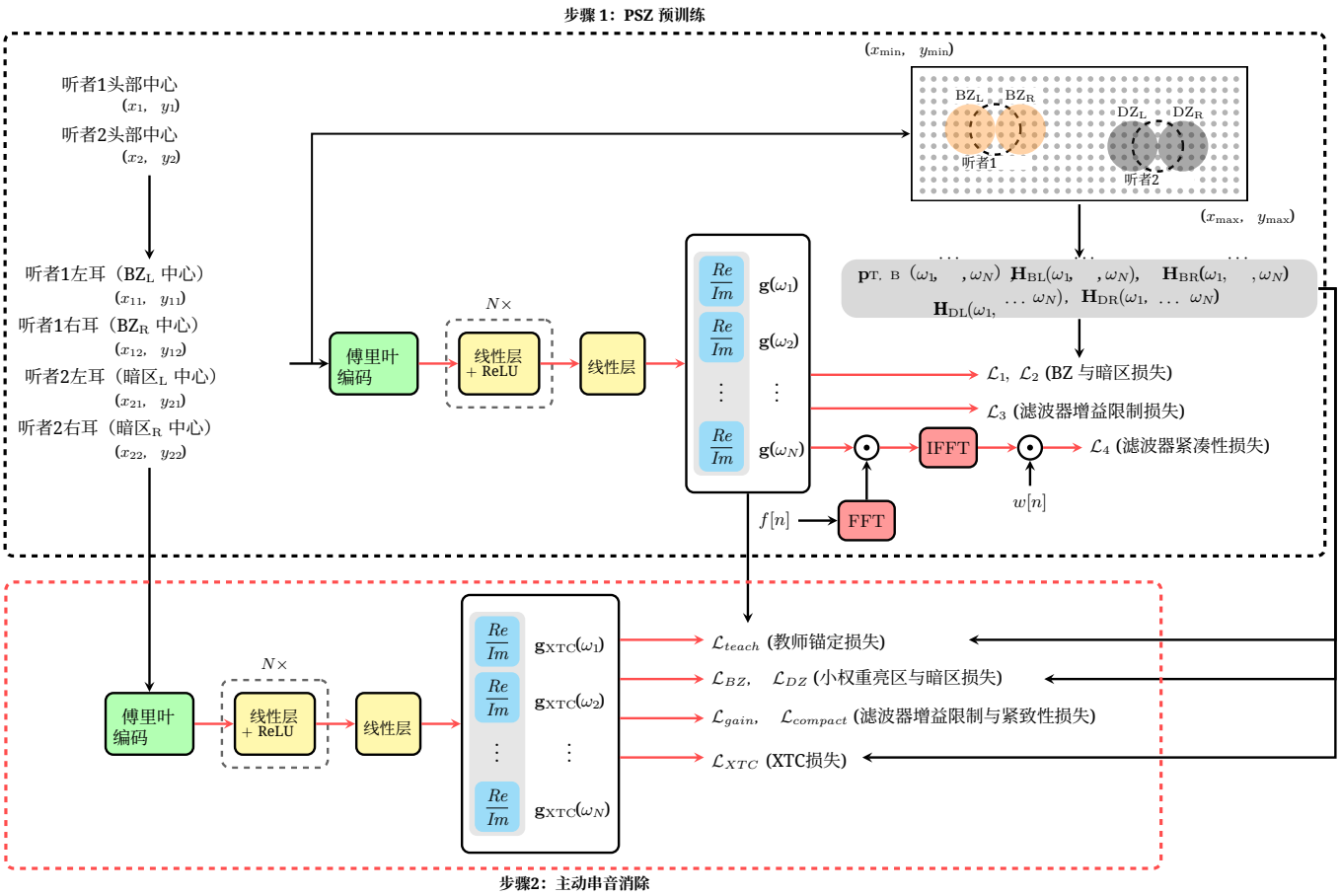


图1. 所提出的两阶段BSANN框架概览。步骤1执行PSZ预训练，以获得频域扬声器滤波器$\mathbf{g}(\omega_n)$；其损失函数由亮区和暗区目标以及增益约束和紧致性约束组成。步骤2将滤波器优化为$\mathbf{g}_{\mathrm{XTC}}(\omega_n)$；其损失函数以主动串音消除项为主，并通过教师锚定、小权重亮区/暗区保护项以及增益约束和紧致性约束进行正则化。

其中， M_B 表示每只耳朵的亮区控制点数量（假设左右耳的数量相同）， N表示计算损失平均值时所涉及的离散频率频点数量。

2) 暗区抑制：定义

$$\tilde{\mathbf{H}}_{\mathrm{D}}(\omega_n) = \begin{bmatrix} \mathbf{H}_{\mathrm{DL}}(\omega_n) \\ \mathbf{H}_{\mathrm{DR}}(\omega_n) \end{bmatrix}, \tag{8}$$

对应于非活动听者的暗区耳朵。暗区损失为

$$\mathcal{L}_2 = \frac{1}{NM_{\mathrm{D}}} \sum_{n=1}^{N} \left\| \tilde{\mathbf{H}}_{\mathrm{D}}(\omega_n)\mathbf{g}(\omega_n) \right\|_2^2, \tag{9}$$

其中，假设暗区的目标声压为零。*3)* 增益限制：为避免滤波器幅值过大，施加增益约束：

$$\mathcal{L}_3 = \frac{1}{NL} \sum_{n=1}^{N} \left\| \max\left(\mathbf{0}, |\mathbf{g}(\omega_n)| - g_{\max}\mathbf{1}\right) \right\|_2^2, \tag{10}$$

其中， $|\mathbf{g}(\omega_n)|$ 是逐元素幅度， $g_{\max}$ 是增益上限， $\mathbf{0}$ 和$\mathbf{1}$ 分别是长度为 L 的全零向量和全一向量，L 是扬声器的数量。

4) 时域紧致性：为抑制时域中较长且振荡的滤波器，在逆快速傅里叶变换后应用渐缩窗，以惩罚晚期能量：

$$\mathcal{L}_4 = \frac{1}{\hat{N}L} \sum_{\ell=1}^{L} \left\| w[n] \cdot \left(f[n] * \hat{g}_\ell[n] \right) \right\|_2^2, \tag{11}$$

其中， $f[n]$ 是一个带通加权滤波器，$\hat{}N$ 是时域滤波器长度。

综合目标为

$$\mathcal{L} = \alpha\,\mathcal{L}_1 + (1-\alpha)\,\mathcal{L}_2 + \beta\,\mathcal{L}_3 + \gamma\,\mathcal{L}_4, \tag{12}$$

其中， α、 β和 γ 是用于平衡四个项贡献的加权参数。

B. 物理信息*ATF* 构造

用于 BSANN训练的声学传递函数由三个基于物理信息的组件构成：在消声环境中测得的扬声器频率响应、用于近似物理换能器指向性的解析活塞指向性模型，以及刚性球头相关传递函数。与理想点声源声学传递函数相比，这种物理信息建模能够更真实地捕捉频率—角度特性。解析活塞指向性

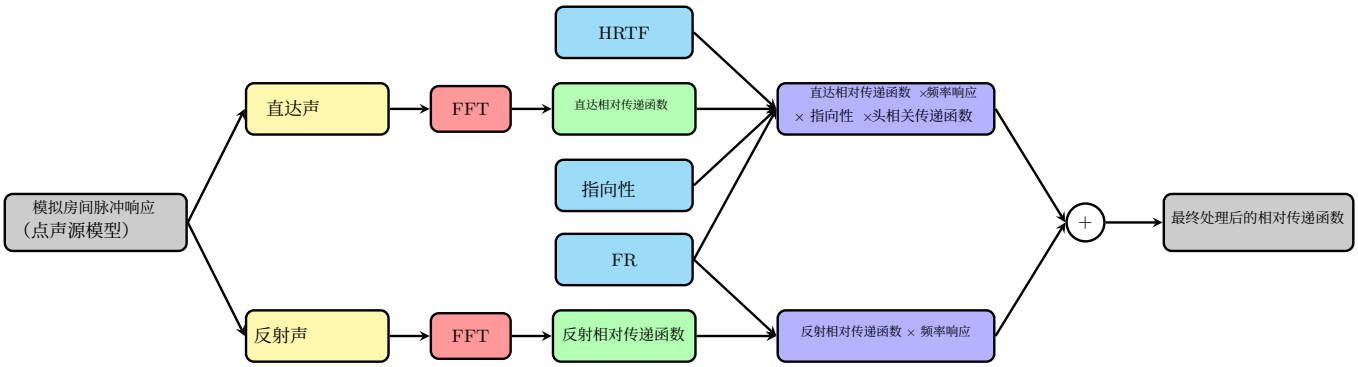


图2. 用于构建BSANN训练所需物理信息声学传递函数（ATF）的处理流程。模拟房间脉冲响应（RIR）被分解为直达分量和反射分量，转换为频域房间传递函数（RTF），并与测得的消声响应、解析活塞指向性和刚性球头相关传递函数相结合，从而得到用于BSANN损失函数的声学传递函数。

