物理声学建模对神经网络个性化声区渲染的影响剖析：一项消融研究

HaoJiang[1] Edgar Choueiri[1]1 3D 音频与应用声学实验室，普林斯顿大学，美国
新泽西州普林斯顿 hj3737@princeton.edu, choueiri@princeton.edu

arXiv:2603.02508v1 [eess.AS] 3 Mar 2026

摘要

基于深度学习的个人声区（PSZ）依赖模拟声学传递函数（ATF）进行训练，但理想化的点声源模型存在较大的仿真到现实差距。尽管物理信息组件有助于提升泛化能力，但各组件的独立贡献尚不明确。本文采用 Bin-aural Spatial Audio Neural Network（BSANN），针对一种以头部姿态为条件的双耳 PSZ 渲染器开展了受控消融研究。我们逐步为模拟声学传递函数加入三种组件：（i）特定扬声器消声室测量的频率响应（FR）；（ii）解析圆形活塞指向性（DIR）；以及（iii）刚性球头部相关传递函数（RS-HRTF）。我们使用两个仿真人头进行现场测量，评估四种配置。性能指标包括 100–20,000 赫兹范围内的区间隔离度（IZI）、节目间干扰（IPI）和串音消除（XTC）。结果表明，FR 提供频谱校准，适度提升 XTC，并减少听者间 IPI 不平衡。DIR 带来最稳定的声区分离增益（IZI/IPI 平均提高 10.05 分贝）。RS-HRTF 对双耳分离的提升最为显著，使 XTC 提高 +2.38/+2.89 分贝（平均值从 4.51 提高至 7.91 分贝），主要体现在 2 千赫兹以上；同时也会造成轻微的、因听者而异的 IZI/IPI 变化。这些发现有助于在预算有限的情况下，确定构建训练用声学传递函数时测量与模型的优先顺序。

1 引言

个人声区（PSZs）旨在通过扬声器阵列塑造声场，在共享的声学空间中为多位听众提供彼此独立的音频节目，无需佩戴耳机 [1, 2]。这一能力可在家庭和汽车环境等共享空间场景中实现私密或个性化的聆听体验 [3, 4]。传统的个人声区重现通常被表述为一个基于测得的声学传递函数（ATFs）的优化问题，典型方法包括声学对比度控制（ACC） [5, 6] 和声压匹配（PM） [7]。各种变体和扩展方法持续改进目标函数的表述与感知权衡，例如时域优化和可变跨度折衷滤波 以及可变跨度折衷滤波 [8–11]。

PSZ 系统面临的一个持续挑战，是对用于设计的实测“被控对象”声学传递函数与运行期间所遇到的传递函数之间失配的鲁棒性。此类失配可能源于听者移动、扬声器重新摆放、房间条件变化以及测量噪声，从而可能大幅降低隔离度并增加干扰 [12, 13]。因此，一系列鲁棒化策略已得到研究，包括正则化以及借助声学建模的降测量方案 [12, 14]。针对动态场景，自适应方法与位置相关方法也已被探索，以适应移动中的听者与时变条件 [4, 15, 16]。

近年来，基于深度学习的方法已成为 PSZ 渲染与滤波器设计的一种互补范式。状态条件神经渲染器，例如空间音频神经网络（SANN） [17], ，将滤波器设计摊销到学习过程中，从而降低运行时计算负担，并在经过足够多样的声学条件训练后可能提升鲁棒性。然而，此类数据驱动方法的有效性取决于用于构建训练与评估数据的传递函数的质量与覆盖范围。尤其对于头部跟踪与双耳渲染而言，获取密集的、依赖听者的实测 ATF 数据集往往耗时过长，因而促使人们依赖模拟 ATF。然而，基于理想化点声源假设的仿真往往难以良好迁移到实际部署，从而导致显著的仿真到现实差距。

这种差距主要源于理想化点声源模型所忽略的物理行为，包括设备特定的频谱着色、频率相关的扬声器指向性，以及决定双耳线索的头部相关双耳传播效应。为了以切实可行的方式缩小这一差距，我们近期在双耳空间音频神经网络（BSANN） [18] 的研究中采用了基于物理信息的 ATF，通过可迁移的组成部分丰富仿真；这些组成部分可以离线测量或建模，例如在消声环境中测得的扬声器响应、解析活塞指向性，以及刚性球头部相关传递函数（RS-HRTF） [19–21]。这些选择颇具吸引力，因为它们无需

进行全面的室内测量，同时还能捕捉主导仿真到现实失配的关键非理想扬声器效应和双耳效应。然而，这些组成部分通常被整体引入，因此每一层建模带来的增量收益并不明确；例如，串音消除与声区分离方面的改进，究竟主要由换能器特性驱动，还是由头部散射效应驱动。

本文通过一项受控的分阶段消融研究来探讨这一问题，量化逐步加入符合物理规律的声学组件对神经声区渲染的影响。我们采用 BSANN 架构 [18] 作为代表性框架，并使用逐步丰富方案：从基线点声源仿真出发，依次纳入：(i) 特定扬声器在消声环境下测得的频率响应（FR），(ii) 解析活塞指向性建模（DIR） [20, 22],以及 (iii) 刚性球头部相关传递函数（ RS-HRTF ）建模 [21, 23]。由此得到四种配置：基线、基线+FR、基线+FR+DIR，以及基线+FR+DIR+RS-HRTF。在所有配置中，我们固定神经网络架构、优化设置、数据划分和训练方案，仅改变 ATF 生成流程。尽管这种分阶段设计无法将组件效应与加入顺序完全分离，但它反映了符合物理规律的传播链，并能够在这一顺序下，将性能变化归因于逐步纳入各个建模层。性能采用区间隔离度（IZI）、[24], 中定义的节目间干扰（IPI）和串音消除（XTC）进行量化，并报告宽带指标及代表性频段上的指标。

