多通道房间脉冲响应的快速随机近似

Yi Luo*，Rongzhi Gu*

arXiv:2304.08052v1 [cs.SD] 17 Apr 2023

摘要——**现代基于神经网络的语音处理系统通常需要具备抵抗混响的能力，因此训练这类系统需要大量混响数据。如今，系统训练更倾向于采用在线仿真流程，因为这种流程无需预先生成并保存数据到硬盘，就能让模型使用无限量的数据样本进行训练。因此，房间脉冲响应仿真方法不仅需要生成更逼真的人工房间脉冲响应滤波器，还需要快速生成这些滤波器，以加速训练过程。现有的房间脉冲响应仿真工具已在广泛的语音处理任务和神经网络架构中证明了其有效性，但由于计算复杂度或所生成的房间脉冲响应滤波器的质量问题，它们在在线仿真流程中的应用仍存疑。本文提出 FRAM-RIR，这是一种对广泛使用的镜像源法（ISM）进行快速随机近似的方法，可高效生成逼真的多通道房间脉冲响应滤波器。FRAM-RIR 根据若干启发式假设，随机采样每个虚拟声源的位置和反射次数，从而绕过基于镜像源法的算法对声音传播路径的显式计算，同时仍能准确保留所有声源的到达方向信息。理想波束图和方向特征的可视化结果表明，与现有广泛使用的基于镜像源法的工具相比，FRAM-RIR 能生成更逼真的房间脉冲响应滤波器；此外，在多通道噪声语音分离和去混响任务上的实验结果显示，采用多种神经网络架构时，与其他房间脉冲响应仿真工具相比，使用 FRAM-RIR 训练的模型在真实房间脉冲响应上的表现也能达到同等或更优水平，同时显著加快训练过程。FRAM-RIR 的 Python 实现可在线获取**[1]。

关键词—数据增强、镜像源法、房间脉冲响应、语音处理

I. 引言

当说话者或其他声源处于室内环境中时，日常对话中会出现混响。为确保近期神经语音处理和建模系统在此类场景下的鲁棒性，通常会在系统训练中使用人工混响，以增加可用训练数据的数量，并扩大数据对不同房间条件的覆盖范围 [1]。因此，用于模拟人工混响的方法通常通过房间脉冲响应滤波器来表示，并且需要模仿真实混响的模式。此外，随着越来越多的现代系统选择实时数据仿真流程，这种流程允许无需预先生成并保存到硬盘，即可生成无限量的训练数据，并能够提升系统在广泛任务中的性能 [2]–[6], ，RIR滤波器的仿真速度在训练流程中也变得至关重要。

过去几年中，人们研究并提出了多种 RIR 仿真工具，以提高仿真质量并加快仿真过程 [7], [8]。基于物理建模的方法一直是主流，因为它们能够根据房间条件准确地模拟声音反射。最流行的方法之一是镜像声源法 *(ISM)* [9], ，该方法通过房间边界（例如地板和墙壁）对原始声源进行镜像来计算声传播路径。为了加快仿真过程并提高生成滤波器的质量，人们提出了基于扩散的方法，以绕过晚期混响声传播路径的显式计算 [10], [11]。基于 ISM 的算法和工具已广泛用于大多数神经网络语音处理系统的训练，并已证明在多种任务中有效，例如自动语音识别 (ASR) [12]–[14], 语音增强 [15], [16],语音分离 [17], [18],和语音去混响 [19],[20]。除了基于 ISM 的工具外，人们还探索了基于光线追踪的方法，以支持对更真实房间的建模，从而生成更好的 RIR 滤波器 [21], [22]。天鹅绒噪声和重缩放脉冲列等随机近似方法也被用于以更快的方式模拟人工混响 [8], [23]–[25]。随着生成模型的近期发展，神经网络也被用于优化仿真 RIR 滤波器，使其更好地匹配实录 RIR 滤波器的分布 [26]–[28]。

尽管上述方法已成功应用于各种任务和应用中，但在现代基于神经网络的语音处理系统训练中使用这些方法仍面临若干困难。基于镜像声源法的方法通常假设房间为空且呈矩形或平行六面体形状，而这一假设可能导致“扫掠回声效应”，使模型在使用此类脉冲响应训练后，在真实场景中的泛化能力受损 [29], [30]。为减轻扫掠回声效应，在各种统计假设下随机化虚拟声源位置已被证明有效 [30], [31], ，但仍需显式计算声传播路径。声传播路径的计算也可能耗时，且需要借助图形处理器等硬件进一步加速 [32], [33], ，这使得在分布式模型训练中需要快速进行数据仿真时，这些方法更难并行运行。基于射线追踪的方法计算开销可能很大，而且要仿真逼真的RIR滤波器，可能需要预先进行房间建模，这

*贡献相同。[1] https://github.com/tencent-ailab/FRA-RIR/tree/fram rir

仍不适用于即时数据生成。随机近似或噪声整形方法能够快速生成RIR滤波器，但这些滤波器的表现通常与真实RIR滤波器不同，而且这些方法通常难以推广到多通道场景，因为需要准确保留每个声源-麦克风对的到达时间差（TDOA）。基于神经网络的方法在神经网络的规模和复杂度适中的情况下可以实现快速运行，但对生成的RIR滤波器进行精细控制相对困难，例如确保生成的RIR滤波器准确保留所需的混响时间（T60）或每个声源的到达方向（DOA）。这些缺点使现有方法难以应用于神经网络模型的即时训练流程。

本文扩展了我们此前关于单通道 RIR 滤波器快速仿真的工作 [34]，并提出了多通道 RIR 的快速随机近似方法（***F***ast ***R***andom***A***pproximation of ***M***ulti-channel ***RIR***，***FRAM-RIR***）。这是一种简单的方法，能够仿真多通道 RIR 滤波器，并准确保留声源的到达方向信息，速度足以支持即时数据仿真。FRAM-RIR 遵循标准基于镜像源法的算法的问题定义，但通过若干启发式假设，根据每个虚拟声源的反射位置和数量进行随机采样，从而近似计算明确的声传播路径。所有虚拟声源的采样过程都可以并行进行，因此 FRAM-RIR 速度快，适合在中央处理器上并行处理。使用标准台式机级别的中央处理器，为 4 麦克风阵列生成 RIR 滤波器时，FRAM-RIR 可在单个中央处理器线程上于 140 毫秒内生成逼真的 RIR 滤波器，速度最高可比现有基于镜像源法的工具快 40 倍。此外，对真实 RIR 滤波器和不同仿真 RIR 滤波器卷积后的语音的理想波束图与方向特征进行可视化后发现，与其他工具相比，FRAM-RIR 能更好地模拟真实的空间模式。在多种神经网络架构上的多通道噪声语音分离和去混响任务实验结果表明，与其他 RIR 仿真工具相比，使用 FRAM-RIR 训练的模型在真实房间脉冲响应上的表现也能达到相当或更好的水平。

本文其余部分安排如下。第二节简要回顾基于镜像源法的房间脉冲响应仿真方法。第三节介绍所提出的 FRAM-RIR 方法。第四节介绍实验配置。第五节分析 FRAM-RIR 的行为，并展示采用各种神经网络架构进行分离和去混响任务的结果。第六节总结全文。

II. 镜像源法回顾

我们采用 [1] 中由镜像法生成的房间脉冲响应滤波器的定义：

$$h^m[n] = \frac{1}{D_{0,m}}\delta\left[n - \left\lceil \frac{D_{0,m}f_s}{c_0} \right\rceil\right] + \sum_{i=1}^{I} \frac{r^{g_{i,m}}}{D_{i,m}}\delta\left[n - \left\lceil \frac{D_{i,m}f_s}{c_0} \right\rceil\right] \tag{1}$$