和刚性球头相关传递函数完全基于数学模型，因此不依赖任何特定的扬声器阵列或聆听环境。扬声器频率响应是唯一经过测量的组件，与特定的扬声器阵列设计相关；不过，由于这些响应是在消声环境中获得的，因此除网络训练期间已有的假设外，不会施加任何聆听环境限制。图2概述了整体处理流程，其中包含三个主要组件：

1) 消声扬声器频率响应：令 $a_\ell[n]$ 表示扬声器ℓ测得的消声脉冲响应。相应的频率响应为

$$A_\ell(\omega) = \mathrm{FFT}\{a_\ell[n]\}, \tag{13}$$

这些响应体现了扬声器特有的频谱着色。这样可以确保BSANN使用能够代表部署硬件的频谱进行训练，而不是使用理想的平坦响应。

2) 解析活塞指向性：每个扬声器的方向性辐射采用刚性挡板中的圆形活塞模型进行解析建模。对于半径为 a的活塞，其指向性因子随角频率 ω 和观测角 θ 变化的函数表示为

$$D_\ell(\omega,\theta) = \frac{2J_1(ka\sin\theta)}{ka\sin\theta}, \qquad k = \frac{\omega}{c}, \tag{14}$$

其中， $J_1(\cdot)$为一阶贝塞尔函数， c 为声速。该模型无需进行指向性测量，即可再现有限尺寸扬声器的频率相关的波束收窄现象。

3) 刚性球头相关传递函数。耳旁控制点采用归一化刚性球散射公式建模。令$\mathbf{r}_{\mathrm{src}}$ 表示从扬声器指向头部中心的方向（即在刚性球模型中，每个扬声器均视为点声源），并令 $\mathbf{r}_{e,m}$ 表示相对于头部中心、位于耳朵 e 周围的控制点 m 的位置，其半径为 $r_{\mathrm{src}} = \|\mathbf{r}_{\mathrm{src}}\|$, $r_{e,m} = \|\mathbf{r}_{e,m}\|$，角间距为$\gamma_{e,m} = \angle(\mathbf{r}_{\mathrm{src}},\mathbf{r}_{e,m})$。对于波数 $k = \omega/c$，定义散射核

$$\begin{aligned} S_n(k) = {} & j_n(kr_<)\, h_n^{(1)}(kr_>) \\ & - \alpha_n(ka)\, h_n^{(1)}(kr_{\mathrm{src}})\, h_n^{(1)}(kr_{e,m}), \end{aligned} \tag{15}$$

其中， $r_< = \min(r_{\mathrm{src}}, r_{e,m})$、 $r_> = \max(r_{\mathrm{src}}, r_{e,m})$、 $j_n(\cdot)$和 $h_n^{(1)}(\cdot)$分别为球贝塞尔函数和汉克尔函数，而$\alpha_n(ka) = j_n'(ka)/h_n^{(1)\prime}(ka)$表示 n阶刚性球散射系数。

从扬声器方向到耳朵 e周围控制点 m 的归一化刚性球头相关传递函数为

$$H_{\mathrm{HRTF},e,m}(\omega) = \frac{jk}{G_{\mathrm{ff},e,m}(\omega)} \sum_{n=0}^{N_{\max}} (2n+1)\, S_n(k)\, P_n(\cos\gamma_{e,m}), \tag{16}$$

where $P_n(\cdot)$ is the Legendre polynomial and

$$G_{\mathrm{ff},e,m}(\omega) = \frac{e^{-jk\|\mathbf{r}_{\mathrm{src}}-\mathbf{r}_{e,m}\|}}{\|\mathbf{r}_{\mathrm{src}}-\mathbf{r}_{e,m}\|} \tag{17}$$

是用于归一化的自由场格林函数。

4) 最终*ATF*组装。对于每个扬声器–耳朵–控制点三元组 $(\ell,\ e,\ m)$，将模拟点声源房间脉冲响应分解为直达分量$h_{e,\,m,\,\ell}^{\mathrm{dir}}[n]$ 和反射分量$h_{e,\,m,\,\ell}^{\mathrm{refl}}[n]$，如图2所示。相应的频域传递函数为

$$\begin{aligned} H_{e,m,\ell}^{\mathrm{dir}}(\omega) &= \mathrm{FFT}\{h_{e,m,\ell}^{\mathrm{dir}}[n]\}, \\ H_{e,m,\ell}^{\mathrm{refl}}(\omega) &= \mathrm{FFT}\{h_{e,m,\ell}^{\mathrm{refl}}[n]\}. \end{aligned} \tag{18}$$

随后，对直达路径应用扬声器频率响应、解析活塞指向性和刚性球头相关传递函数，并对反射路径应用扬声器频率响应，从而得到用于BSANN训练的物理信息ATF：

$$\begin{aligned} \tilde{H}_{e,m,\ell}(\omega) = {} & H_{e,m,\ell}^{\mathrm{dir}}(\omega)\, A_\ell(\omega)\, D_\ell(\omega,\theta_{e,m,\ell})\, H_{\mathrm{HRTF},e,m}(\omega) \\ & + H_{e,m,\ell}^{\mathrm{refl}}(\omega)\, A_\ell(\omega), \end{aligned} \tag{19}$$

其中， $A_\ell(\omega)$ 是在消声环境下测得的扬声器频率响应， $D_\ell(\omega,\theta_{e,\,m,\,\ell})$ 是在角度$\theta_{e,\,m,\,\ell}$ 处计算得到的解析活塞指向性，其中 ℓ为扬声器， m 为耳朵 e的控制点；$H_{\mathrm{HRTF},\,e,\,m}(\omega)$ 则是上文定义的归一化刚性球体 HRTF。直达项捕捉指向性效应和耳部邻近滤波，而反射项则贡献无方向性后期能量。针对每只耳朵 e，对所有控制点 $m = 1, \ldots, M$和所有扬声器 ℓ 计算式（19），即可得到用于 BSANN损失函数的分耳ATF子矩阵$\mathbf{H}_{\mathrm{BL}}$、$\tilde{\mathbf{H}}_{\mathrm{BR}}$、$\tilde{\mathbf{H}}_{\mathrm{DL}}$和$\tilde{\mathbf{H}}_{\mathrm{DR}}$。

C. 采用教师保护的个人声区目标的主动 XTC

上述逐耳 BSANN 首先仅使用式（12）中的损失函数针对个人声区进行训练。为方便起见，我们定义

$$\begin{aligned}\mathcal{L}_{\mathrm{BZ}} &\equiv \mathcal{L}_1, \quad \mathcal{L}_{\mathrm{DZ}} \equiv \mathcal{L}_2,\\ \mathcal{L}_{\mathrm{gain}} &\equiv \mathcal{L}_3, \quad \mathcal{L}_{\mathrm{compact}} \equiv \mathcal{L}_4.\end{aligned} \tag{20}$$

PSZ 预训练阶段（图1中的步骤1）最小化

$$\mathcal{L}_{\mathrm{PSZ}} = \alpha\mathcal{L}_{\mathrm{BZ}} + (1-\alpha)\mathcal{L}_{\mathrm{DZ}} + \beta\mathcal{L}_{\mathrm{gain}} + \gamma\mathcal{L}_{\mathrm{compact}}, \tag{21}$$