本研究的贡献主要有三点：

1. 提出一种受控的分阶段消融方案，量化在神经 PSZ 训练中逐步加入具有物理依据的声学组件后，对模拟 ATF 的影响；

2. 定量分析频谱校正（FR）、指向性建模（DIR）和刚性球头相关传递函数建模（RS-HRTF）对声区分离和串音消除的增量效益，并分析宽带及不同频率范围内的表现；

3. 针对测量和实现预算有限的情况，提供构建训练用声学传递函数时优先选择物理测量和模型的实用建议。

2 方法

本节详细介绍用于分离各个物理信息声学组件贡献的实验框架。我们从整体层面概述固定的 BSANN 渲染器（完整架构细节将在 [18] 中介绍），描述具有物理信息的 ATF 构建流程和累积消融方案，并定义评估度量指标。

2.1 神经 PSZ 渲染框架

我们采用双耳空间音频神经网络（BSANN） [18] 作为所有实验的固定渲染框架。网络根据听者瞬时的头部姿态进行条件化，预测一组控制滤波器，以实现动态双耳重放。令 L 表示扬声器数量， K 表示听者（节目）数量，其中每个节目 k 由立体声源通道$s_{k,\ c}[n]$ 组成，适用于 $c \in \{L, R\}$。在本研究中，每位听者 k 都对应一个专属节目 k。对于给定状态，渲染器会生成一个滤波器组

$$\mathbf{w} = \{w_{\ell,k,c}[n]\}_{\substack{\ell=1\ldots L\\ k=1\ldots K\\ c\in\{L,R\}}}. \tag{1}$$

单元 ℓ 的扬声器驱动信号由所有卷积后的声源通道叠加而成：

$$x_\ell[n] = \sum_{k=1}^{K} \sum_{c\in\{L,R\}} (s_{k,c} * w_{\ell,k,c})[n], \tag{2}$$

其中 $*$ 表示卷积。

BSANN采用与 [18], 相同的目标函数和训练流程进行端到端训练，使用的复合损失包括亮区准确度和暗区抑制项，以及增益限制和时域紧凑性正则项。尤为重要的是，为确保消融研究公平，各实验配置中的渲染器架构、损失权重、优化超参数以及训练集/验证集划分均保持不变。唯一变化的是训练期间用于模拟声学环境的声学传递函数集合。

2.2 物理信息引导的声学传递函数建模

为合成训练用声学传递函数，我们遵循 [18], 第 II-B 节中的物理信息引导的声学传递函数构建流程，其中将模拟的点源房间脉冲响应分解为直达分量和反射分量。对于每个听者–耳朵–控制点–扬声器组合（k, e, m, ℓ）（听者 k、耳朵 $e \in \{L, R\}$、控制点m、扬声器 ℓ），使用 gpuRIR [25] 生成点源房间脉冲响应，并将其分离为直达分量 $h^{\mathrm{dir}}_{k,\ e,\ m,\ \ell}[n]$ 和反射分量 $h^{\mathrm{refl}}_{k\,e,\ ,\ m,\ \ell}[n]$。其频域传递函数为

$$\begin{aligned} H^{\mathrm{dir}}_{k,e,m,\ell}(\omega) &= \mathrm{FFT}\{h^{\mathrm{dir}}_{k,e,m,\ell}[n]\},\\ H^{\mathrm{refl}}_{k,e,m,\ell}(\omega) &= \mathrm{FFT}\{h^{\mathrm{refl}}_{k,e,m,\ell}[n]\}. \end{aligned} \tag{3}$$

随后，我们使用三个物理信息组件逐步增强直达分量和反射分量：(i) 特定扬声器的消声测量频率响应（FR），(ii) 解析活塞指向性（DIR），以及 (iii) 刚性球头相关传递函数（RS-HRTF）。

(i) 特定扬声器的消声室扬声器频率响应（FR）。为捕捉与扬声器相关的频谱染色，我们将每个扬声器单元经消声测量的复频率响应 $A_\ell(\omega)$ 纳入其中。测得的响应同时应用于直达路径和反射路径。

(ii) 解析活塞指向性（指向性）。为对频率相关的指向性进行建模，我们采用刚性（无限）挡板中的解析圆形活塞模型 [20, 22]。令 $\theta_{k,\ e,\ m,\ \ell}$ 表示扬声器声轴与从扬声器ℓ 指向听者-k 控制点 m （位于耳朵 e周围）的方向之间的离轴角。远场幅度指向性因子定义为

$$D_\ell(\omega,\theta_{k,e,m,\ell}) = \frac{2J_1(k_0 a_\ell \sin\theta_{k,e,m,\ell})}{k_0 a_\ell \sin\theta_{k,e,m,\ell}},\ k_0 = \frac{\omega}{c}, \quad (4)$$

其中， $J_1(\cdot)$ 是第一类一阶贝塞尔函数， a_ℓ 是有效活塞半径， c 是声速。根据连续性，轴上取值定义为 $D_\ell(\omega, 0) = 1$。该指向性模型仅应用于直达分量，而反射分量则贡献无方向性晚期能量。