其中，$m \in [1, \ldots, M]$ 表示接收器索引，I 表示虚拟声源总数，$D_{0,\,m}$ 表示直达声源到第 m 个接收器的距离，$D_{i,\,m}$ 表示第 i 个虚拟声像到第 m个接收器的距离，r 表示表面的反射系数，g_i 表示第i个声源到第 m 个接收器的反射次数，f_s 表示目标采样率，c_0 表示声速。图 1(a) 展示了房间内点声源 P_0 的声像示意图。我们采用相同的方法，通过 Eyring 经验公式 [1], [35] 估算反射系数：

$$r = \sqrt{1 - \left(1 - e^{-0.16R/T_{60}}\right)^2} \tag{2}$$

其中，R 表示房间体积与总表面积的比值，T_{60} 表示声音在房间中衰减 60 分贝所需的混响时间。

由于 $h^m[n]$ 是脉冲序列，其采样率会影响时间分辨率，因此采样率需要足够高，以确保即使使用间距较小的麦克风阵列，也能有效地对不同虚拟声源的到达时间差进行建模。因此，我们采用与 [1] 相同的策略：首先以很高的采样率 $r_h f_s$ 生成 $h^m[n]$，然后将其降采样至中间采样率 $r_l f_s$，其中 $1 < r_l < r_h$ 为两个重采样因子；接着应用截止频率为 80 赫兹的高通滤波器，以去除不需要的低频成分 [1], [9]。随后再次对滤波后的房间脉冲响应滤波器进行降采样，使其达到目标采样率 f_s，作为最终输出，与实际声源（其采样率同样为 f_s）进行卷积。在实践中，我们将 $r_h = \lfloor 10^6/f_s \rfloor$ 和 $r_l = \lfloor \sqrt{r_h} \rfloor$ 设定为相应数值，以平衡仿真速度和房间脉冲响应质量。 —

III. 多通道RIR的快速随机近似

在我们此前关于单通道RIR快速随机近似（FRA-RIR）的研究中，[34],我们提出了对公式 1 的计算进行的三项核心改进：

1) 对房间相关统计量 R 进行随机采样，而不是基于空矩形房间或长方体房间的假设显式计算。2) 不再显式计算$D_{i,\,m}$，而是从概率分布中对其进行采样。3) 不再显式计算 $g_{i,\,m}$，而是将其定义为$D_{i,\,m}$ 的函数，并加入随机扰动。

尽管 FRA-RIR 的性能与其他 RIR 仿真工具相当或更优，且仿真速度显著更快，但其核心问题在于，它在多通道场景中的应用受到限制，因为不同麦克风对应的镜像必须对齐，才能正确表示其空间位置。双耳相位差（IPDs）和角度特征（AF）等空间特征也应根据正确的麦克风间距进行对齐。然而，由于所有仿真都是随机进行的，原始 FRA-RIR 方法难以实现这些特性。在此，我们扩展 FRA-RIR，使其支持多通道房间脉冲响应模拟，同时满足上述要求，并将其称为多通道 *RIR* 快速随机近似（*FRAM-RIR*）方法。

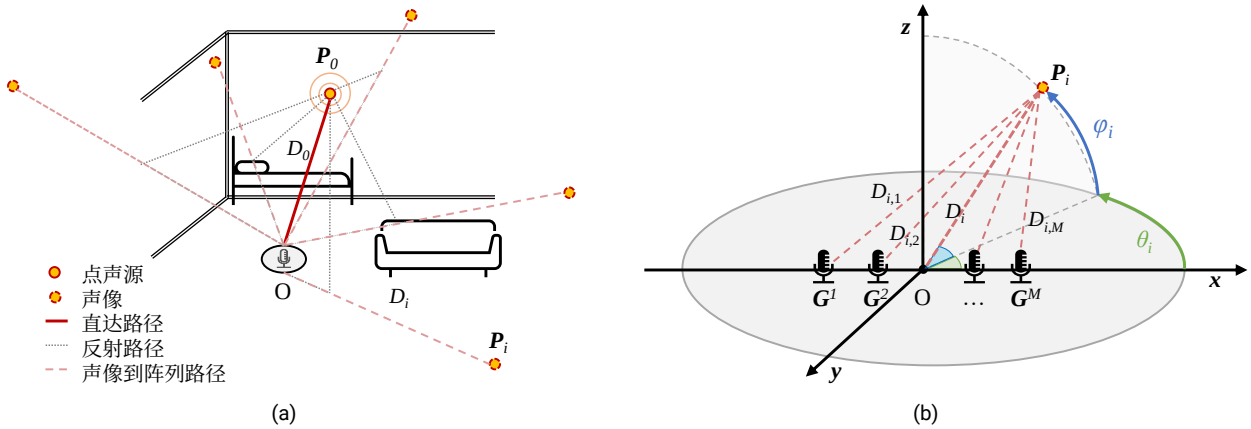


图 1. (a) 鞋盒形房间中的一个点声源 P_0 及其通往接收器（例如麦克风阵列）的直达路径 D_0 、若干早期声像 P_i 及其对应的传播路径；(b) 第 i 个声像可通过声像到接收器的距离 D_i、方位角 θ_i 和仰角 ϕ_i唯一确定。

A. 房间相关统计量的模拟

与 FRA-RIR 通过在方程 2 中随机采样 R 和 T_{60} 来仿真房间相关统计量不同，我们回到 R 的原始定义：假设房间为空的矩形或长方体房间，并通过随机采样房间尺寸来计算该量。我们选择 R 的标准配置，不仅是因为经验表明，其性能与随机采样的 R相当，而且现实房间的特性（例如，不同家具或装饰物具有不同的吸声系数，以及不规则的房间尺寸）也可以通过设置非整数反射次数来仿真，即将$g_i,\ m$ 设为有理数而非整数，并对反射次数施加随机扰动。我们将在第III-C节讨论这一点。

B. 声像的声源—接收器距离模拟

多通道RIR模拟中的一个关键问题是，对于所有虚拟声源，都应准确模拟麦克风之间的TDOA。FRAM-RIR通过在房间内随机生成虚拟声像的三维坐标来模拟虚拟声像的位置，具体做法是在以选定麦克风或房间内预定义三维坐标为中心的球面上，模拟声像到接收器的距离 D、方位角 θ和仰角 ϕ （在不存在歧义时省略下标 m ）。图1（b）展示了一个线性麦克风阵列的示例，其中将阵列中心作为球心，并根据麦克风的位置定义x轴。第 i个虚拟声源的坐标在半径为 D_i、方位角为 θ_i、仰角为 ϕ_i的球面上采样，其中 θ_i在 $[0, 2\pi]$和 ϕ_i 之间均匀采样， $[-\pi/2, \pi/2]$在 D_i 和 D_i 之间均匀采样，而 $DR_i \triangleq D_i/D_0$则通过对与直达路径距离比, $DR_i > 1$ 之间的距离比（DR）进行采样得到，以确保所有虚拟声源的传播距离都长于直达声源。我们进一步假设 DR_i 可以从一个预定义的概率密度函数（PDF）中采样。尽管难以针对不同房间或环境中的虚拟声源距离准确估计通用的PDF，但在此我们做出一个简单假设：由于反射条件日益复杂，随着采样的声源-接收器距离增加，虚拟声源的数量也会增加。因此，我们采用二次函数作为PDF来对距离 D_i进行采样：

$$P(x) = \begin{cases} \dfrac{3x^2}{\beta^3 - \alpha^3}, & \alpha < x \le \beta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

其中， $0 \le \alpha < \beta \le 1$ 是控制ˆ分布范围的标量。为确保 $DR_i > 1$ $DR_i \in$ ，我 $[\alpha, \beta]$ $P\ x$ DR_i DR_i 们首先从 ( ) 中采样，并通过线性缩放将ˆ调整到范围 $[1, c_0T_{60}/d_0]$:

$$DR_i = 1 + \frac{\alpha}{\beta - \alpha}\left(\frac{\hat{DR}_i}{\alpha} - 1\right)\left(\frac{c_0T_{60}}{d_0} - 1\right) \tag{4}$$