从而得到一个基线 BSANN，它在不显式考虑主动 XTC 的情况下优化亮区和暗区性能。

为了实现带有 XTC 的基于扬声器的双耳渲染，我们引入一个 XTC 目标函数，在保留预期耳部响应的同时惩罚耳间泄漏。对于每个频率分箱 ω_n，令$\mathbf{P}(\omega_n) \in \mathbb{C}^{M\times 2\times L}$ 表示仅针对两个亮区耳朵的声学系统，每只耳朵有 M 个控制点，且有 L 个扬声器。令$\mathbf{W}(\omega_n) \in \mathbb{C}^{L\times 2}$ 表示 BSANN 预测并应用于这两个目标耳朵的立体声滤波器。由此得到的有效耳部传递矩阵为

$$\mathbf{T}_{\mathrm{eff}}(\omega_n) = \mathbf{P}(\omega_n)\,\mathbf{W}(\omega_n) \in \mathbb{C}^{M\times 2\times 2}, \tag{22}$$

其对角元素（R_{LL}， R_{RR}）对应预期的亮区耳部响应，而非对角元素（R_{LR}， R_{RL}）则表示耳间串音。

1) 非对角串音损失：对于每个频率分箱 ω_n，相对耳间泄漏定义为

$$r(\omega_n) = \frac{1}{2}\left[\frac{|R_{RL}(\omega_n)|^2}{|R_{LL}(\omega_n)|^2} + \frac{|R_{LR}(\omega_n)|^2}{|R_{RR}(\omega_n)|^2}\right]. \tag{23}$$

为避免低能量频率区域主导损失，我们采用对数压缩和基于能量的加权。定义对角能量

$$E(\omega_n) = \frac{1}{2}\Big(|R_{LL}(\omega_n)|^2 + |R_{RR}(\omega_n)|^2\Big), \tag{24}$$

在实现中，该能量被视为停止梯度权重。非对角 XTC损失计算为归一化加权平均：

$$\mathcal{L}_{\mathrm{off}} = \frac{1}{NM}\sum_{n=1}^{N}\frac{E(\omega_n)\,\log\big(1+r(\omega_n)\big)}{\sum_{n'=1}^{N}E(\omega_{n'})}. \tag{25}$$

该项会惩罚串音，同时抑制意图耳朵能量极低的频率分箱所产生的影响。

2) 对角匹配损失：为防止预期的耳部响应在主动串音消除阶段发生退化，我们将其幅度与目标对角响应$\mathbf{T}^{\mathrm{target}}(\omega_n)$进行匹配。对于每个频率分箱 ω_n，令 $R_{LL}^{\mathrm{target}}(\omega_n)$ 和 $R_{RR}^{\mathrm{target}}(\omega_n)$ 表示相应的目标幅度。对角匹配损失定义为

$$\begin{aligned}\mathcal{L}_{\mathrm{diag}} = \frac{1}{NM}\sum_{n=1}^{N}\frac{1}{2}\Bigg[&\left(\frac{|R_{LL}(\omega_n)|}{|R_{LL}^{\mathrm{target}}(\omega_n)|}-1\right)^2\\ &+\left(\frac{|R_{RR}(\omega_n)|}{|R_{RR}^{\mathrm{target}}(\omega_n)|}-1\right)^2\Bigg].\end{aligned} \tag{26}$$

该项通过在逐频率基准上惩罚偏离目标幅度的情况，保留所需的对角耳部响应。

3) 通过声学系统条件性进行正则化：为避免主动串音消除过程中出现不稳定的滤波器解，我们引入自适应努力正则项，其权重取决于局部声学系统格拉姆矩阵 [34] 的条件性。对于每个频率分箱 ω_n 和每个控制点 $m \in \{1,\dots,M\}$，令 $\mathbf{P}_m(\omega_n) \in \mathbb{C}^{2\times L}$ 表示该控制点处的逐耳声学系统。相应的格拉姆矩阵为

$$\mathbf{G}_m(\omega_n) = \mathbf{P}_m(\omega_n)^{\mathrm{H}}\mathbf{P}_m(\omega_n) \in \mathbb{C}^{L\times L}. \tag{27}$$

令 $\lambda_{\max,\,m}(\omega_n)$ 和 $\lambda_{\min,\,m}(\omega_n)$ 分别表示$\mathbf{G}_m(\omega_n)$ 的最大和最小特征值，并定义条件性比率

$$\kappa_m(\omega_n) = \frac{\lambda_{\max,m}(\omega_n)}{\lambda_{\min,m}(\omega_n)}. \tag{28}$$

根据上述实现，条件性相关权重为

$$\beta_m(\omega_n) = \beta_0\,\mathrm{ReLU}\left(\frac{\kappa_m(\omega_n)-\kappa_{\min}}{\kappa_{\min}}\right)\frac{\mathrm{tr}\big(\mathbf{G}_m(\omega_n)\big)}{L}, \tag{29}$$

其中， β_0 控制基础强度， $\kappa_{\min}$ 决定激活条件性惩罚项的阈值。随后，正则化损失计算如下：

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\beta_m(\omega_n)\,\big\|\mathbf{W}(\omega_n)\big\|_F^2, \tag{30}$$

该损失会在系统条件较差的频率–控制点组合处，更强烈地惩罚较大的滤波器幅值。

4) 最终串音消除目标函数：完整的串音消除目标函数结合了非对角串音惩罚项、对角响应匹配项以及条件性相关正则项：

$$\mathcal{L}_{\mathrm{XTC}} = \lambda_{\mathrm{off}}\,\mathcal{L}_{\mathrm{off}} + \lambda_{\mathrm{diag}}\,\mathcal{L}_{\mathrm{diag}} + \lambda_{\mathrm{reg}}\,\mathcal{L}_{\mathrm{reg}}. \tag{31}$$

5) 教师保护式自适应：直接优化（31）可能会大幅扰动个人化静区预训练解，并破坏亮区与暗区的能量平衡。为防止隔离性能下降，我们引入第二个训练阶段（图1中的步骤2），在该阶段中，BSANN基于个人化静区预训练滤波器，使用复合目标函数进行自适应：

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{xtc}}\,\mathcal{L}_{\mathrm{XTC}} + w_{\mathrm{BZ}}\,\mathcal{L}_{\mathrm{BZ}} + w_{\mathrm{DZ}}\,\mathcal{L}_{\mathrm{DZ}}\\ + \beta\,\mathcal{L}_{\mathrm{gain}} + \gamma\,\mathcal{L}_{\mathrm{compact}} + \eta\,\mathcal{L}_{\mathrm{teach}}.\end{aligned} \tag{32}$$

其中， λ_{xtc} 控制串音消除目标函数的相对重要性，而 w_{BZ} 和 w_{DZ} 是较小的保护权重，用于在主动串音消除过程中保留原有的亮区和暗区特性。

(32) 中的最后一项是教师锚定损失，用于使更新后的滤波器保持接近预训练的 PSZ 解。令 $\mathbf{F}_{\mathrm{cur}}(\omega_n)$ 和 $\mathbf{F}_{\mathrm{teach}}(\omega_n)$ 分别表示当前 BSANN 和预训练 BSANN 在频率 ω_n处的堆叠频域滤波器（涵盖所有扬声器和节目通道）：

$$\begin{aligned}\mathbf{F}_{\mathrm{cur}}(\omega_n) &= \mathrm{vec}\big(\mathbf{W}_{\mathrm{cur}}(\omega_n)\big),\\ \mathbf{F}_{\mathrm{teach}}(\omega_n) &= \mathrm{vec}\big(\mathbf{W}_{\mathrm{teach}}(\omega_n)\big).\end{aligned} \tag{33}$$

教师锚定损失定义如下：

$$\mathcal{L}_{\text{teach}} = \frac{1}{N}\sum_{n=1}^{N} \left\|\mathbf{F}_{\text{cur}}(\omega_n) - \mathbf{F}_{\text{teach}}(\omega_n)\right\|_2^2, \qquad (34)$$

这可以抑制相对于预训练 PSZ 滤波器的大幅偏离，同时仍允许网络沿着能够改善 $\mathcal{L}_{\text{XTC}}$的方向进行调整。

在实际应用中， w_{BZ} 和 w_{DZ} 的取值均小于λ_{xtc}，因此亮区项和暗区项仅作为较弱的正则化项，在主动串音消除阶段保留原始 PSZ 声场。教师锚定项 $\mathcal{L}_{\text{teach}}$ 为滤波器形状提供了额外的全局约束，从而在串音消除阶段实现稳定训练。如第 IV 节所示，最终采用两阶段流程：先通过 $\mathcal{L}_{\text{PSZ}}$ 进行 PSZ 预训练，再通过 $\mathcal{L}_{\text{total}}$进行主动串音消除，从而得到兼具高 IZI、IPI 和强串音消除效果的滤波器。