(iii) 刚性球头部相关传递函数（刚性球头部相关传递函数）。为捕捉头部引起的散射和遮蔽效应，耳朵附近的控制点采用归一化刚性球散射公式 [21, 23]进行建模。我们将头部建模为半径为 R_k的刚性球体。令 $\mathbf{r}_{k,\ell}$ 表示从听者-k 头部中心指向扬声器ℓ的向量，令 $\mathbf{r}_{k,e,m}$ 表示控制点 m（位于耳朵 e 周围）相对于同一头部中心的位置。我们定义$r_{k,\ell} = \|\mathbf{r}_{k,\ell}\|$, $r_{k,e,m} = \|\mathbf{r}_{k,e,m}\|$, 以及角间隔 $\gamma_{k,e,m,\ell} = \angle(\mathbf{r}_{k,\ell}, \mathbf{r}_{k,e,m})$。

对于声波数 $k_0 = \omega/c$，散射核定义为

$$\begin{aligned} S_n(k_0) = {} & j_n(k_0 r_<)\, h_n^{(1)}(k_0 r_>) \\ & - \alpha_n(k_0 R_k)\, h_n^{(1)}(k_0 r_{k,\ell})\, h_n^{(1)}(k_0 r_{k,e,m}), \end{aligned} \quad (5)$$

其中， $r_< = \min(r_{k,\ell}, r_{k,e,m})$， $r_> = \max(r_{k,\ell}, r_{k,e,m})$， $j_n(\cdot)$ 和 $h_n^{(1)}(\cdot)$ 分别为球贝塞尔函数和汉克尔函数，$\alpha_n(\cdot)$ 表示刚性球散射系数。

扬声器 ℓ 到耳朵 e 周围控制点 m （听者 k）的归一化刚性球HRTF表示为

$$\begin{aligned} H^{\mathrm{HRTF}}_{k,e,m,\ell}(\omega) = {} & \frac{\mathrm{i}k_0}{G_{\mathrm{ff},k,e,m,\ell}(\omega)} \\ & \cdot \sum_{n=0}^{N_{\max}} (2n+1)\, S_n(k_0)\, P_n(\cos\gamma_{k,e,m,\ell}), \end{aligned} \quad (6)$$

其中， $P_n(\cdot)$ 是勒让德多项式， $N_{\max}$ 是截断阶数，而

$$G_{\mathrm{ff},k,e,m,\ell}(\omega) = \frac{e^{-\mathrm{i}k_0\|\mathbf{r}_{k,\ell} - \mathbf{r}_{k,e,m}\|}}{\|\mathbf{r}_{k,\ell} - \mathbf{r}_{k,e,m}\|} \quad (7)$$

表1：累积消融配置。每个阶段都在前一阶段的基础上，通过启用一个额外的物理信息组件来构建。

ID	频率响应（A_ℓ）	指向性（D_ℓ）	刚性球头部相关传递函数（H^{HRTF}）
C0	No	No	No
C1	Yes	No	No
C2	Yes	Yes	No
C3	Yes	Yes	Yes

是用于归一化的自由场格林函数。该项仅应用于直达路径分量。

2.3 累积消融方案

我们采用一种累积消融策略，其设置遵循符合物理原理的传播链：声源换能器响应（频率响应） $\rightarrow$ 声源辐射（指向性） $\rightarrow$ 头部散射（刚性球头部相关传递函数）。我们定义了四种配置（C0–C3），汇总于表1。以基线点声源仿真（C0）为起点，我们依次启用消声室测量的频率响应、解析指向性和刚性球头部散射项（刚性球头部相关传递函数）。

对于每种配置，训练用声学传递函数通过增强第2.2节所述的点声源直达与反射分量合成。为简洁起见，以下公式中省略频率依赖关系（ω）和路径索引 $\{k, e, m, \ell\}$。令 θ和Ω分别表示特定的扬声器离轴角和头部入射角。累积配置定义如下:

$$\begin{aligned} H^{\mathrm{C0}} &= H^{\mathrm{dir}} + H^{\mathrm{refl}}, \\ H^{\mathrm{C1}} &= A\Big[H^{\mathrm{dir}} + H^{\mathrm{refl}}\Big], \\ H^{\mathrm{C2}} &= A\Big[D(\theta)H^{\mathrm{dir}} + H^{\mathrm{refl}}\Big], \\ H^{\mathrm{C3}} &= A\Big[D(\theta)H^{\mathrm{HRTF}}(\Omega)H^{\mathrm{dir}} + H^{\mathrm{refl}}\Big]. \end{aligned} \quad (8)$$

等价地，将 $A = 1$, $D(\theta) = 1$，并令 $H^{\mathrm{HRTF}}(\Omega) = 1$，即可从 C3 恢复基线 C0。与第 2.2 节一致，频谱响应 A 作用于直达路径和反射路径，而空间项 D和 H^{HRTF} 仅调制直达分量。

2.4 度量指标

我们使用区间隔离度（IZI）、 [24], 中定义的节目间干扰（IPI）以及串音消除（XTC）来量化声区分离与双耳性能。度量指标按频率定义，基于在左耳和右耳测得的重放声压。在本研究中，我们考虑双听者场景，即 $K = 2$。

IZI与IPI。按照 [24], , 我们分别单独激活每个节目。令 $P_{(j)\,k,e}(\omega)$表示仅以单位能量播放节目 j 时，听者 k 的耳朵 e 处的复重放声压。对于每个频率 ω，我们定义

$$E_k^{\mathrm{tar}}(\omega) = \sum_{e\in\{L,R\}} \left|P_{k,e}^{(k)}(\omega)\right|^2, \tag{9}$$

$$E_k^{\mathrm{int}}(\omega) = \sum_{e\in\{L,R\}} \sum_{j\neq k} \left|P_{k,e}^{(j)}(\omega)\right|^2, \tag{10}$$

$$E_k^{\mathrm{leak}}(\omega) = \sum_{i\neq k} \sum_{e\in\{L,R\}} \left|P_{i,e}^{(k)}(\omega)\right|^2. \tag{11}$$