其中， c_0T_{60} 是虚拟声源的最大传播距离。实际传播距离 D_i 随后可计算为 $D_i = D_0 \times DR_i$。

C. 声像反射次数的模拟

接下来，我们需要生成每个虚拟声源的反射次数。给定混响时间 T_{60}、直达路径距离 D_0、混响系数 r 和声速 c_0, 我们首先计算最远虚拟声衰减 60 分贝时的反射次数 RR_{max} ：

$$RR_{max} = (\log_{10} c_0T_{60} - \log_{10} d_0 - 3)/\log_{10} r \tag{5}$$

其中， c_0T_{60} 对应最长的可能传播距离。 RR_{max} 可视为虚拟声源可能遇到的最大反射次数，因为模拟RIR通常只考虑 T_{60}范围内的情况。随后，我们将 i个虚拟声源的反射次数 $g_i \in [1, RR_{max}]$ 定义为 D_i的函数，并进一步

对其添加随机扰动：

$$p_i \sim \mathcal{U}(a,b)$$
$$g_i = 1 + (\frac{D_i}{c_0 T_{60}})^2 \cdot (RR_{max} - 1) + p_i \cdot DR_i^{\tau} \quad (6)$$
$$g_i = \max(\min(g_i, RR_{max}), 1)$$

其中，$\mathcal{U}$ 表示均匀分布，p_i 表示反射次数上的随机扰动，$\tau > 0$表示距离收缩因子。这里我们假设，传播距离较长的虚拟声源可能会经历更多次反射。扰动项可以这样解释：总传播距离相近的虚拟声源，其反射次数也可能不同。此外，请注意，由于我们允许 g_i 为非整数，因此，第三节 A 中提到的此类分数次反射，可以解释为当前反射次数为 $\lfloor g_i \rfloor$的虚拟声源具有不同的反射系数$\hat{r}$，其定义为$\hat{r} \triangleq r \cdot e^{q_i \lfloor q_i \rfloor}$。这样，我们便能模拟更多样、更复杂的声源反射模式，并缓解标准镜像源法中反射系数 r 在所有虚拟声源中均保持相同这一不切实际的假设。这里，我们基于这样一种启发式假设，将扰动设为传播距离 D_i 的函数：传播距离较长的虚拟声源也可能遇到材料不同的表面，因此其反射次数（以及反射系数）的变化幅度更大。

D. 房间脉冲响应滤波器的生成

对 D_i 和 g_i进行采样后，将所有虚拟声源经过重缩放的脉冲序列相加，即可生成 $h[n]$。我们首先将 $h[n]$ 初始化为长度为 $L \triangleq \lceil T_{60} r_h f_s \rceil$ 的全零向量，然后将每个虚拟声源依次加到 $h[n]$上：

$$q_i = \min(\lceil \frac{D_i}{c_0} r_h f_s \rceil, L-1) \quad (7)$$
$$h[q_i] = h[q_i] + \frac{r^{g_i}}{D_i} \quad (8)$$

我们将 $g_i = 0$ 设为 $i = 0$（即直达声源）。显然，可以针对所有虚拟声源并行执行这种索引与求和过程，从而加快整体仿真过程。在需要执行去混响的任务中，需要使用早期混响房间脉冲响应滤波器作为早期混响分量的目标。我们将直达声源周围 $[-6, 50]$ ms 的上下文定义为早期混响分量：

$$h_e[n] = \begin{cases} h[n], & -\lceil \frac{6 r_h f_s}{1000} \rceil \le n - \lceil \frac{D_0}{c_0} r_h f_s \rceil \le \lceil \frac{50 r_h f_s}{1000} \rceil \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

随后，将 $h[n]$ 和 $h_e[n]$ 传入相同的下采样–高通滤波–下采样处理流程，该流程如第二节所述。

IV. 实验配置

A. 数据仿真

在两项任务上评估所提出的 FRAM-RIR 的有效性：多通道噪声语音分离，以及多通道联合噪声语音分离与去混响。对于每条语句，我们首先从 Librispeech 数据集的 *train-clean-100* 子集中随机抽取两段语音信号[36], ，然后从*100Nonspeech* 数据集 [37] 和 *MUSAN* 数据集中随机抽取一个噪声信号 [38]。随后使用 RIR 滤波器对两段语音进行空间化处理，并将说话人之间的重叠率随机设定为至少 0.5。噪声模拟为贯穿整条语句的点声源干扰。信干比（SIR）随机设定在[-6, 6] 分贝范围内，信噪比（SNR）则随机设定为 [10, 20] 分贝。

训练期间，使用不同的RIR仿真工具模拟多通道 RIR 滤波器，以进行比较，包括 GPU-RIR [33],PRA-RIR [22] 和所提出的 FRAM-RIR。每个说话人与麦克风阵列中心之间的距离范围为 0.3—6 米。混响时间 T_{60} 在 [0.1 至 0.7] 秒范围内随机采样。房间尺寸按长、宽、高^3 从 3×3×2.5 到 10×10×4 米×××进行采样。假设说话人静止不动。

对于评估集，采用真实录制的多通道房间脉冲响应数据集 [39] 生成 1,000 条测试语句。其中一种录音麦克风阵列配置为包含 8 个阵元的线性阵列，阵元间距为 4-4-4-8-4-4-4 厘米。为简便起见，我们使用其中 4 个阵元，间距为 4-8-4 厘米。测得的混响时间 T_{60}分别为 160 毫秒、360 毫秒和 610 毫秒。声源位于空间网格上，方位角范围为 -90° 至 90°，步长为 15°；声源与麦克风阵列的距离为 1 米或 2 米。

B. 训练配置

我们比较训练期间的两种数据仿真配置，即离线 配置和在线 配置。对于离线 配置，每个训练样本的语音和噪声信号、信噪比、信号干扰比以及对应的 RIR 滤波器都是预先定义的，并在整个训练阶段保持不变。我们总共生成并存储了 100,000 个 RIR 滤波器（12,500 个房间 × 8 声源位置）。这些 RIR 滤波器与语音和噪声信号进行卷积，生成 50,000 个多通道混合信号。相反，在线 配置会为每个样本动态采样所有内容，并在训练期间生成无限数量的语句。

C. 模型配置

为了更全面地评估和比较房间脉冲响应仿真方法，我们训练了若干具有不同输入输出配置和模型架构的单通道及多通道语音分离模型。需要指出的是，我们并非旨在比较模型性能，而是要展示所提出的 FRAM-RIR 仿真方法在相对广泛的常见语音分离和降噪模型中的有效性与效率。对于所有频域模型，我们分别将帧长和帧移设为 32 ms 和 8 ms。

- 单通道双向长短期记忆网络（单通道双向长短期记忆网络）。单通道双向长短期记忆网络的输入特征

是参考通道处混合语谱图的实部和虚部（RI）拼接而成。该网络包含 6 个残差 BLSTM 层，每个方向各有 256 个隐藏单元。输出目标是每个说话人的复数比值掩码（cRM），由 BLSTM 层之后的两个全连接（FC）层估计。