III. 实验设置与评估

本节介绍用于评估所提出的 BSANN 框架的实验设置，包括扬声器配置、听者几何布局、训练数据集、模型参数和评估指标。实验旨在量化真实声学条件下的 IZI、IPI 和串音消除效果。

A. 扬声器布局与听者几何布局

图 3 展示了用于评估的物理设置。由 24 个扬声器组成的阵列安装在一块刚性挡板上，位于听者高度，并排列成上下两排：下排为 8 个低音扬声器（100–2000 赫兹），上排为 16 个高音扬声器（2–20 千赫兹）。驱动单元沿水平弧线均匀间隔排列，朝向听者。阵列几何结构和驱动单元间距与 BSANN 训练期间所采用的配置一致。

两个头躯干模拟器（B&K HATS）对称地放置在阵列中心线两侧，分别位于中心左侧或右侧 0.5 米处，距阵列前方 1.0 米。两个 HATS 均与扬声器处于相同高度，双耳轴垂直于阵列，并相对于阵列中点居中。

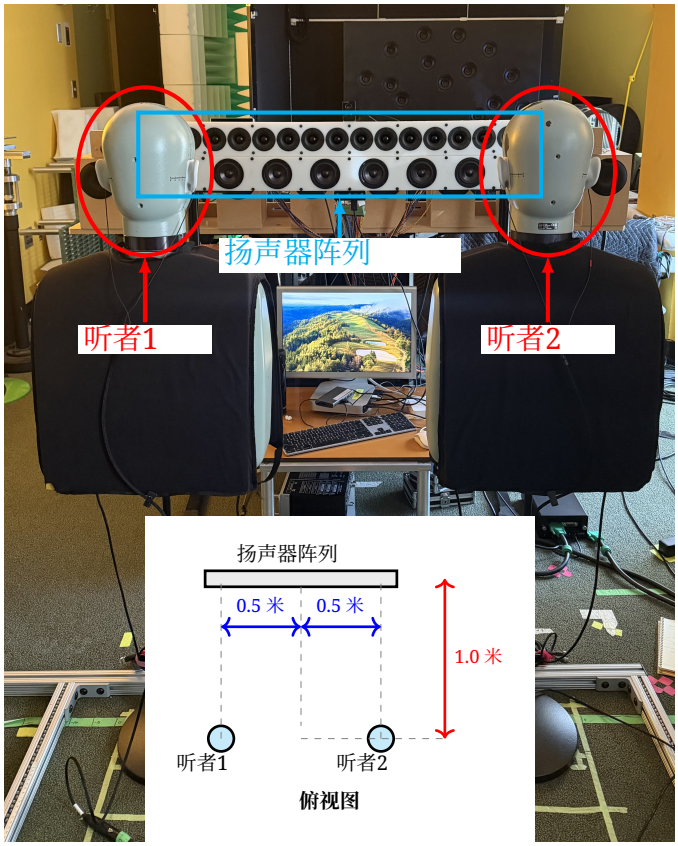


图3. 物理实验装置。扬声器系统由一个安装在听者高度的24单元扬声器阵列组成，下排有8个低音扬声器，上排有16个高音扬声器。两个头躯干模拟器（B&K HATS）对称地放置在阵列中心两侧，分别位于扬声器阵列左侧或右侧0.5 m处、前方1.0 m处，如俯视图所示。

B. 训练数据集与流程

训练数据集使用模拟房间脉冲响应（RIRs）生成，采用了原始 SANN-PSZ 工作 [23] 中的鞋盒形房间流程。对于每个训练样本，随机抽取房间尺寸、混响时间、阵列位置和听者区域中心，并使用 gpuRIR [35]模拟从所有扬声器到控制点的房间脉冲响应。除随机化策略外

用于 [23], 当前数据集纳入了听者几何布局的变化：刚性球头部的半径在合理范围内均匀采样；耳朵位置从球面偏移一小段随机距离，以考虑耳廓深度；耳旁控制点则通过在每只耳朵周围的圆形区域内均匀采样获得

，并剔除位于刚性球边界内部的点。

随后，按照˜第II-B节，利用模拟房间脉冲响应构建物理信息声学传递函数（声学传递函数）。每个声学传递函数 $H_{e,\ m,\ \ell}(\omega)$ 都由模拟的直达声和反射声房间响应、测得的无混响扬声器响应 $A_\ell(\omega)$、解析活塞指向性$D_\ell(\omega,\ \theta_{e,\ m,\ \ell})$以及刚性球头相关传递函数 $H_{\text{HRTF},\ e,\ m}(\omega)$组合而成。这些物理信息传递函数构成了完整的训练数据集，用于PSZ 预训练及后续的主动串音消除阶段。

利用这些物理信息传递函数，BSANN 分两个阶段进行训练。第一阶段使用式（21）中定义的损失函数对网络进行预训练，以用于个人声区，超参数为 $\alpha = 0.5$、$\beta = 0.5$ 和 $\gamma = 0.5$。

在第二阶段，根据式（32）中定义的、以 XTC损失为主导的复合目标函数对模型进行了优化。通过广泛的参数搜索来选择超参数，旨在最大化隔离度和串音消除性能。由于发现最终性能对超参数的具体取值仅有较弱的敏感性，因此我们没有对搜索过程进行系统化，也未详细报告。本研究采用的取值为$\lambda_{\text{xtc}} = 0.14$、 $\lambda_{\text{off}} = 1.5$、 $\lambda_{\text{diag}} = 1.0$、$\lambda_{\text{reg}} = 1.0$、 $\beta_0 = 10^{-4}$,

$\kappa_{\min} = 10^3$和 $\varepsilon = 10^{-8}$。亮区和暗区的保护器权重设为 $w_{\mathrm{BZ}} = w_{\mathrm{DZ}} = 0.075$。增益权重和紧致度权重均保持原始取值，分别为 $\beta = 0.5$ 和 $\gamma = 0.5$。教师锚定系数设为 $\eta = 1.0$。

C. 评估指标

我们采用三种客观指标进行评估：区间隔离度（IZI）、节目间隔离度（IPI）和˜串音消除（XTC） [23], [29], [36]。令 $\mathbf{H}_1(\omega) = [\mathbf{H}_{1,\mathrm{L}}(\omega), \mathbf{H}_{1,\mathrm{R}}(\omega)]^{\mathsf{T}}$ 表示听者1的逐耳传递函数子矩阵，同理，令 $\tilde{}\mathbf{H}_2(\omega) = [\mathbf{H}_{2,\mathrm{L}}(\omega), \mathbf{H}_{2,\mathrm{R}}(\omega)]^{\mathsf{T}}$表示听者2对应的子矩阵。令$\tilde{}\mathbf{g}_1(\omega) = [\mathbf{g}_{1,\mathrm{L}}(\omega), \mathbf{g}_{1,\mathrm{R}}(\omega)]^{\mathsf{T}}$ 表示节目1的BSANN生成的立体声滤波器，$\tilde{}\mathbf{g}_2(\omega) = [\mathbf{g}_{2,\mathrm{L}}(\omega), \mathbf{g}_{2,\mathrm{R}}(\omega)]^{\mathsf{T}}$则表示节目2的滤波器。

IZI用于评估与两位不同听者相关的亮区和暗区之间的声学分离度。IZI定义为 [23], [36]

$$\begin{aligned} \mathrm{IZI}_1(\omega) &= 10\log_{10}\frac{\|\tilde{\mathbf{H}}_1(\omega)\tilde{\mathbf{g}}_1(\omega)\|_2^2}{\|\tilde{\mathbf{H}}_2(\omega)\tilde{\mathbf{g}}_1(\omega)\|_2^2}, \\ \mathrm{IZI}_2(\omega) &= 10\log_{10}\frac{\|\tilde{\mathbf{H}}_2(\omega)\tilde{\mathbf{g}}_2(\omega)\|_2^2}{\|\tilde{\mathbf{H}}_1(\omega)\tilde{\mathbf{g}}_2(\omega)\|_2^2}. \end{aligned} \tag{35}$$