频率相关的声区隔离指数和 IPI 为

$$\mathrm{IZI}_k(\omega) = 10\log_{10}\left(\frac{E_k^{\mathrm{tar}}(\omega)}{E_k^{\mathrm{leak}}(\omega)+\epsilon}\right). \tag{12}$$

$$\mathrm{IPI}_k(\omega) = 10\log_{10}\left(\frac{E_k^{\mathrm{tar}}(\omega)}{E_k^{\mathrm{int}}(\omega)+\epsilon}\right), \tag{13}$$

其中 ϵ 是一个较小的正则化常数，用于确保数值稳定性。

串音消除（XTC）。对于每位听者 k，我们估计一个 2×2 端到端传递矩阵 $\mathbf{T}_k(\omega)$，用于将预期的立体声输入通道映射到重放的耳部信号。频率相关的串音消除定义为

$$\mathrm{XTC}_k(\omega) = 10\log_{10}\left(\frac{|T_{k,L,L}(\omega)|^2+|T_{k,R,R}(\omega)|^2}{|T_{k,L,R}(\omega)|^2+|T_{k,R,L}(\omega)|^2+\epsilon}\right). \tag{14}$$

宽带结果通过对感兴趣频率范围内以分贝表示的度量指标进行对数均值平均得到。

3 实验设置与评估

为量化纳入物理信息组件的影响，我们通过真实环境实验验证了四种累积配置（C0–C3）。本节详细介绍物理测试平台和基于测量的评估流程。所有评估均在听音室中使用两个静止的 Br¨uel & Kjær 头躯干模拟器（HATS）进行。尤为重要的是，为确保比较公平，所有实验中的 BSANN 模型架构、训练超参数和评估流程均保持不变；唯一的变量是用于生成训练数据的合成策略，如第 2.3 节所述。

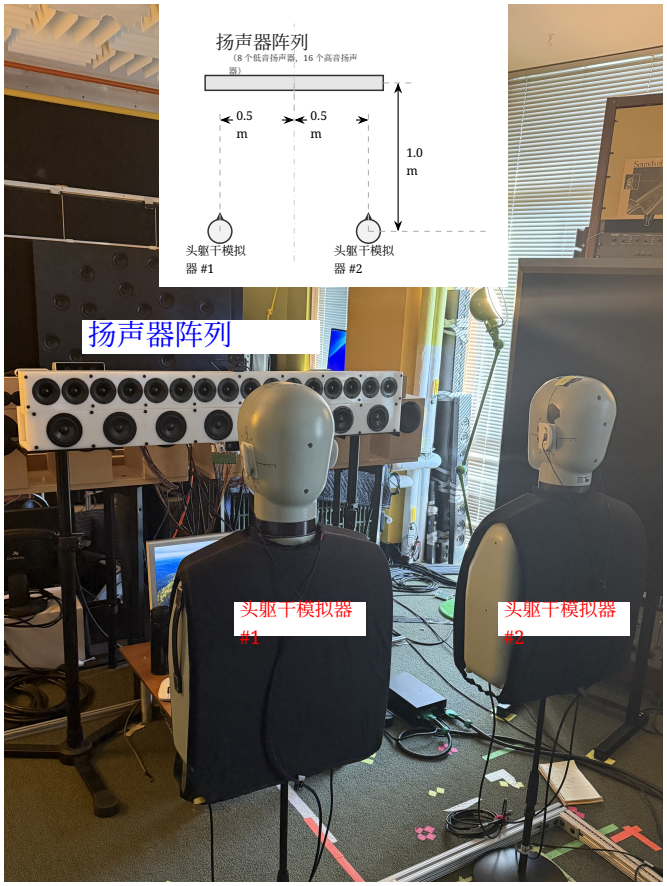


图1：实验测试平台的几何配置。回放系统采用安装在刚性障板上的24单元扬声器阵列，分为两排：8个低音扬声器（100–2000赫兹）和16个高音扬声器（2–20千赫兹）。两个头躯干模拟器对称放置，距阵列1.0 m，相对中心线的横向偏移为±0.5m，朝向垂直于阵列中点。

3.1 物理设置

扬声器系统由安装在刚性障板上、位于听者高度的 $L = 24$ 单元线性阵列组成。该阵列包含上下两排：8 个低音扬声器，工作频率约为 100–2000 赫兹；以及 16 个高音扬声器，覆盖 2–20 千赫兹。两个头躯干模拟器关于阵列中心线对称放置，分别位于中心左侧和右侧 0.5 米处，距阵列 1.0 米。头躯干模拟器与扬声器垂直对齐，其耳间轴垂直于阵列中点。试验台的详细几何配置如图 1 所示。

3.2 训练条件

我们训练了四个 BSANN 模型，分别对应累积消融配置 C0–C3（表1）。对于每个

模型，训练用声学传递函数均采用相同的仿真框架合成，仅改变声学传递函数构建流程：基线点声源响应（C0）、加入特定扬声器实测的扬声器频率响应（C1）、在直达路径上加入解析活塞指向性（C2）、在直达路径上加入刚性球头相关传递函数（C3）。C0–C3 的网络架构选择、损失权重、优化设置和数据集划分均保持一致。