- 单通道频带拆分循环神经网络（SC-BSRNN） [40]。BSRNN 是一种近期提出的、用于音乐源分离 [40]和语音增强 [41]的先进架构。BSRNN 的特点是采用明确的面向任务的子带划分模块，并交替进行频带级和序列级建模。这里，我们实现了原始模型的轻量版本，其中 BSRNN 模块数量设为 8，子带特征维度设为 48，RNN 层中的隐藏单元数量设为 96。
- 单通道 Conformer（SC-Conformer） [42]。SC-Conformer 是一种功能强大的 Conformer 模型，专为频域语音增强而设计。编码器的输入是参考通道上混合语谱图的实部、虚部和幅度部分的组合。解码器将输出估计分解为幅度掩码估计和复数语谱图细化。原始模型包含一个指标判别器，以提升感知质量分数；但为确保与其他模型的训练流程一致，我们在此不纳入该判别器。
- 多通道 BLSTM（MC-BLSTM）。除了 RI 特征外，还将复数双耳相位差（ℂIPD）拼接为额外的输入特征，以提供空间信息 [43]。
- 采用 BLSTM 和 BSRNN 的序列广义维纳滤波器（seq-GWF） [44]。序列波束成形流程，即由预分离、波束成形和后增强串联组成的流程，近期已被提出 [45], [46] ，与传统神经波束成形流程相比，已显示出稳定的性能提升。对于 BLSTM-seq-GWF 模型的预分离模块，其输入特征和输出目标与单通道双向长短期记忆网络相同，而 BLSTM 层数减半。随后，利用预分离模块的粗略估计和多通道混合语谱图应用 GWF 模块。参考通道的混合语谱图 RI 特征、预分离语谱图以及 GWF 模块输出的语谱图会拼接起来，作为后增强模块的输入；该模块与预分离模块采用相同的网络架构。对于 BSRNN-seq-GWF 模型，则以 BSRNN 模块替代 BLSTM 层。BSRNN 模块总数为 8，子带特征维度设为 32。
- 多通道 Conformer（MC-Conformer）。我们只需将 ℂIPD 特征与单通道输入特征拼接，作为额外的空间特征。输出目标与 SC-Conformer 相同。
- FaSNet-TAC [47]。FaSNet-TAC 是对原始滤波求和网络（FaSNet）的扩展 [48] ，其中应用了变换与拼接（TAC）模块，以联合估计所有通道中滤波求和运算所需的滤波器。我们采用与原始文献相同的配置，使用双路径循环神经网络（DPRNN）模块进行序列建模 [49]。

建议感兴趣的读者参阅原始文献，了解上述模型和方法的详细信息。

D. 训练与评估配置

我们采用信噪比结合话语级排列不变训练（uPIT） [50] 来训练所有模型。使用 Adam优化器[51], 初始学习率为 1e-3。每训练 2 个 epoch，学习率便衰减 0.98；若连续 10 个 epoch 未能在验证数据上找到最佳模型，则采用提前停止策略。所有实验均在配备 8 块 NVIDIA Tesla P40 GPU 的单 GPU 服务器上使用 PyTorch 工具包完成。对于在线数据仿真，房间脉冲响应滤波器仿真以及在线信号采样与混合均使用中央处理器或图形处理器并行执行，每个数据加载器配备 8 个工作进程。

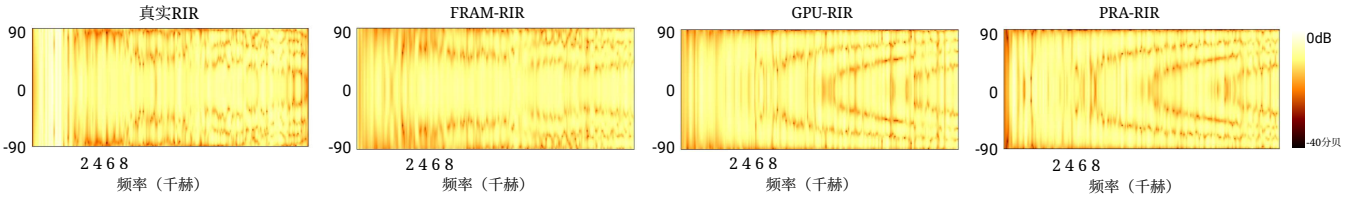
在评估中，我们使用尺度不变信噪比（SI-SNR） [52],信噪比、语音质量感知评估（PESQ）和短时客观可懂度（STOI）作为指标。为了全面评估模型性能，我们还报告了不同方位角范围 [43] （即 <15°、15-45°、45-90° 和 >90°）及重叠比值 [53]（即 50-75% 和 >75%）下的性能。

V. 结果与分析

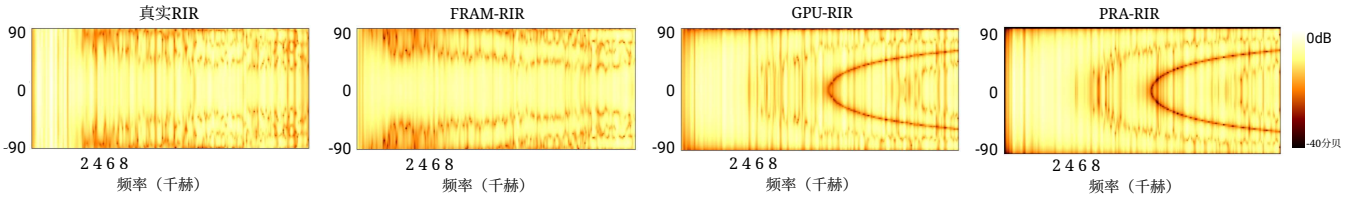
A. 方向属性的验证

我们首先考察 FRAM-RIR 是否能够在第 III 节所述的启发式假设和随机近似下，保留声源准确的空间位置。为此，我们将两个空间特征可视化：理想波束图 和方向特征。理想波束图可视为直达路径声源及其反射的到达方向线索，随后可用于验证仿真 RIR滤波器是否具有正确的入射方向和逼真的反射模式。同样，方向特征还可以进一步确认仿真 RIR滤波器的方向性模式。由于许多现有模型和系统将此类方向特征作为输入特征，因此，真实 RIR滤波器与仿真 RIR滤波器之间的相似性也可以作为一个指标，用于判断仿真 RIR滤波器是否有潜力更好地泛化到真实的录音条件。

图 2 展示了采用理想比率掩码（IRM）的最小方差无失真响应（MVDR）波束形成器 [54] 和采用理想复数谱映射（CSM）的 MVDR 波束形成器 [55], [56]在不同模拟或真实 RIR 滤波器下生成的理想波束图。我们对 RIR 滤波器进行模拟，使房间、麦克风阵列和声源的配置与真实 RIR 滤波器一致。我们可以轻易观察到，由 FRAM-RIR 计算得出的波束图与真实房间脉冲响应更加吻合，而 GPU-RIR 和 PRA-RIR 均



(a) 基于IRM的MVDR波束形成器的波束图



(b) 基于CSM的MVDR波束形成器的波束图

图2. 使用不同RIR滤波器时，(a) 基于IRM的最小方差无失真响应波束形成器和 (b) 基于CSM的最小方差无失真响应波束形成器的波束图。波束形成器系数是根据使用不同RIR滤波器进行空间化的双说话人混响混合信号计算得出的。真实RIR的测量 T_{60} 为360 ms，目标说话人和干扰说话人的到达方向分别为0°和90°。

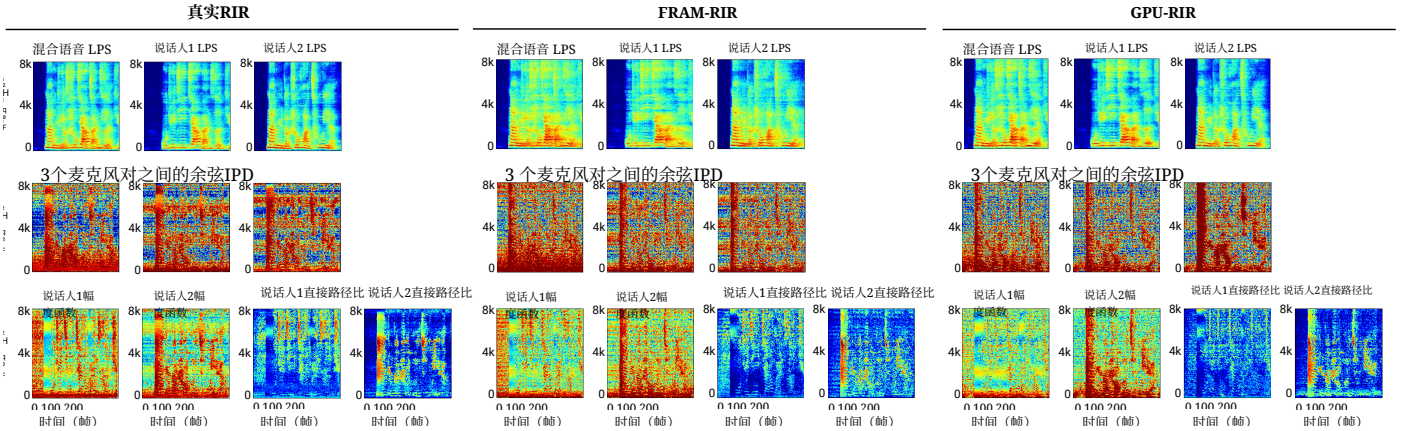


图3. 从使用不同多通道 RIR 滤波器进行空间化的双说话人多通道混合信号中提取的对数功率比（LPS）、空间特征（余弦IPD）和方向特征（角度特征 & DPR）的可视化结果。