同样，IPI 用于评估渲染至同一听者的两个独立节目通道之间的干扰。IPI 定义为 [23], [36]

$$\begin{aligned} \mathrm{IPI}_1(\omega) &= 10\log_{10}\frac{\|\tilde{\mathbf{H}}_1(\omega)\tilde{\mathbf{g}}_1(\omega)\|_2^2}{\|\tilde{\mathbf{H}}_1(\omega)\tilde{\mathbf{g}}_2(\omega)\|_2^2}, \\ \mathrm{IPI}_2(\omega) &= 10\log_{10}\frac{\|\tilde{\mathbf{H}}_2(\omega)\tilde{\mathbf{g}}_2(\omega)\|_2^2}{\|\tilde{\mathbf{H}}_2(\omega)\tilde{\mathbf{g}}_1(\omega)\|_2^2}. \end{aligned} \tag{36}$$

XTC 用于评估对活动听者双耳之间对侧泄漏的抑制效果。对于两种可能的节目分配，令$\mathbf{T}_{\mathrm{eff},1}(\omega_n)$ 和 $\mathbf{T}_{\mathrm{eff},2}(\omega_n)$ 分别表示节目1和节目2的有效耳传递矩阵。对角线元素（$R_{LL,k}$,$R_{RR,k}$），（$k \in \{1,2\}$）表示预期的同侧响应，非对角线元素（$R_{LR,k}$,$R_{RL,k}$）表示残余串音。两种 XTC 测量指标定义为 [29]

$$\begin{aligned} \mathrm{XTC}_1(\omega) &= 10\log_{10}\frac{|R_{LL,1}(\omega)|^2 + |R_{RR,1}(\omega)|^2}{|R_{LR,1}(\omega)|^2 + |R_{RL,1}(\omega)|^2}, \\ \mathrm{XTC}_2(\omega) &= 10\log_{10}\frac{|R_{LL,2}(\omega)|^2 + |R_{RR,2}(\omega)|^2}{|R_{LR,2}(\omega)|^2 + |R_{RL,2}(\omega)|^2}. \end{aligned} \tag{37}$$

D. 测试条件

实验旨在评估立体声建模、ATF 构造中的声学真实性以及主动串音消除对整体 PSZ 和双耳性能的影响。实验考虑了三组条件：

- **SANN 与 BSANN 对比：**单声道 SANN-PSZ 模型与使用理想点声源声学传递函数训练、且不含物理信息组件或主动 XTC 阶段的 BSANN-PSZ 模型之间的比较，

表I 单声道 SANN-PSZ 模型与使用理想点声源声学传递函数训练、且不含物理信息组件或主动 XTC 阶段的 BSANN-PSZ 模型的对数频率加权均值 IZI、IPI 和 XTC。下标 1 和 2 分别表示听者1和听者2。 Δ 列表示指标差值（BSANN 减去 SANN）。

指标	SANN	BSANN	Δ
IZI_1 [分贝]	9.99	9.54	−0.45
IZI_2 [分贝]	5.11	9.18	+4.07
IPI_1 [分贝]	9.18	8.98	−0.20
IPI_2 [分贝]	5.92	9.74	+3.82
XTC_1 [分贝]	0.00	4.97	+4.97
XTC_2 [分贝]	0.01	4.38	+4.37

以便量化单独归因于逐耳控制的增益。

- **模拟点声源声学传递函数与物理信息传递函数：**比较分别使用理想点声源声学传递函数和物理信息传递函数训练的 BSANN-PSZ模型；后者纳入了测得的扬声器响应、解析活塞指向性和刚性球头相关传递函数。两个模型均未使用主动 XTC 阶段进行训练，以便量化物理信息声学建模带来的益处。

- **使用与未使用主动 XTC：**比较未使用主动 XTC 阶段训练的 BSANN-PSZ模型与使用主动 XTC 阶段训练的 BSANN-PSZ模型。两个模型均使用相同的物理信息传递函数进行训练，以便量化主动 XTC 阶段对整体性能的贡献。

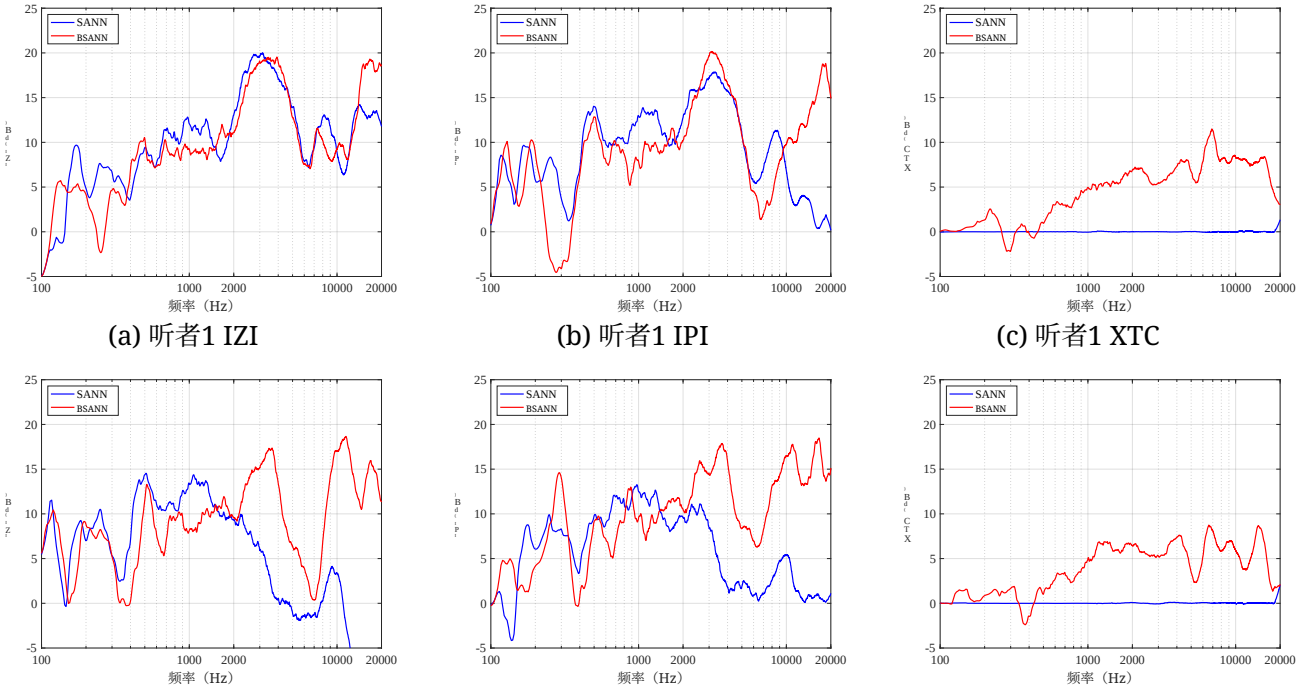
所有条件均在静态头部姿态下进行评估。评估时，将宽带扫频正弦信号通过各模型的滤波器处理后，经扬声器阵列播放。使用微型耳内麦克风（Theoretica Applied Physics 的 BACCH-BM Pro Mk III）记录听者1和听者2（均由 B&K HATS 表示）的耳部信号；麦克风放置在每位听者左右耳道入口处。随后使用测得的耳部信号计算 IZI、IPI 和 XTC 指标。本研究在边界不对称的听音室中进行，该环境更能代表真实的聆听空间。

IV. 实验结果

本节报告了在第III节所述评估条件下测得的 BSANN-PSZ 模型性能。结果根据第III-C节定义的指标，以 IZI、IPI 和 XTC 表示。除非另有说明，所有结果均为100–20,000赫兹范围内按对数频率加权的平均值。

A. SANN-PSZ 与 BSANN-PSZ 的比较

SANN 和 BSANN 的 IZI、IPI 和 XTC 对数频率加权均值汇总于表I，相应曲线见图4。这些结果表明



(a) 听者1 IZI (b) 听者1 IPI (c) 听者1 XTC

(d) 听者2 IZI (e) 听者2 IPI (f) 听者2 XTC

图4。单声道SANN-PSZ模型（蓝色）与使用理想点声源声学传递函数训练、且不含物理信息组件或主动串音消除的BSANN-PSZ模型（红色）之间的比较。每个面板分别显示听者1和听者2的对数频率加权 IZI、IPI 或 XTC 曲线。