3.3 基于测量的评估方案

所有配置均在静态头部姿态下进行评估。对于每个测试用例，经过训练的 BSANN 会针对固定的听者几何构型生成特定的扬声器控制滤波器。为验证性能，使用这些预测滤波器处理宽带指数正弦扫频（ESS）激励信号，并通过扬声器阵列播放。随后，使用经过自由场均衡的 BACCH-BM Pro 入耳式麦克风采集双耳信号；麦克风放置在头躯干模拟器耳道入口处。利用录制的扫频信号计算第 2.4 节定义的 IZI、IPI 和串音消除度量指标。除非另有说明，结果均报告为 100–20,000 赫兹范围内按对数频率加权的平均值。

4 结果与分析

按照第 2.3 节的定义，对四种累积训练配置（C0–C3）在固定评估几何构型下的单一静态位置进行现场测量评估。性能指标按频率计算（第 2.4 节），并汇总为 100–20,000 赫兹带宽内的宽带对数均值（即对数间隔频率区间对应的分贝值的算术平均值）。

4.1 总体测量性能

图 2(a–c) 总结了每位听者的宽带对数均值度量指标，而图 3 展示了相应的频率相关特性。对测量数据的分析揭示了三个关键观察结果。

首先，结合头部散射（刚性球头部相关传递函数）可显著增强串音消除。配置 C3 为两位听者带来了最高的串音消除效果（XTC_1：7.95 分贝；XTC_2：7.87 分贝），使听者平均串音消除从基线（C0）中的 4.51 分贝提升至 C3 [图 2(c)] 中的 7.91 分贝。

这一趋势在图 3 的串音消除频谱中体现得十分明显：在中高频范围（约从 2 千赫兹开始），C3 与其他配置明显不同。相反，在 1 千赫兹以下，各配置的性能总体相当，表明在此设置中，低频双耳分离对是否包含头部散射不太敏感。

值得注意的是，与 C0 相比，C1 和 C2 仅使串音消除发生了小幅变化，而主要的性能提升出现在 C3 启用刚性球头部相关传递函数时。

其次，声区分离指标（IZI 和 IPI）主要由声源指向性建模决定。最优的听者平均 IZI 和 IPI 均由配置 C2（+频率响应+指向性）实现，两个度量指标均达到 10.05 分贝 [图 2(a,b)]。这一观察结果与图 3 一致：C2 在中频区域展现出持续的提升，这对应于能量导向以及泄漏/干扰抑制最为有效的带宽。

最后，图 2(a,b) 揭示了听者之间声区分离的不对称性。对于听者1，随着物理保真度提高，IZI 和 IPI 单调改善，并在 C3 时达到峰值（IZI_1：$8.98 \rightarrow 10.34$ 分贝；IPI_1:$8.33 \rightarrow 11.06$ 分贝）。相比之下，对于听者2，声区分离从 C0 到 C2 持续改善，C2 的整体表现最佳（IZI_2:$9.59 \rightarrow 10.06$ 分贝；IPI_2:$10.23 \rightarrow 10.01$ 分贝）；而物理保真度进一步提高至 C3 后，声区分离反而下降，尤其是 IPI。频率相关曲线进一步支持了这一非单调趋势：在特定子频带中，C3 响应低于 C2。

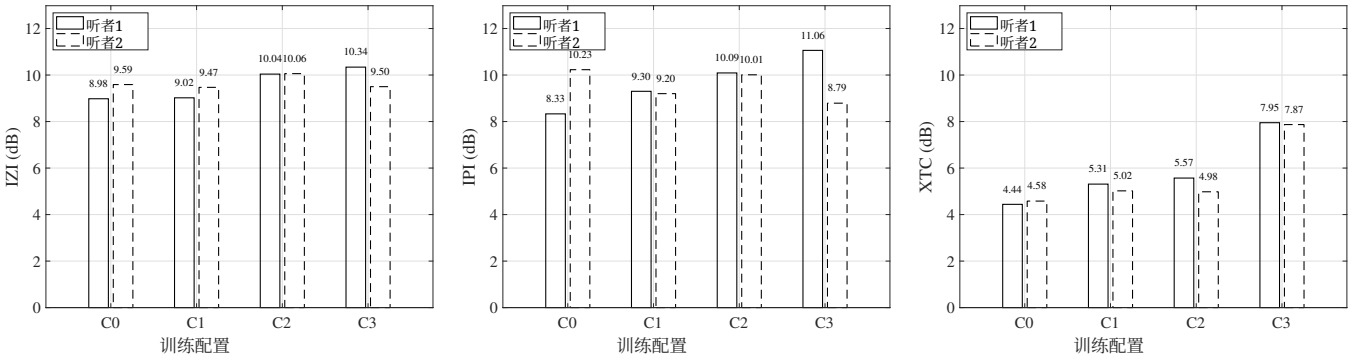
4.2 频率响应、指向性和刚性球头部相关传递函数的增量贡献

为了在累积协议下分离各物理分量的增量效应，图 2(d–f) 分别报告了听者1和听者2的宽带对数均值变化（100–20,000赫兹），对应于依次加入频率响应（C1−C0）、指向性（C2−C1）和刚性球头部相关传递函数（C3−C2）。