受到空旷鞋盒形房间扫掠回声效应的影响。

图3进一步展示了频谱特征和方向特征，其中我们选取对数功率比（LPS）作为频谱特征，并选取余弦IPD（cosIPD）、角度特征（AF） [57]和方向功率比（DPR） [58]作为空间特征。AF和DPR定义如下：

$$\mathrm{TPD}^{(p)}(\theta, f) = \frac{2\pi f}{2(F-1)}\tau_p(\theta) \tag{10}$$

$$\mathrm{IPD}^{(p)}(t, f) = \angle \mathbf{Y}^{p_1}(t, f) - \angle \mathbf{Y}^{p_2}(t, f) \tag{11}$$

$$\mathrm{AF}(\theta, t, f) = \sum_{p=1}^{M} \left\langle \mathrm{TPD}^{(p)}(\theta, f), \mathrm{IPD}^{(p)}(t, f) \right\rangle \tag{12}$$

$$\mathrm{DPR}(v_t, t, f) = \frac{\left|\mathbf{w}_{v_t}^{\mathsf{H}}(f)\mathbf{Y}(t, f)\right|_F^2}{\sum_v^V \left|\mathbf{w}_v^{\mathsf{H}}(f)\mathbf{Y}(t, f)\right|_F^2} \tag{13}$$

其中，$\tau_p(\theta)$ 表示声音从目标方向传来时的 TDOA 经验 y 在第 p 个麦克风对 (p_1, p_2) 上的方向 θ ^{H} 表示多通道混合语谱图，$(\cdot)^{\mathsf{H}}$ 表示矩阵的复共轭，v_t 和 v 分别是指向目标方向和候选方向的固定波束形成器的索引，而 $\mathbf{w} \in \mathbb{C}^{M\times F}$ 是固定波束形成器的系数。

角度特征（AF）衡量理论目标双耳相位差（TPD） [59] 与观测到的耳间相位差之间的相似度，其中，每个目标相位差都是单位脉冲从目标方向入射到麦克风对时产生的理论相位差。目标相位差与耳间相位差越相似，相应时频单元由目标方向声音主导的概率就越高。DPR 定义为目标方向与所有其他方向之间的功率比，其中每个方向的声功率可以通过固定的波束形成器估算，该波束形成器指向相应方向，例如超指向性波束形成器（SD-BF）[58]。该方法假设存在空间多样性，并将空间划分为

固定分辨率的空间网格。从真实房间脉冲响应可以看出，只有在计算方向特征所用的到达方向（即，公式 10 中的 θ 和公式 13 中的 v_t）与直达声的实际到达方向一致时，方向性模式才清晰且具有区分度；后者由脉冲响应模式表示。通过比较模拟RIR的模式，我们可以发现，FRAM-RIR能够实现细粒度到达方向控制，并生成更真实的方向性模式。

表I. 使用不同RIR仿真工具生成四通道RIR的仿真速度。

方法	#线程	#RIR（房间×声源）	速度（秒）
RIR-GEN [60]	1	3k (1k × 3)	5.665
GPU-RIR [33]			0.017*
PRA-RIR [22]			0.647
FRAM-RIR			0.139
GPU-RIR [33]	1	30k (10k × 3)	0.016*
PRA-RIR [22]	8		0.116
FRAM-RIR	8		0.019

表 II. 使用不同RIR仿真工具进行即时批次生成的仿真速度。批次大小设为 1。

方法	工作线程数	速度（秒/批次）
GPU-RIR [33]	1	0.204*
PRA-RIR [22]	1	2.036
PRA-RIR [22]	2	1.070
PRA-RIR [22]	4	0.594
PRA-RIR [22]	8	0.298
FRAM-RIR	1	0.347
FRAM-RIR	2	0.165
FRAM-RIR	4	0.087
FRAM-RIR	8	0.054

B. 仿真速度比较

除了能够生成更逼真的 RIR滤波器外，FRAM-RIR 的另一项优势是，即使在中央处理器上，其 RIR 生成速度也足够快，使在线数据仿真变得容易得多。我们在此比较并报告不同 RIR仿真工具的 RIR 仿真速度，以及将这些工具用于模型训练流程时的混合信号仿真速度。表I比较了四种启用或未启用多线程的 RIR仿真工具，其中 GPU-RIR 使用 NVIDIA Tesla P40 GPU，其他方法使用 Intel Xeon Platinum 8255C CPU @ 2.50 GHz。可以看出，使用单线程时，FRAM-RIR 明显快于其他基于中央处理器的工具；启用多线程后，其速度可与 GPU-RIR 相当[2]。表 II 报告了使用不同工具和不同数量的工作进程，通过 PyTorch数据加载器在线仿真生成一批数据的速度。我们为每个样本在同一房间内随机采样 3 个声源位置。可以发现，使用 2 个工作进程时，FRAM-RIR 的训练样本生成速度可以超过 GPU-RIR；增加工作进程数量还能进一步加快生成速度，这证明了它的高效性和易用性。

[2]请注意，GPU-RIR 不支持多线程，因为它默认使用 CUDA级并行化来加速仿真。

C. 多通道噪声语音分离和去混响任务的性能

表 III 列出了上述单通道和多通道语音分离模型在多通道噪声语音分离任务上的结果。所有模型均使用通过不同 RIR 仿真工具生成的数据进行训练，数据仿真流程采用离线方式或即时（在线）方式。所有评估均在由实录RIR生成的多通道混合信号上进行。文中还报告了理想比率掩码以及基于理想比率掩码的最小方差无失真响应波束形成器的性能，以供参考。

首先可以观察到，与离线数据仿真相比，即时数据仿真始终能提升所有模型和所有 RIR仿真工具的性能，这进一步证实了即时训练在此类任务中的重要性。此外可以看到，在大多数模型架构中，GPU-RIR 和 FRAM-RIR 的表现优于 PRA-RIR。由于 GPU-RIR 和 FRAM-RIR 的仿真速度都快于 PRA-RIR，因此可以考虑用这两种工具替代 PRA-RIR，以加速并改进即时数据仿真。对 GPU-RIR 和 FRAM-RIR 的比较表明，使用这两种工具训练的模型在不同模型架构和系统流程下表现各异；总体而言，两种工具的性能相当。对于表现最佳的模型，即 MC-Conformer 和 BSRNN-seq-GWF，GPU-RIR 的表现更好，尤其是在方位角差较小的语句上。原因之一是，由于对虚拟声源传播距离进行采样，FRAM-RIR 模拟出的早期反射声传播路径比基于标准 ISM 的方法更多；当 T_{60} 较小时且说话人距离较近时，这可能会降低方向特征的区分度。重新平衡仿真数据的分布，或根据房间、说话人距离或 T_{60} 进行仿真，或许可以缓解这一问题，但我们将其留作未来工作。

我们进一步在更具挑战性的联合分离与去混响任务上比较这些 RIR 工具。通过评估模型在执行去混响时能否很好地泛化到真实RIR滤波器，进一步检验这些工具模拟的混响模式。我们使用与表 III 相同的数据和流程，但将训练目标从混响的干净源信号改为干净源信号的早期反射成分，这些成分由第二节 III-D 中定义的早期混响 RIR 滤波器计算得出。根据表 III 的结果，我们选择 SC-BSRNN 和 BSRNN-seq-GWF 模型分别作为单通道和多通道模型，因为它们在模型复杂度与性能之间取得了良好的平衡。表 IV 显示，在联合降噪、分离和去混响任务中，这两个模型在不同条件下的性能趋势相似；FRAM-RIR 的模型性能仍与 GPU-RIR 不相上下。

表 III. 多通道噪 声分离在使用真实实录房间脉冲响应模拟的测试数据上的性能比较 .