单声道公式在两位听者之间造成了显著不均衡的性能：听者1实现了较为理想的隔离度，而听者2的 IZI 和 IPI 均低约 4–5 分贝。这种不均衡主要是由房间右侧的非对称反射环境造成的，强烈的早期反射降低了听者2位置的隔离度。

BSANN基本消除了这种不平衡。相较于SANN基线，BSANN使听者2的隔离度在IZI和IPI中分别提高了4.07分贝和3.82分贝，基本消除了两位听者之间的性能差距。相比之下，对于听者1，变化极小（IZI中为−0.45分贝，IPI中为 −0.20分贝），这表明，立体声逐耳公式在单声道模型已表现良好的情况下能够保持SANN的性能，同时显著改善处于不利位置的听者的性能。

立体声逐耳控制的优势在XTC指标中更加突出。对于SANN而言，两个程序的对数频率加权XTC实际上均为零；而BSANN则使听者1和听者2的XTC分别提高了4.97分贝和4.37分贝。这些在未采用主动 XTC 阶段的情况下获得的被动串扰消除增益，凸显了立体声逐耳公式固有的串音抑制能力。

总的来说，这些改进表明，BSANN-PSZ架构不仅能在非对称房间中平衡不同听者的性能，还能提供采用单声道SANN-PSZ架构无法实现的显著隔离度提升和串音增益。

表II
采用模拟点声源声学传递函数和物理信息传递函数训练的BSANN-PSZ模型的对数频率加权均值 IZI、IPI 和 XTC，均未使用主动串音消除。下标1和2分别表示听者1和听者2。 Δ 列表示指标差值（物理信息减去模拟值）。

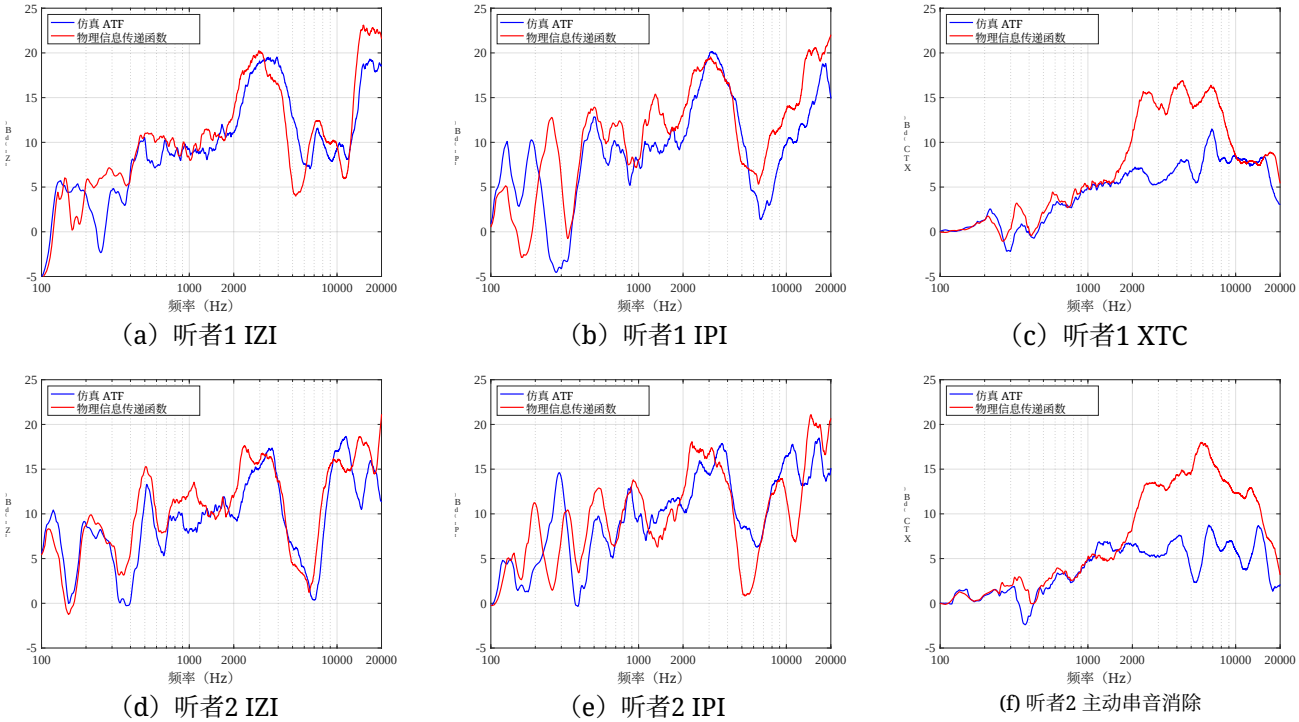
指标	仿真 ATF	物理信息 ATF	Δ
IZI_1 [分贝]	9.54	10.23	+0.69
IZI_2 [分贝]	9.18	10.43	+1.25
IPI_1 [分贝]	8.98	10.87	+1.89
IPI_2 [分贝]	9.74	9.79	+0.05
XTC_1 [分贝]	4.97	7.93	+2.96
XTC_2 [分贝]	4.38	8.19	+3.81

B. 物理信息声学传递函数的影响

图5和表II总结了物理信息声学建模对BSANN性能的影响。当使用模拟点声源声学传递函数进行训练时，BSANN已能为两位听者提供均衡的隔离度，其对数频率加权值分别为 $IZI_1 = 9.54$分贝、IZI_2 =9.18 分贝、$IPI_1 = 8.98$ 分贝和I $PI_2 = 9.74$ 分贝。

采用由实测扬声器响应、解析活塞指向性和刚性球头相关传递函数组成的物理信息ATF，可提升两位听者的隔离度。与仿真 ATF 相比，物理信息传递函数使 IZI_1 和 IZI_2 分别提高 0.69 分贝和 1.25 分贝，并使 IPI_1 和 IPI_2 分别提高 1.89 分贝和 0.05 分贝。这些增益表明，增强的声学真实性提高了 BSANN 实现的耳间隔离度。

最显著的差异出现在 XTC 指标上。使用仿真 ATF 时，对数频率加权串音消除水平



（a）听者1 IZI （b）听者1 IPI （c）听者1 XTC

（d）听者2 IZI （e）听者2 IPI (f) 听者2 主动串音消除

图5. 对比采用理想点声源声学传递函数（蓝色）训练的 BSANN-PSZ模型与采用物理信息传递函数（红色）训练的 BSANN-PSZ模型声学传递函数（红色）训练的两个模型均未进行主动串音消除。每个面板均显示听者1和听者2的对数频率加权 IZI、IPI 或 XTC 曲线

表III

采用和未采用主动 XTC 阶段训练的 BSANN-PSZ模型的对数频率加权 IZI、IPI 和串音消除水平。两个模型均使用相同的物理信息传递函数进行训练。下标 1 和 2 分别表示听者 1 和听者 2。 Δ 列表示指标差值（主动串音消除减去未进行串音消除）。

指标	无主动串音消除	主动串音消除	Δ
IZI_1 [分贝]	10.23	10.23	+0.00
IZI_2 [分贝]	10.43	10.03	−0.40
IPI_1 [分贝]	10.87	11.11	+0.24
IPI_2 [分贝]	9.79	9.16	−0.63
XTC_1 [分贝]	7.93	10.55	+2.62
XTC_2 [分贝]	8.19	11.13	+2.94

对于听者1和听者2，数值分别为4.97分贝和4.38分贝。采用物理信息传递函数后，这些数值分别提高至7.93分贝和8.19分贝，对应的增益为2.96分贝和3.81分贝。这些结果表明，纳入解析活塞指向性和依赖耳朵的散射效应，能够增强双耳间隔离度，并提高双耳（ITD、双耳声级差）线索传输的保真度。

总体而言，物理信息建模通过提高隔离性能，并借助更准确的扬声器频率响应、辐射模式和依赖耳朵的散射线索提供更强的被动串扰消除，从而提升了 BSANN 在真实声学条件下的性能。

C. 主动 *XTC* 阶段的影响

图6和表III展示了主动 XTC 阶段对 BSANN 性能的影响。在应用主动串音消除之前，使用物理信息传递函数进行训练的 BSANN 已实现较高的隔离度，其对数频率加权值分别为 $IZI_1 = 10.23$ 分贝、$IZI_2 = 10.43$ 分贝、$IPI_1 = 10.87$ 分贝和 $IPI_2 = 9.79$ 分贝。

应用主动串音消除可保持该 PSZ 性能。隔离度的变化很小：两位听者的 IZI 最多降低 0.40 分贝，IPI 最多降低 0.63 分贝，这表明 XTC 阶段不会损害 PSZ 预训练期间实现的亮区或暗区控制。

XTC 阶段的主要影响体现在 XTC 指标上。在未采用主动串音消除时，对数均值串扰消除水平分别为 7.93 分贝（听者1）和 8.19 分贝（听者2）。采用主动串音消除后，这些数值分别增至 10.55 分贝和 11.13 分贝，对应的增幅分别为 2.62 分贝和 2.94 分贝。这些增幅在图 6 的串扰消除曲线中清晰可见。

总体而言，主动 XTC 阶段增强了双耳 XTC，同时保持了通过 PSZ 预训练实现的隔离性能。

D. 总结

在所有评估条件下，所提出的 BSANN-PSZ 始终优于单声道SANN-个人声区基线，并体现出物理信息声学建模和主动串音消除带来的明显优势。与 SANN 相比，BSANN 在两位听者之间实现了更均衡的 IZI 和 IPI 表现，并显著提升了串音消除效果。纳入物理信息声学传递函数后，通过捕捉真实环境中的扬声器频率响应、指向性模式和依赖耳朵的散射效应，隔离度和双耳间隔离度均进一步提升。

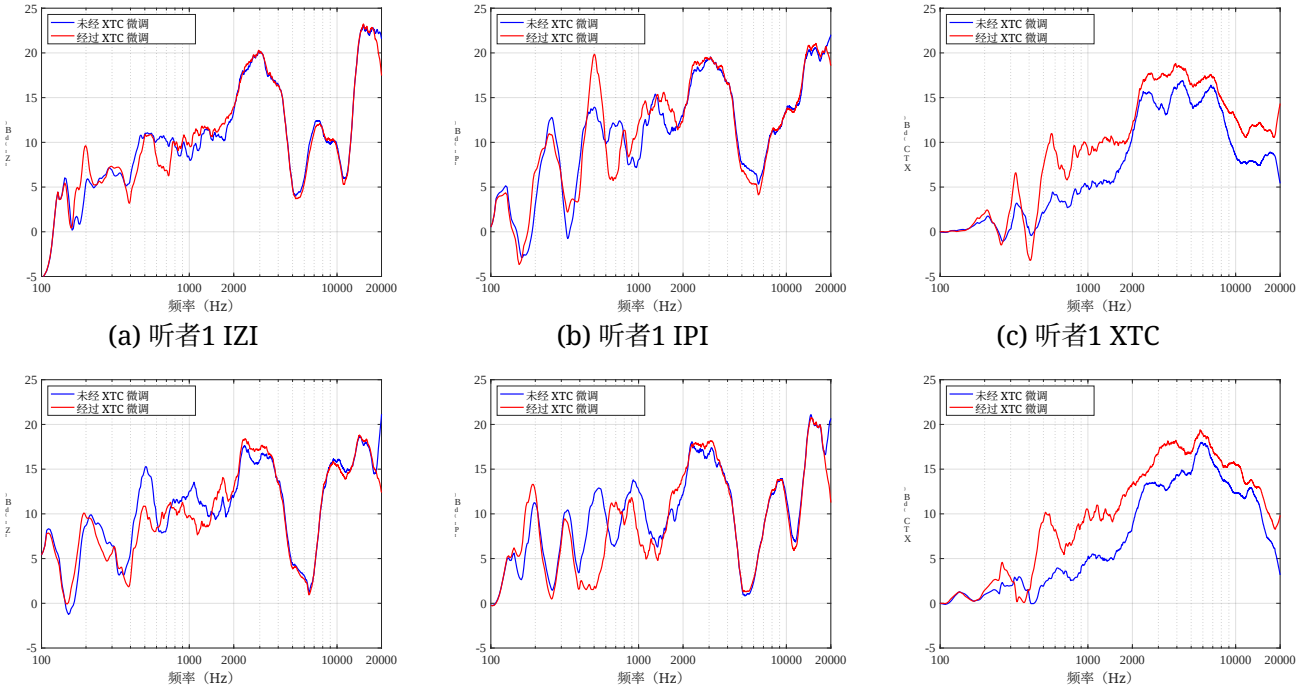


图 6. BSANN-PS Z 模型的比较：未采用主动 XTC 阶段训练（蓝色）与采用主动 XTC 阶段训练（红色），两者使用相同的物理信息传递函数。每个面板显示听者1和听者2的对数-频率加权 IZI、IPI 或 XTC 曲线。

最后，主动 XTC 阶段在保留预训练期间获得的 PSZ 性能的同时，将串音消除水平额外提高了 2.62–2.94 分贝。总之，这些结果客观地证明，逐耳立体声控制、物理信息建模和主动串音消除相辅相成，能够在真实声学条件下实现稳健的多听者立体声和双耳 PSZ 渲染。

第五节：讨论与结论

本研究提出了双耳空间自适应神经网络（BSANN），这是一个统一的框架，用于立体声和双耳 PSZ 渲染，可优化多个头部跟踪听者每只耳朵处的声场。通过将单声道 SANN-PSZ 公式扩展为逐耳架构，BSANN 可直接控制所有耳朵处的声压，而不只是声区中心处的声压。实验结果表明，这种方法能使非对称房间中不同听者的 IZI 和 IPI 更为均衡，同时提升 XTC，说明立体声逐耳控制所提供的额外自由度得到了有效利用，能够抵消房间引起的非对称性和耳间泄漏。

这一特性的关键在于使用物理信息传递函数，它融合了消声室测量的扬声器响应、解析活塞指向性和刚性球头相关传递函数。测得的扬声器响应将阵列实际的频谱着色融入训练数据，因此，网络学到的滤波器天然适配部署硬件，而非理想化声源。活塞模型引入了辐射的频率—角度依赖性，使高频下的主瓣更加锐利，并对旁瓣进行塑形，从而改善空间能量导向，减少溢出到暗区的能量。刚性球头相关传递函数还通过引入耳间声级差和时间差，以及头部周围依赖于方向的散射，将声场进一步分解为与耳朵相关的分量，从而增强耳间分离度，并在真实声学条件下稳定双耳线索。以物理信息传递函数取代点声源声学传递函数后，IZI、IPI 和被动串扰消除均有所改善，这与物理直觉相符：物理信息传递函数能够更准确地刻画扬声器如何在空间中辐射能量，以及每只耳朵如何接收这些能量，使网络能够学到更符合真实声学行为的滤波器。

主动 XTC 阶段通过明确塑造有效的耳间传递关系来优化这一基线，在保留预期同侧响应的同时抑制对侧泄漏。对非对角项进行惩罚可减少对侧泄漏，而对角匹配和教师锚定则可防止优化“过度校正”，从而避免使预期的同侧响应或底层 PSZ 声场失真。条件数感知正则项进一步抑制病态频率-位置区域中的过大滤波器增益，否则这些增益会放大噪声或房间不确定性。测得的 XTC 提升约为 2.6–2.9 分贝，而 IZI 和 IPI 仅有轻微变化，这与上述解释一致。

本研究重点介绍 BSANN-PSZ 框架并通过实验展示其优势，而非全面剖析各组成部分的贡献。具体而言，对物理信息

ATF 构造（消声室测量的扬声器频率响应、解析活塞指向性和刚性球体头相关传递函数）以及 XTC 目标函数中各项的详细消融分析留待未来研究，以便系统评估它们的相对重要性及可能的简化方案。

基于当前研究结果，仍有若干方向值得进一步探究。首先，尽管 BSANN 框架在当前的二维姿态表述中支持头部跟踪渲染，但一个自然的扩展方向是通过建模三维平移和三维旋转，将其扩展至完整的六自由度头部运动。其次，与仅在单一评估姿态下开展实验相比，在动态头部运动轨迹下进行实验将更全面地刻画渲染声场的时间稳定性和鲁棒性。第三，以从大型数据集中获取的个体化或群组头相关传递函数取代刚性球体模型，或可进一步提升双耳准确度，尤其是在较高频率下，此时耳廓和躯干效应占主导地位。第四，将 BSANN 生成的频域滤波器映射为高效的 IIR 实现，可降低嵌入式或移动设备部署中的计算成本和内存占用。最后，尽管非正式的主观聆听测试（本文未报告）表明，本文所述的大多数性能提升确实可以被听觉感知，但仍需开展系统且全面的主观评估，才能充分验证所提出的基于 BSANN 的 PSZ 框架。

致谢

作者感谢曾晓峰在开发本研究所用的刚性球头相关传递函数计算代码过程中给予的帮助。

参考文献

[1] W. Druyvesteyn 和 J. Garas，“个人化声音”，美国音频工程学会期刊，第45卷，第9期，第685–701页，1997年。[2] T. Betlehem、W. Zhang、M. Poletti 和 T. Abhayapala，“个人声区：为多位听者提供无需界面的音频”，*IEEE* 信号处理杂志，第32卷，第2期，第81–91页，2015年。[3] R. M. Jacobsen、K. F. Skov、S. Johansen、M. B. Skov 和 J. Kjeldskov，“ 与声区共处：在家庭环境中对动态声区开展的长期实地研究”，载于*2023*年计算机系统中的人为因素*CHI*大会论文集，2023年，第1–14页。[4] J. Cheer、S. J. Elliott 和 M. F. S. G´alvez，“汽车座舱个人音频系统的设计与实现”，美国音频工程学会期刊，第61卷，第6期，第412–424页，2013年。[5] L. Vindrola、M. Melon、J.-C. Chamard 和 B. Gazengel，“利用滤波-x 最小均方算法在汽车座舱中自适应个人声区”，美国声学学会期刊，第150卷，1793，2021 3，1779 第 页 – 。[6] K. F. Skov、P. A. Nielsen 和 J. Kjeldskov，“调适共享医院空间：医疗保健中的声区”，载于第*18*届国际*Audio Mostly*大会论文集，2023年，第63–70页。[7] F. M. Heuchel、D. Caviedes-Nozal、J. Brunskog、F. T. Agerkvist 和 E. Fernandez-Grande，“大规模户外声场控制”，美国声学学会期刊，第148卷，第4期，第2392–2402页，2020年。[8] J.-W. Choi 和 Y.-H. Kim，“利用多个声源生成带有照明区域的声学亮区”，美国声学学会期刊，第111卷，第4期，第1695–1700页，2002年。[9] M. G´alvez、S. Elliott 和 J. Cheer，“个人音频滤波器的时域优化”，*IEEE/ACM* 音频、语音与语言处理汇刊，第23卷，第11期，第1869–1878页，2015年。[10] M. Poletti，“二维多区域环绕声系统研究”，载于音频工程学会大会论文集，2008年。[11] J.-H. Chang 和 F. J acobsen，“采用圆形双层扬声器阵列的声场控制”，美国声学学会期刊，第131卷，第6期，第4518–4525页，2012年。[12] T. Abe、S. Koyama 和 H. Saruwatari，“多区域声场控制的振幅匹配”，IEEE/ACM 音频、语音与语言处理汇刊，第31卷，第656–669页，2023年。[13] T. Lee、J. Nielsen 和 M. Christensen，“采用可变跨度折衷滤波器的信号自适应与感知优化声区”，IEEE/ACM 音频、语音与语言处理汇刊，第28卷，第2412–2426页，2020年。[14] J. Brunnstr¨om、S. Koyama 和 M. Moonen，“采用核插值加权进行声区控制的可变跨度折衷滤波器”，载于 ICASSP 论文集，2022年。[15] S. Elliott、J. Cheer、J.-W. Choi 和 Y.-H. Kim，“个人音频系统的鲁棒性与正则化”，IEEE 音频、语音与语言处理汇刊，第20卷，第7期，第2123–2133页，2012年。[16] P. Coleman、P. J. Jackson、M. Olik、M. Møller 和 J. Pedersen，“使用圆形扬声器阵列的声区方法的声学对比度、平面度与鲁棒性”，美国声学学会期刊，第135卷，1940，2014 4，1929 第 页 – 。[17] Q. Zhu、P. Coleman、M. Wu 和 J. Yang，“通过声学建模减少现场测量的鲁棒声学对比度控制”，美国音频工程学会期刊，第65卷，第6期，第460–473页，2017年。[18] M. B. Møller、J. K. Nielsen、E. Fernandez-Grande 和 S. K. Olesen，“传递函数噪声对房间声区控制的影响”，IEEE/ACM 音频、语音与语言处理汇刊，第27卷，第9期，第1405–1418页，2019年。[19] Y. Qiao 和 E. Y. Choueiri，“采用头部跟踪的双耳房间脉冲响应神经建模与插值”，载于音频工程学会第155届大会（AES 155），纽约州纽约市，2023年。[20] M. G´alvez、D. Menzies 和 F. Fazi，“线性扬声器阵列的动态音频重放”，美国音频工程学会期刊，第67卷，第4期，第190–200页，2019年。[21] G. Pepe、L. Gabrielli、S. Squartini、C. Tripodi 和 N. Strozzi，“个人声区的数字滤波器设计：一种神经网络方法”，载于国际联合神经网络大会（IJCNN）论文集，2022年，第1–8页。[22] R. Alessandri，“一种用于多区域声场合成的基于深度学习的方法”，硕士论文，米兰理工大学，2021年。[23] Y. Qiao 和 E. Y. Choueiri，“Sann-psz：用于头部跟踪个人声区的空间自适应神经网络”，IEEE/ACM 音频、语音与语言处理汇刊，第33卷，第2735–2748页，2025年。[24] P. M. Morse 和 K. U. Ingard，《理论声学》。普林斯顿大学出版社，1968年。[25] P. R. Stepanishen，“无限平面挡板上活塞的瞬态辐射”，美国声学学会期刊，第49卷，第5期，第1629–1638页，1971年。[26] G. F. Kuhn，“方位平面上的双耳时间差模型”，美国声学学会期刊，第62卷，第1期，第157–167页，1977年。[27] R. O. Duda 和 W. L. Martens，“球形头部模型响应的距离依赖性”，美国声学学会期刊，第104卷，第5期，第3048–3058页，1998年。[28] X. Ma、C. Hohnerlein 和 J. Ahrens，“采用线性扬声器阵列的听者位置自适应超指向性交叉串音消除：概念与感知验证”，美国音频工程学会期刊，第67卷，第11期，第871–881页，2019年。[29] E. Choueiri，“通过扬声器播放双耳音频”，载于《沉浸式声音：双耳与多声道音频的艺术与科学》，A. Roginska 和 P. Geluso 编。纽约州纽约市：泰勒与弗朗西斯出版社，2018年，第6章，第124–179页。[30] T. Kabzinski 和 P. Jax，“一种采用耳部麦克风的自适应交叉串音消除系统”，载于 AES第147届大会论文集，2019年。[31] J. Lindfors、J. Liski 和 V. V¨alim¨aki，“移动听者的扬声器均衡”，美国音频工程学会期刊，第70卷，第9期，第722–730页，2022年。[32] M. Tancik、P. P. Srinivasan 等，“傅里叶特征使网络能够学习低维域中的高频函数”，载于 NeurIPS，2020年。[33] I. Goodfellow、Y. Bengio 和 A. Courville，《深度学习》。麻省理工学院出版社，2016年。[34] G. H. Golub 和 C. F. Van Loan，《矩阵计算》。约翰·霍普金斯大学出版社，2013年。[35] D. D´ıaz-Guerra、A. Miguel 和 J. R. Beltran，“gpurir：一种用于房间脉冲响应仿真并支持 GPU 加速的 Python 库”，《多媒体工具与应用》，第80卷，5671，2021 4，5653 第 页 – 。[36] Y. Qiao、L. Guadagnin 和 E. Choueiri，“个人声区重放系统的隔离性能指标”，JASA Express Letters，第2卷，第10期，文章编号104601，2022年。