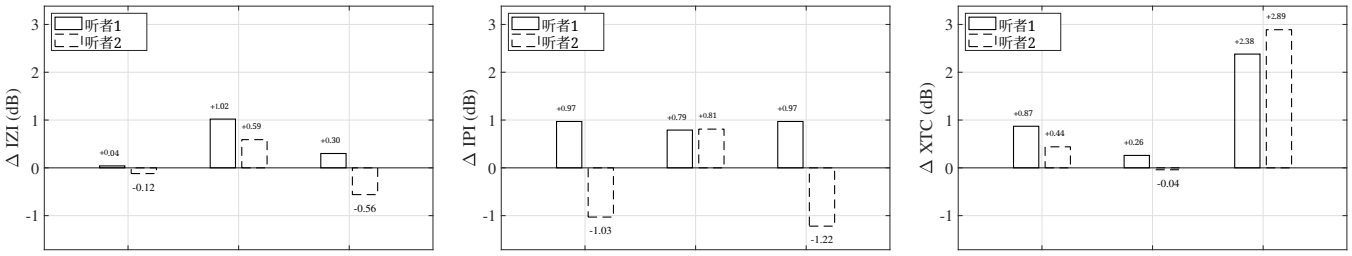
频率响应（C1 与 C0）。加入特定扬声器的消声室测量频率响应后，两位听者的串音消除均获得适度但一致的改善（Δ串音消除$_1$=+0.87分贝，Δ串音消除$_2$=+0.44 分贝；图 2(f)）。在图 3 中，这表现为相对 C0 在部分频段上的中等曲线偏移。对于声区分离指标，频率响应对 IZI 影响较小，但显著降低了听者间 IPI 的不对称性。尽管 C0 表现出较大差异（IPI_2 比 IPI_1 高出 1.90 分贝），加入频率响应后 IPI_1 增加 0.97 分贝、IPI_2 降低 1.03 分贝，从而使性能接近均衡（IPI_1=9.30 分贝，IPI_2=9.20 分贝） [图 2(b)]。

这种趋同表明，在固定目标函数下，纳入频率响应可降低扬声器引起的频谱失配，并在两位听者之间重新分配干扰抑制权衡。

指向性（C2 对比 C1）。引入解析活塞指向性后，两名听者的声区分离均获得最为一致的提升：IZI 提高



(a) 宽带 IZI (b) 宽带 IPI (c) 宽带 XTC



增量加法 增量加法 增量加法

（d）增量 ΔIZI (e) 增量 ΔIPI （f）增量 Δ串音消除

图 2：四种累积配置的宽带性能总结。上排：听者1和听者2在 C0–C3 配置下，100–20,000 赫兹范围内的对数均值指标值。下排：累积消融方案中的增量变化（频率响应：C1−C0；指向性：C2−C1；刚性球头部相关传递函数：C3−C2）。

了 $\Delta IZI_1 = +1.02$分贝和 $\Delta IZI_2 = +0.59$ 分贝，而 IPI 提高了 $\Delta IPI_1 = +0.79$ 分贝和 $\Delta IPI_2 = +0.81$ 分贝（图 2(d,e)）。在图 3 中，这对应于分离曲线的宽带抬升，与建模非全向辐射后更为真实的频率相关能量导向以及泄漏/干扰的降低相一致。

此阶段串音消除的变化较小，且因听者而异（Δ串音消除$_1 = +0.26$分贝， Δ串音消除$_2 = -0.04$ 分贝；图 2(f)），这表明指向性主要通过空间能量约束改善声区分离（IZI/IPI），而对该设置下的双耳通道解耦影响有限。

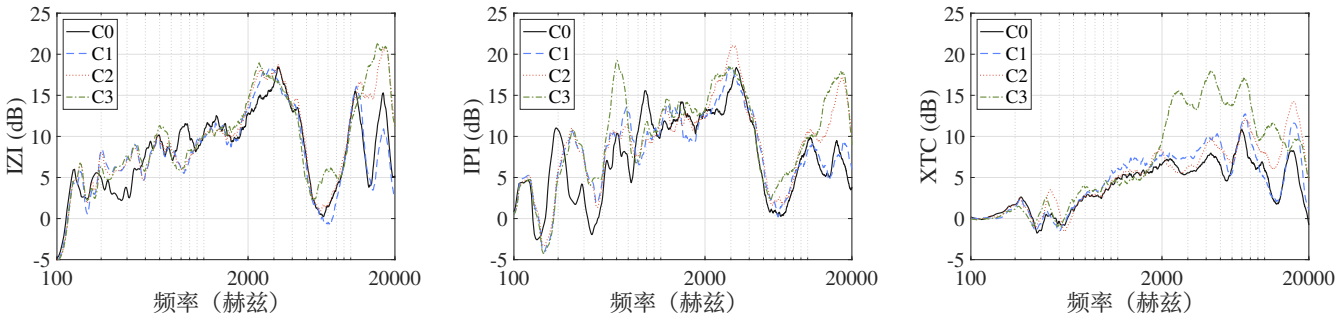
刚性球头部相关传递函数（C3 与 C2 对比）。加入刚性球头散射是双耳分离性能提升的主要因素，使串音消除提高了 Δ串音消除$_1 = +2.38$ 分贝和Δ串音消除$_2 = +2.89$分贝（图 2(f)）。图 3 显示，这种提升在中频区域及以上最为明显（约从 2 千赫兹开始），此处头部引起的双耳差异较为显著。与此同时，对声区分离指标的影响因听者而显著不同：对听者1而言，声区隔离和节目间干扰均有所改善（$\Delta IZI_1 = +0.30$ 分贝，$\Delta IPI_1 = +0.97$ 分贝）；而对听者2而言，两者均有所下降（$\Delta IZI_2 = -0.56$ 分贝， $\Delta IPI_2 = -1.22$分贝）[图 2(d,e)]。

串音消除的显著提升表明，纳入头部引起的散射和声影效应至关重要——尤其是在实测播放条件下实现中高频双耳分离。IZI 和 IPI 表现出的差异表明存在轻微的几何相关的权衡：对串音消除有益的局部双耳散射结构建模，可能同时改变能量导向和泄漏/干扰抑制效果，进而导致 IZI 和 IPI 随听者而异。

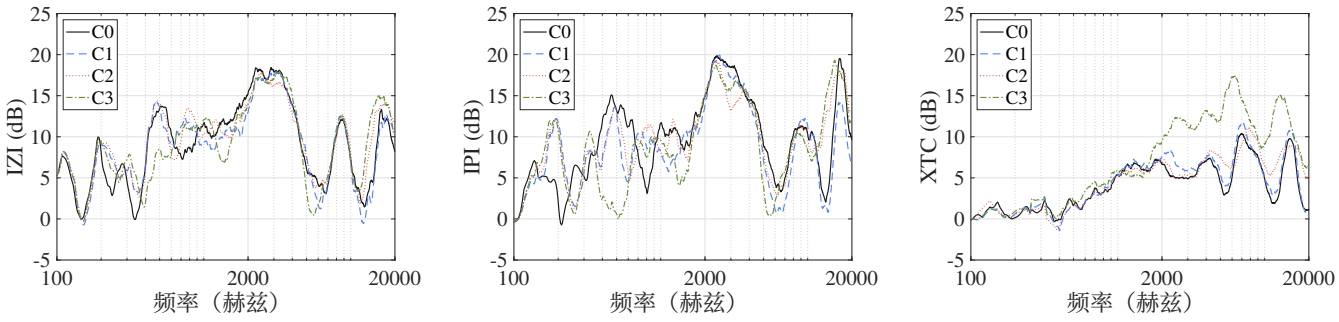
5 结论

本文提出了一项受控的累积消融研究，以量化ATF生成流程中基于物理信息的声学分量如何影响神经个人声区渲染。在BSANN框架 [18] 下，所有网络与训练设置保持固定，我们针对单一静态位置上两台头躯干模拟器的原位测量评估了四种训练配置（C0–C3）。性能采用频率相关的声区隔离指数、IPI和串音消除进行评估，宽带汇总值以100–20,000赫兹范围内的对数均值报告。

结果清晰地呈现了组件级归因。加入特定扬声器的实测频率响应（FR）后，两位听者的串音消除（XTC）均有小幅但一致的改善；更重要的是，IPI 中的听者间不平衡有所降低，表明



(a) 听者1：IZI　(b) 听者1：IPI　(c) 听者1：串音消除



（d）听者2：IZI　（e）听者2：IPI　（f）听者2：串音消除

图 3：在单一静态位置下评估四种累积训练配置的频率相关的声区隔离指数、IPI 和串音消除。图例对应于 C0（基线点声源仿真）、C1（+频率响应）、C2（+频率响应+指向性）、以及 C3（+频率响应+指向性+刚性球头部相关传递函数）。上排：听者1。下排：听者2。

准确的换能器频谱表征主要起到实用校准步骤的作用，即使峰值隔离的改善有限，也能减少频谱失配和听者间不平衡。引入解析活塞指向性（DIR）后，两位听者的声区分离（IZI 和 IPI）均获得最一致的提升，而对 XTC 的影响微乎其微。这表明，指向性主要改善空间能量约束，并抑制泄漏和干扰，而非实现双耳通道解耦。加入刚性球头散射（RS-HRTF）则在双耳分离方面带来主要提升，产生了幅度最大的 XTC 增益，尤其是在 2 千赫兹以上。与此同时，纳入 RS-HRTF 会导致 IZI/IPI 随听者而异，表明在固定渲染器和目标函数下，最大化 XTC 与维持声区分离之间存在轻微且取决于几何条件的权衡。

从实际应用角度来看，这些发现表明，在测量或实施资源有限的情况下，可采用以下优先级策略：

- 对于主要针对声区分离（IZI/IPI）的应用，对扬声器指向性进行建模是最具影响力的物理组件；
- 对于双耳重放质量和串音消除（XTC），引入头部散射模型至关重要；以及
- 实测扬声器频谱响应有助于减少仿真到现实的频谱失配，并平衡不同听者之间的性能。

未来的工作将把这一分析从单一静态位置扩展到动态和多位置评估，并将研究以来源于大型数据库的个体化或群组头相关传递函数替代刚性球头部相关传递函数，以进一步提升双耳精度，尤其是在耳廓和躯干效应占主导的较高频率下，所有这些工作均将在相同的累积消融框架下进行。

参考文献

[1] W. F. Druyvesteyn 和 J. Garas。《个人声音》。《音频工程学会期刊》，45(9):685–701，1997年9月。

[2]Terence Betlehem、Wen Zhang、Mark A. Poletti 和 Thushara D. Abhayapala。个人声区：为多位听众提供无界面音频。《IEEE信号处理杂志》，32(2):81– 91，2015年3月。doi: https://doi.org/10.1109/MSP. 2014.2360707。

[3] Jordan Cheer 和 Stephen J. Elliott。《汽车座舱中个人音频系统的设计与实现》。载于《声学会议论文集》，第19卷，页码055009，蒙特利尔，加拿大，2013年6月。doi：https://doi.org/10.1121/1.4798948。

[4]Lucas Vindrola、Manuel Melon、Jean-Christophe Chamard 和 Bruno Gazengel。《采用滤波 X 最小均方算法调整汽车座舱中的个人声区》。《美国声学学会期刊》，150(3)：1779–1793，2021年9月。doi：https://doi.org/10.1121/10.0005875。

[5] Joung-Woo Choi 和 Yang-Hann Kim。使用多个声源生成具有照明区域的声学亮区。《美国声学学会期刊》，111(4):1695–1700，2002年4月。doi: https://doi.org/ 10.1121/1.1456926。

[6] Philip Coleman、Philip J. B. Jackson、Marek Olik、Martin Møller、Martin Olsen 和 Jan Abildgaard Pedersen。使用圆形扬声器阵列的声区方法的声学对比度、平面性和鲁棒性。《美国声学学会期刊》，135(4):1929–1940，2014年4月。doi: https://doi.org/10.1121/1.4866442。

[7] Mark A. Poletti。二维多区域环绕声系统研究。发表于音频工程学会第125届大会，2008年10月。论文7551。

[8]Marcos F. Sim´on G´alvez、Stephen J. Elliott 和 Jordan Cheer。用于个人音频扬声器阵列的滤波器时域优化。IEEE/ACM 音频、语音与语言处理汇刊，23(11):1869–1878，2015年11月。doi: https: // doi.org/10.1109/TASLP.2015.2456428。

[9]Taewoong Lee、Jesper Kjær Nielsen 和 Mads Græsbøll 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，28:2412–2426，2020年7月。doi: https://doi.org/10.1109/TASLP.2020.3013397。

[10]Jesper Brunnstr¨om、Shoichi Koyama 和 Marc Moonen。采用核插值加权的声区控制可变跨度折衷滤波器。发表于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，第1071–1075页，新加坡，2022年5月。doi: https: //doi.org/10.1109/ICASSP43922.2022.9746550。

[11] Takumi Abe、Shoichi Koyama、Natsuki Ueno 和 Hiroshi Saruwatari。多区域声场控制中的幅度匹配。《IEEE/ACM 音频、语音与语言处理汇刊》，31:656–669，2023年1月。doi：https://doi.org/10.1109/TASLP.2022.3231715。

[12] Stephen J. Elliott、Jordan Cheer、Jung-Woo Choi 和 Youngtae 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》，20(7):2123–2133，2012年9月。doi：https://doi.org/10.1109/TASL.2012.2197613。

[13]Martin Bo Møller、Jesper Kjær Nielsen、Efren Fernandez-Grande 和 Søren Krarup Olesen。传递函数噪声对房间内声区控制的影响。IEEE/ACM 音频、语音与语言处理汇刊，27(9):1405–1418，2019年9月。doi：https://doi.org/10.1109/TASLP.2019.2921151。

[14]Qiaoxi Zhu、Philip Coleman、Ming Wu 和 Jun Yang。通过声学建模减少现场测量的稳健声学对比度控制。《音频工程学会期刊》，65(6):460–473，2017年6月。doi：https://doi.org/10.17743/jaes.2017.0016。

[15] Marcos F. Sim´on G´alvez、Dylan Menzies 和 Filippo Maria Fazi。使用线性扬声器阵列的动态音频重放。《音频工程学会期刊》，67(4):190–200，2019年4月。doi：https://doi.org/10.17743/jaes.2019.0007。

[16] Joel Lindfors、Juho Liski 和 Vesa V¨alim¨aki。面向移动听者的扬声器均衡。J. Au- dio Eng. Soc.，70(9):722–730，2022年9月。doi: https://doi.org/10.17743/jaes.2022.0020。

[17] Yue Qiao 和 Edgar Choueiri。SANN-PSZ：用于头部跟踪个人声区的空间自适应神经网络。IEEE Trans. Audio Speech Lang. Pro- cess.，33:2735–2748，2025年6月。doi: https://doi.org/ 10.1109/TASLPRO.2025.3581123。

[18]Hao Jiang 和 Edgar Choueiri。使用双耳空间自适应神经网络（BSANN）为个人声区进行立体声音频渲染。arXiv 预印本，2026年1月。URL https://arxiv.org/abs/2601.06621。

[19] Edgar Choueiri. 通过扬声器播放双耳音频。载于 Agnieszka Roginska 和 Paul Geluso 编，《沉浸式声音：双耳与多声道音频的艺术与科学》，第 5 章，第 124–179 页。Taylor and Francis，纽约，纽约州，2018年。

[20] Peter R. Stepanishen. 无限平面挡板中活塞的瞬态辐射。美国声学学会期刊，49(5):1629–1638，1971年5月。doi: https://doi.org/10. 1121/1.1912541.

[21] RichardO. Duda 和 William L. Martens. 球形头部模型响应的距离依赖性。美国声学学会期刊，104(5):3048–3058，1998年11月。doi: https://doi.org/10.1121/1.423886.

[22] Philip M. Morse 和 K. Uno Ingard。《理论声学》。普林斯顿大学出版社，新泽西州普林斯顿，1986年。

[23] George F. Kuhn。方位平面中耳间时间差的模型。美国声学学会期刊，62(1):157–167，1977年7月。doi: https://doi.org/10. 1121/1.381498。

[24] Yue Qiao、L´eo Guadagnin 和 Edgar Choueiri。个人声区重放系统的隔离性能指标。JASA Express Lett.，2(10): 104801，2022年10月。doi: https://doi.org/10.1121/ 10.0014604。

[25]David D´ıaz-Guerra、Antonio Miguel 和 Jose R. Beltran。gpuRIR：一款利用 GPU 加速进行房间脉冲响应仿真的 Python 库。《多媒体工具与应用》，80(4):5653–5671，2021年2月。doi：https://doi.org/10.1007/s11042-020-09905-3。