模型	参数数量	MAC (G/s)	RIR	data	尺度不变信号失真比 (分贝) 方位角差 <15°	15-45°	45-90°	>90°	重叠率 50-75%	>75%	Avg	SNR (dB)	短时客观可懂度	感知语音质量评估
混合	-	-	-	-	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.52	1.10
IRM	-	-	-	-	11.1	11.0	11.2	11.1	11.3	10.9	11.1	11.4	0.89	1.76
理想比率掩蔽最小方差无失真响应	-	-	-	-	4.0	4.5	6.2	6.6	5.8	5.8	5.8	5.8	0.70	1.28
单通道双向长短期记忆网络	5.6M	0.74	GPU	离线	4.8	3.9	4.2	4.3	4.3	4.1	4.2	5.8	0.67	1.24
			PRA		3.8	3.0	3.3	3.4	3.5	3.1	3.3	5.1	0.66	1.24
			FRAM		4.6	3.6	4.1	4.2	4.5	3.6	4.1	5.7	0.68	1.27
			GPU	在线	6.3	5.5	5.8	6.1	6.1	5.6	5.8	7.0	0.72	1.30
			PRA		5.2	4.5	4.9	5.1	5.1	4.6	4.9	6.3	0.68	1.27
			FRAM		6.5	5.5	5.8	6.1	6.2	5.5	5.9	7.0	0.71	1.31
MC-BLSTM	5.8M	0.75	GPU	在线	4.4	6.7	8.9	9.7	8.8	8.0	8.4	9.1	0.82	1.46
			PRA		4.9	7.1	9.5	10.2	9.3	8.5	8.9	9.5	0.84	1.49
			FRAM		5.3	7.3	9.7	10.3	9.5	8.6	9.1	9.7	0.83	1.51
BLSTM-seq-GWF	7.0M	9.71	GPU	离线	5.4	4.7	5.1	7.0	6.1	5.9	6.1	7.2	0.74	1.34
			PRA		6.0	5.1	6.7	7.6	6.8	6.3	6.5	7.7	0.75	1.37
			FRAM		6.3	5.7	7.2	7.9	7.2	6.8	7.0	8.0	0.76	1.38
			GPU	在线	10.1	9.8	10.7	11.0	11.0	10.0	10.5	11.0	0.84	1.62
			PRA		8.4	7.7	9.3	9.7	9.2	8.7	9.0	9.6	0.81	1.48
			FRAM		10.3	9.6	10.7	11.0	10.9	10.1	10.5	11.0	0.85	1.62
FaSNet-TAC	2.8M	10.05	GPU	离线	4.4	4.6	7.2	9.0	7.2	6.9	7.0	8.0	0.76	1.37
			PRA		2.6	3.0	6.3	8.1	6.0	5.8	5.9	7.1	0.72	1.32
			FRAM		6.3	4.9	5.4	5.8	5.8	5.1	5.5	6.8	0.71	1.31
			GPU	在线	4.9	5.4	7.9	9.3	7.9	7.2	7.6	8.4	0.77	1.37
			PRA		4.2	4.6	6.8	7.6	6.6	6.2	6.4	7.4	0.73	1.32
			FRAM		6.9	6.3	7.6	8.6	7.8	7.2	7.6	8.4	0.77	1.39
SC-Conformer	1.8M	31.62	GPU	在线	13.2	12.4	13.2	13.0	13.3	12.4	12.9	13.2	0.92	1.91
			PRA		12.7	12.3	12.8	12.7	13.0	12.2	12.6	12.9	0.91	1.87
			FRAM		13.2	12.4	13.1	13.0	13.3	12.4	12.9	13.1	0.90	1.89
MC-Conformer	1.8M	31.63	GPU	在线	13.3	13.3	14.2	14.5	14.4	13.6	14.0	14.2	0.86	2.05
			PRA		12.8	12.9	14.0	14.3	14.1	13.3	13.7	13.9	0.94	2.06
			FRAM		11.5	12.6	13.8	13.9	13.8	12.9	13.4	13.6	0.93	1.99
SC-BSRNN	3.3M	6.38	GPU	在线	12.4	11.8	12.4	12.4	12.6	11.8	12.2	12.5	0.91	1.95
			PRA		12.0	11.3	11.9	11.9	12.2	11.3	11.7	12.1	0.90	1.94
			FRAM		12.5	11.8	12.4	12.3	12.6	11.8	12.2	12.5	0.91	1.96
BSRNN-seq-GWF	3.7M	14.62	GPU	在线	13.6	13.4	14.2	14.3	14.3	13.6	14.0	14.2	0.93	2.13
			PRA		12.8	12.6	13.7	13.9	13.8	13.0	13.4	13.6	0.92	2.04
			FRAM		13.2	12.9	13.9	14.2	14.1	13.3	13.7	13.9	0.92	2.07

表 IV. 在采用真实录制的房间脉冲响应模拟的测试数据上，多通道噪声语音分离与去混响联合处理性能对比。

模型	参数数量	RIR	data	尺度不变信号失真比（分贝） 方位角差 <15°	15-45°	45-90°	>90°	重叠率 50-75%	>75%	Avg	SNR	STOI	PESQ
混合	-	-	-	-1.8	-1.6	-1.7	-1.5	-1.6	-1.6	-1.6	-0.9	0.50	1.08
IRM	-	-	-	7.0	7.3	7.3	7.6	7.6	7.2	7.4	8.3	0.85	1.53
理想比率掩蔽最小方差无失真响应	-	-	-	2.6	3.0	4.3	4.9	4.1	4.1	4.1	5.1	0.67	1.23
SC-BSRNN	3.3M	GPU	在线	7.5	7.4	7.6	7.9	7.9	7.4	7.7	8.7	0.84	1.58
		PRA		7.5	7.0	7.4	7.8	7.8	7.1	7.5	8.5	0.84	1.58
		FRAM		7.7	7.5	7.8	8.1	8.2	7.5	7.8	8.9	0.85	1.60
BSRNN-seq-GWF	3.7M	GPU	在线	8.9	9.1	9.5	10.2	9.9	9.3	9.6	10.4	0.88	1.82
		PRA		8.5	8.7	9.2	9.9	9.5	9.0	9.3	10.1	0.88	1.80
		FRAM		8.8	8.9	9.5	10.1	9.8	9.2	9.5	10.4	0.88	1.78

D. T_{60} 课程训练

某些应用可能要求模型在房间大小范围很广的场景中保持鲁棒性，例如从小卧室或会议室到大型教室或音乐厅。我们通过实验发现，直接使用如此大的T_{60} 范围训练模型，可能导致性能欠佳且收敛速度较慢。一个可能的原因是，较大的T_{60} 会使方向特征呈现更复杂的模式，导致模型在训练初期难以充分利用频谱和空间特征。在此，我们验证了课程训练这一训练技术，该技术旨在缓解这一问题 [18], ，并应用于基于 FRAM-RIR 的在线训练流程。给定在线数据仿真流程所需覆盖的最大 T_{60} 值，我们先在相对较小的 T_{60} 范围内采样以开始模型训练，并随着训练推进逐步提高该范围的上限。实验发现，将初始 T_{60} 范围设为 [50, 100]毫秒，并在之后的每个训练轮次将上限增加 50 毫秒，直至达到预先设定的最大 T_{60} ，能够有效加快模型收敛并提升性能。图 4 展示了采用或未采用 T_{60}课程训练的两个系统——MC-BLSTM 模型和 BSRNN-seq-GWF 模型——的性能。我们对两个模型均进行了最多 100 个训练轮次，并报告了不同轮次下的测试集尺度不变信噪比得分。可以看出，课程训练能有效加快两个模型的收敛速度，尤其是在训练初期；对于 MC-BLSTM，采用 T_{60} 课程训练后性能提升显著。结果表明， T_{60} 课程训练可以有效补充在线训练流程。

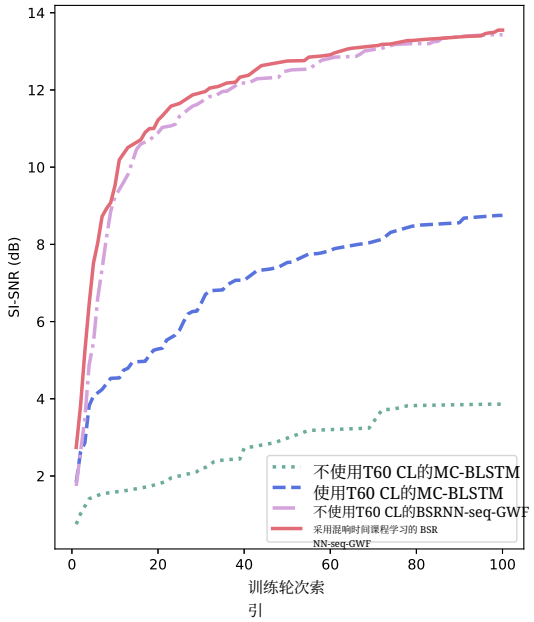


图 4. 应用T_{60} 课程学习（w/ T_{60} CL）或未应用（w/o T_{60} CL）时，MC-BLSTM 和 BSRNN-seq-GWF 模型在测试数据上的尺度不变信噪比（分贝）随训练轮次的变化。

VI. 结论与未来工作

本文提出了多通道房间脉冲响应的快速随机近似（FRAM-RIR），这是一种用于语音处理系统训练中数据增强、能够快速模拟逼真RIR滤波器的方法。FRAM-RIR不仅能够模拟出比现有工具更符合真实RIR滤波器特征的RIR滤波器，而且在单线程下显著快于其他基于中央处理器的方法，在多线程下也快于图形处理器加速的方法。实验结果表明，FRAM-RIR可实现快速的在线训练，且性能不逊于甚至优于

在多通道噪声语音分离和去混响任务中，与多种模型广泛使用的其他房间脉冲响应仿真工具相比。此外，我们提出了一种 T_{60} 课程训练策略，以加快训练阶段的收敛速度并提升模型性能。未来的工作包括将 FRAM-RIR 应用于并验证其在其他类型语音处理任务中的效果，以及扩展其设计与应用，使之适用于更复杂的场景，例如支持移动声源的房间脉冲响应仿真和麦克风建模。

参考文献

[1] C. Kim、A. Misra、K. Chin、T. Hughes、A. Narayanan、T. N. Sainath 和 M. Bacchiani，“在虚拟

用于训练深度神经网络，以实现 Google Home 的远场语音识别，”*Interspeech*会议论文集，第379–383页，2017年。[2] T.-S. Nguyen、S. Stueker、J. Niehues 和 A. Waibel，“通过在线数据增强改进序列到序列语音识别训练”，载于 *ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第7689–7693页。[3] S. Zhang、Y. Kong、S. Lv、Y. Hu 和 L. Xie，“基于 FT-LSTM 的复数网络，用于联合声学回声消除和语音增强”，载于 *Interspeech*会议论文集，2021年，第4758–4762页。[4] N. Zeghidour 和 D. Grangier，“WaveSplit：通过说话人聚类实现端到端语音分离”，*IEEE/ACM*音频、语音与语言处理汇刊（*TASLP*），第29卷，第2840–2849页，2021年。[5] C. Subakan、M. Ravanelli、S. Cornell、M. Bronzi 和 J. Zhong，“语音分离只需注意力机制”，载于 声学、语音与信号处理（*ICASSP*），*2021*年*IEEE*国际会议。IEEE，2021年，第21–25页。[6] T. K. Lam、M. Ohta、S. Schamoni 和 S. Riezler，“用于序列到序列自动语音识别的在线对齐数据增强”，*arXiv*预印本 *arXiv:2104.01393*，2021年。[7] V. Valimaki、J. D. Parker、L. Savioja、J. O. Smith 和 J. S. Abel，“人工混响的五十年”，*IEEE*音频、语音与语言处理汇刊（*TASLP*），第20卷，第5期，第1421–1448页，2012年。[8] V. V¨alim¨aki、J. Parker、L. Savioja、J. O. Smith 和 J. Abel，“人工混响的五十余年”，载于 音频工程学会会议：第*60*届国际会议：梦想（音频、音乐和语音的去混响与混响）。音频工程学会，2016年。[9] J. B. Allen 和 D. A. Berkley，“用于高效模拟小房间声学的镜像法”，美国声学学会期刊，第65卷，第4期，第943–950页，1979年。[10] E. A. Lehmann 和 A. M. Johansson，“弥散混响模型

用于高效地通过镜像源模拟计算房间脉冲响应，”IEEE/ACM音频、语音与语言处理汇刊（TASLP），第18卷，第6期，第1429–1439页，2009年。[11] Z. Tang、L. Chen、B. Wu、D. Yu和D. Manocha，“利用扩散声学模拟改进混响语音训练”，载于《声学、语音与信号处理（ICASSP）》，2020年IEEE国际会议。IEEE，2020年，第6969–6973页。[12] L. Couvreur和C. Couvreur，“在强混响环境中将人工混响用于自动语音识别”，载于第二届IEEE比荷卢信号处理研讨会（SPS-2000）论文集，希尔瓦伦贝克，荷兰。Citeseer，2000年，第S001–S004页。[13] I. Tashev和D. Allred，“通过混响抑制提升语音识别”，载于免提语音通信与麦克风阵列联合研讨会（HSCMA），2005年。[14] X. Chang、W. Zhang、Y. Qian、J. Le Roux和S. Watanabe，“Mimo-speech：端到端多通道多说话人语音识别”，载于2019年IEEE自动语音识别与理解研讨会（ASRU）。IEEE，2019年，第237–244页。[15] Y. Zhao、Z.-Q. Wang和D. Wang，“一种用于噪声混响语音增强的两阶段算法”，载于《声学、语音与信号处理（ICASSP）》，2017年IEEE国际会议。IEEE，2017年，第5580–5584页。[16] Y. Zhao和D. Wang，“采用具有时频注意力机制的DenseUNet进行噪声混响语音增强”，载于Interspeech论文集，第2020卷，2020年，第3261–3265页。[17] M. Maciejewski、G. Wichern、E. McQuinn和J. Le Roux，“WHAMR!：噪声混响单通道语音分离”，载于《声学、语音与信号处理（ICASSP）》，2020年IEEE国际会议。IEEE，2020年，第696–700页。[18] R. Aralikatti、A. Ratnarajah、Z. Tang和D. Manocha，“通过多阶段训练和课程学习改进混响语音分离”，arXiv预印本 arXiv:2107.09177，2021年。[19] K. Han、Y. Wang、D. Wang、W. S. Woods、I. Merks和T. Zhang，“学习用于语音去混响和降噪的频谱映射”，IEEE/ACM音频、语音与语言处理汇刊（TASLP），第23卷，第6期，第982–992页，2015年。[20] B. Wu、K. Li、M. Yang和C.-H. Lee，“一种基于深度神经网络、感知混响时间的语音去混响方法”，IEEE音频、语音与语言处理汇刊（TASLP），第25卷，第1期，第102–111页，2016年。[21] T. Funkhouser、N. Tsingos、I. Carlbom、G. Elko、M. Sondhi、J. E. West、G. Pingali、P. Min和A. Ngan，“一种用于交互式建筑声学的波束追踪方法”，《美国声学学会期刊》，第115卷，第2期，第739–756页，2004年。[22] R. Scheibler、E. Bezzam和I. Dokmani´c，“Pyroomacoustics：用于音频房间模拟和阵列处理算法的Python软件包”，载于《声学、语音与信号处理（ICASSP）》，1997年IEEE国际会议。IEEE，2018年，第351–355页。[23] H. J¨arvel¨ainen和M. Karjalainen，“使用天鹅绒噪声进行混响建模”，载于音频工程学会会议：第30届国际会议：智能音频环境。音频工程学会，2007年。[24] V. V¨alim¨aki、H.-M. Lehtonen和M. Takanen，“对不同脉冲密度下天鹅绒噪声及其变体的感知研究”，IEEE音频、语音与语言处理汇刊，第21卷，第7期，第1481–1488页，2013年。[25] P. Masztalski、M. Matuszewski、K. Piaskowski和M. Romaniuk，“StoRIR：用于音频数据增强的随机房间脉冲响应生成”，Interspeech论文集，第2857–2861页，2020年。[26] A. Ratnarajah、Z. Tang和D. Manocha，“IR-GAN：用于远场语音识别的房间脉冲响应生成器”，arXiv预印本 arXiv:2010.13219，2020年。[27] ——，“TS-RIR：用于语音增强的转换后合成房间脉冲响应”，arXiv预印本 arXiv:2103.16804，2021年。[28] A. Ratnarajah、S.-X. Zhang、M. Yu、Z. Tang、D. Manocha和D. Yu，“FAST-RIR：快速神经漫反射房间脉冲响应生成器，”载于声学、语音与信号处理（*ICASSP*），*2022*年*IEEE*国际会议。IEEE，2022年，第571–575页。[29] K. Kiyohara、K. Furuya 和 Y. Kaneda，“规则形状混响室中感知到的扫掠回声”，美国声学学会期刊，第111卷，第2期，第925–930页，2002年。[30] E. De Sena、N. Antonello、M. Moonen 和 T. Van Waterschoot，“房间声学仿真中矩形几何形状的建模”，*IEEE/ACM*音频、语音与语言处理汇刊（*TASLP*），第23卷，第4期，第774–786页，2015年。[31] T. Hanyu，“一种基于镜像声源分布并考虑房间形状复杂度的逼真三维混响生成新算法”，载于第*133*届音频工程学会大会。音频工程学会，2012年。[32] Z.-h. Fu 和 J.-w. Li，“用于房间脉冲响应计算的基于GPU的镜像法”，多媒体工具与应用，第75卷，第9期，第5205–5221页，2016年。[33] D. Diaz-Guerra、A. Miguel 和 J. R. Beltran，“gpuRIR：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1–19页，2020年。[34] Y. Luo 和 J. Yu，“FRA-RIR：镜像法的快速随机近似”，*arXiv*预印本 *arXiv:2208.04101*，2022年。[35] L. L. Beranek，“萨宾公式和艾林公式的分析及其在音乐厅观众和座椅吸声中的应用”，美国声学学会期刊，第120卷，1410，2006年3，1399页 – 。[36] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“Librispeech：基于公有领域有声读物的自动语音识别语料库”，载于声学、语音与信号处理（*ICASSP*），*2015*年*IEEE*国际会议。IEEE，2015年，第5206–5210页。[37] G. Hu 和 D. Wang，“基频估计与有声语音分离的串联算法”，*IEEE*音频、语音与语言处理汇刊，第18卷，第8期，第2067–2079页，2010年。[38] D. Snyder、G. Chen 和 D. Povey，“Musan：音乐、语音与噪声语料库”，*arXiv*预印本 *arXiv:1510.08484*，2015年。[39] E. Hadad、F. Heese、P. Vary 和 S. Gannot，“多种声学环境下的多通道音频数据库”，载于*2014*年第*14*届国际声学信号增强研讨会（*IWAENC*）。IEEE，2014年，第313–317页。[40] Y. Luo 和 J. Yu，“使用频带分割循环神经网络进行音乐源分离”，*arXiv*预印本 *arXiv:2209.15174*，2022年。[41] J. Yu、Y. Luo、H. Chen、R. Gu 和 C. Weng，“使用频带分割循环神经网络实现高保真语音增强”，*arXiv*预印本 *arXiv:2212 00406*，2022年。。[42] S. Abdulatif、R. Cao 和 B. Yang，“CMGAN：用于单声道语音增强的基于Conformer的度量生成对抗网络”，*arXiv*预印本 *arXiv:2209.11112*，2022年。[43] L. Chen、M. Yu、D. Su 和 D. Yu，“用于改善多通道语音分离的多频带置换不变训练与模型集成”，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2019年，第705–709页。[44] Y. Luo，“用于多通道神经分离系统的时域实值广义维纳滤波器”，*IEEE/ACM*音频、语音与语言处理汇刊（*TASLP*），第30卷，3019，2022年3008页 – 。[45] Z.-Q. Wang、H. Erdogan、S. Wisdom、K. Wilson、D. Raj、S. Watanabe、Z. Chen 和 J. R. Hershey，“用于语音分离与增强的序列多帧神经波束成形”，载于*2021*年*IEEE*口语语言技术研讨会（*SLT*）。IEEE，2021年，第905–911页。[46] Z.-Q. Wang 和 D. Wang，“用于语音去混响的多麦克风复数频谱映射”，载于*ICASSP 2020-2020*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第486–490页。[47] Y. Luo、Z. Chen、N. Mesgarani 和 T. Yoshioka，“端到端麦克风排列与数量不变的多通道语音分离

”，载于声学、语音与信号处理（*ICASSP*），*2020*年*IEEE*国际会议。IEEE，2020年，第6394–6398页。[48] Y. Luo、E. Ceolini、C. Han、S.-C. Liu 和 N. Mesgarani，“FaSNet：用于多麦克风音频处理的低延迟自适应波束成形”，载于自动语音识别与理解（*ASRU*），*2019*年*IEEE*研讨会。IEEE，2019年。[49] Y. Luo、Z. Chen 和 T. Yoshioka，“双路径循环神经网络：用于时域单通道语音分离的高效长序列建模”，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第46–50页。[50] M. Kolbæk、D. Yu、Z.-H. Tan 和 J. Jensen，“采用深度循环神经网络话语级排列不变训练的多说话人语音分离”，*IEEE/ACM*音频、语音与语言处理汇刊（*TASLP*），第25卷，1913，2017年10，1901页 – 。[51] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于第*3*届学习表示国际会议，2015年。[52] J. Le Roux、S. Wisdom、H. Erdogan 和 J. R. Hershey，“SDR——半生不熟还是恰到好处？”，载于声学、语音与信号处理（*ICASSP*），*2019*年*IEEE*国际会议。IEEE，2019年，第626–630页。[53] Z. Chen、T. Yoshioka、L. Lu、T. Zhou、Z. Meng、Y. Luo、J. Wu、X. Xiao 和 J. Li，“连续语音分离：数据集与分析”，载于*2020*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2020年，第7284–7288页。[54] J. Heymann、L. Drude 和 R. Haeb-Umbach，“用于声学波束形成的神经网络频谱掩蔽估计”，载于声学、语音与信号处理（*ICASSP*），*2016*年*IEEE*国际会议。IEEE，2016年，第196–200页。[55] K. Tan、Z.-Q. Wang 和 D. Wang，“神经频谱空间滤波”，*IEEE/ACM*音频、语音与语言处理汇刊，第30卷，第605–621页，2022年。[56] Z.-Q. Wang、P. Wang 和 D. Wang，“用于单通道和多通道语音增强及鲁棒自动语音识别的复数频谱映射”，*IEEE/ACM*音频、语音与语言处理汇刊（*TASLP*），第28卷，第1778–1787页，2020年。[57] Z. Chen、T. Yoshioka、X. Xiao、L. Li、M. L. Seltzer 和 Y. Gong，“固定波束形成器与语音分离网络的高效集成，用于多通道远场语音分离”，载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第5384–5388页。[58] R. Gu、L. Chen、S.-X. Zhang、J. Zheng、Y. Xu、M. Yu、D. Su、Y. Zou 和 D. Yu，“神经空间滤波器：借助方向信息进行目标说话人语音分离”，载于*Interspeech*论文集，2019年，第4290–4294页。[59] R. Gu、S.-X. Zhang、Y. Zou 和 D. Yu，“迈向用于时域和频域语音分离的统一全神经波束成形”，*IEEE/ACM*音频、语音与语言处理汇刊，第31卷，第849–862页，2023年。[60]“RIR生成器”，
h
t
t
p
s://
w
w
w.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator。