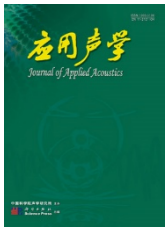


应用声学
Journal of Applied Acoustics
ISSN 1000-310X,CN 11-2121/O4

《应用声学》网络首发论文

题目：房间脉冲响应模拟法及混响时间盲估计应用
作者：郑凯桐，郑成诗，张玉龙，桑晋秋，李晓东
收稿日期：2022-01-22
网络首发日期：2022-07-27
引用格式：郑凯桐，郑成诗，张玉龙，桑晋秋，李晓东．房间脉冲响应模拟法及混响时间盲估计应用[J/OL]．应用声学.
https://kns.cnki.net/kcms/detail/11.2121.O4.20220726.1838.002.html

网络首发时间：2022-07-27 10:58:57
网络首发地址：https://kns.cnki.net/kcms/detail/11.2121.O4.20220726.1838.002.html

房间脉冲响应模拟法及混响时间盲估计应用*

郑凯桐 [1,2]　郑成诗 [1,2]　张玉龙 [1,2]　桑晋秋 [1,2†]　李晓东 [1,2]

(1 中国科学院声学研究所　北京　100190)
(2 中国科学院大学　北京　100049)

摘要：在构建混响语声数据集时，由于缺乏真实长混响房间脉冲响应且模拟的房间脉冲响应与真实不符，因而导致数据驱动的混响时间盲估计模型性能下降。提出了一种基于条件生成对抗网络的房间脉冲响应模拟法，该方法利用真实的房间脉冲响应训练条件生成对抗网络，可以根据指定的混响时间模拟更加真实的房间脉冲响应。使用不同方法模拟的房间脉冲响应构建训练集用于训练盲估计模型，通过声学实验评估模型性能。实验结果表明，由该方法模拟的房间脉冲响应训练的估计模型在不同信噪比下均具有最小的均方根误差且在长混响情况下显著优于其他模型。

关键词：房间脉冲响应模拟；混响时间；盲估计

中图法分类号：TP391.9　　**文献标识码：**A　　**DOI：**

A room impulse response generator and its application on blind reverberation time estimation

ZHENG Kaitong [1,2]　ZHENG Chengshi[1,2]　ZHANG Yulong[1,2]　SANG Jinqiu[1,2]　LI Xiaodong[1,2]

(1 *Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China*)
(2 *University of Chinese Academy of Sciences, Beijing 100049, China*)

Abstract: The performance of data-driven blind reverberation time estimation network degrades because real room impulse response (RIR) lacks data with long reverberation time and there is a gap between real RIRs and simulated RIRs. In this paper, a room impulse response generator based on conditional Generative Adversarial Network is introduced. Trained with real RIRs, this network can generate room impulse responses with desired reverberation time. RIRs simulated by different methods are used to train the blind reverberation time estimators. Acoustic experiments are conducted to evaluate the performance of these estimators. The experimental results show that the estimator trained with simulated RIRs generated by the proposed method has a lower root mean square error than the baseline methods in different noisy scenarios and large reverberation scenarios.

Keywords: Room impulse response generator; Blind estimation; Reverberation time

2022-01-22 收稿; 2022-02-24 定稿
*国家重点研发计划项目(2021YFB3201702)
作者简介：郑凯桐(1997－)，男，福建泉州人，硕士研究生，研究方向：信号与信息处理。
†通信作者 E-mail: sangjinqiu@mail.ioa.ac.cn

0 引言

在房间中，接收点处会接收到从声源发出的直达声和经过房间表面多次反射的反射声。声源在停止发声后，声音在一段时间内仍可被人耳听见的现象叫做混响[1]。声源在停止发声后声压级下降 60 dB 所需要的时间被定义为混响时间(Reverberation time, T_{60})。混响时间是表征房间声学特性的重要参数，主要由房间的尺寸以及表面材料特性所决定。混响时间对语声清晰度、空间感知等人耳主观听觉有重要影响[2]。混响时间可以使用房间脉冲响应(Room impulse response, RIR)通过 Schroder 反向积分法计算[3]。然而，测量 RIR 需要专业的仪器和测量人员并且需要花费较多时间，不便于大规模测量。因此，需要提出更加方便快捷的混响时间盲估计方法。一种简捷方法是直接从混响语声信号中估计混响时间，省去耗时的声学测量和复杂的声学仪器。

近年来，已有许多相关的混响时间盲估计方法被提出[4–7]。这些方法可主要分为基于传统信号处理的方法和基于深度学习的方法。在 2015 年举办的环境声学特性(Acoustic Characterization of Environments, ACE)挑战赛[8]中，各种混响时间盲估计方法被评估，其中基于传统信号处理的混响时间盲估计方法[7]取得了最佳性能。近年来随着深度学习的发展，许多基于深度学习的混响时间盲估计方法[9-10]被提出，并且在仿真数据集下取得超越传统信号处理方法的性能。然而，这些方法主要存在以下两方面问题：首先，估计方法在长混响情况下性能不佳。这是由于在构建混响语声训练集时，通常使用 RIR 和纯净语声卷积模拟混响语声。大部分真实 RIR 的混响时间在 0.6~0.8 s 之间，会导致长混响时间 RIR 的缺失，造成不同混响时间对应的 RIR 数量不均衡，进而影响长混响时间下的模型性能。其次，只使用人工合成的混响语声对方法进行测试。在假设房间系统是线性时不变(Linear time invariant, LTI)系统的前提下，可以使用 RIR 和纯净语声卷积模拟混响语声。然而在真实情况下，房间系统并不严格满足 LTI 系统假设[1]。之前的研究大多使用 RIR 和纯净语声卷积模拟的混响语声而没有使用不同环境的真实语声对模型进行测试，没有验证模型在真实情况下的性能。

针对估计方法在长混响情况下性能不佳的问题，许多计算机模拟 RIR 的方法被提出。传统的模拟方法主要被分为 3 类。第一类是基于波动声学的方法，如有限元法[11]和边界元法[12]。这类方法可以准确地模拟房间中的声波传输，但是对于高频声波的计算复杂度比较高。第二类是基于射线声学的方法，如虚源法[13]和路径追踪法[14]。这类方法因为计算复杂度较低而被广泛使用，但是该类方法对于低频声波的仿真存在局限。第三类是基于 RIR 统计模型的方法，如 Schroder 统计模型，但是该类方法模拟的 RIR 与真实的 RIR 在早期混响部分存在显著差异。以上介绍的传统 RIR 模拟方法均适用于特定的理论假设条件，其模拟的 RIR 与实际的 RIR 存在一定差异，造成深度学习模型在实际场景下性能下降。因此，有学者提出基于生成对抗网络(Generative adversarial network, GAN)的 RIR 模拟方法[15]，提升了深度学习模型在远场语声识别任务下的准确度。然而，该方法无法模拟具有特定混响时间的 RIR 且远场语声识别任务下使用的混响时间大多在 0.8 s 以下，缺乏长混响时间对应的 RIR 数据库。作者随后提出一种快速的 RIR 模拟方法[16]，极大提升了 RIR 的模拟速度。然而，该方法也是主要用于模拟中短混响(0.2~0.7 s)的 RIR，并且与真实房间的 RIR 仍存在差距。

针对混响时间盲估计任务，本文提出了一种基于条件生成对抗网络的 RIR 模拟方法。在训练阶段，真实房间的 RIR 和其对应的混响时间被用于训练条件生成对抗网络。训练完成后，输入所需的混响时间，该网络可以模拟对应混响时间的 RIR。由于使用真实房间的 RIR 进行训练，模型模拟的 RIR 将与真实房间的 RIR 更加接近。为了验证该方法的有效性，将该方法模拟的 RIR 和虚源法、Schroder 统计模型模拟的 RIR 以及真实 RIR 分别构建训练数据集对相同混响时间估计网络框架进行训练。使用真实混响语声对使用不同 RIR 训练的混响估计网络进行测试，间接证明该方法的有效性。

针对先前研究只使用人工合成的混响语声对方法进行测试的问题，本文在 4 个真实房间中进行了混响时间测量及实际环境的语声、噪声录声。4 个具有不同尺寸和不同混响时间的房间包括一间

办公室、两间会议室以及一间隔声室。录声内容包括混响语声和不同类型的噪声。本文使用这些真实数据作为测试集，以验证同一深度学习模型框架经不同 RIR 数据集训练后在真实环境下混响时间盲估计的性能。

1 提出的 RIR 模拟法

GAN[17]是一种将低维数据映射成高维数据的模拟模型。GAN 包含互相交替训练的生成器和判别器。生成器的训练目标是将噪声向量从噪声的分布中映射到目标数据的分布，而判别器的训练目标是区分输入是由生成器产生数据的还是真实数据。在训练过程中，生成器朝着生成使判别器难以区分的样本的方向进行优化，而判别器朝着能够区分生成的数据和真实数据的方向进行优化，生成器和判别器处于对抗博弈的状态。在一段时间的交替训练后，生成器生成的数据将与真实数据难以分辨，从而可以使用生成器进行数据增广等应用。条件生成对抗网络[18]是 GAN 的一种扩展形式。与 GAN 不同之处在于，条件生成对抗网络的生成器在生成样本时需要额外输入条件向量，而判别器在区分样本时也需要额外的输入条件向量，通过样本和对应的条件向量区分生成样本和真实样本。

1.1 算法流程图

本文使用混响时间作为条件生成对抗网络的额外输入条件。通过控制这个输入条件，可以人为控制生成器模拟的 RIR 的混响时间，使模拟的 RIR 数据集涵盖大范围的混响时间。图 1 和图 2 中分别是基于 GAN 的 RIR 模拟方法和基于条件生成对抗网络的 RIR 模拟方法算法流程图。

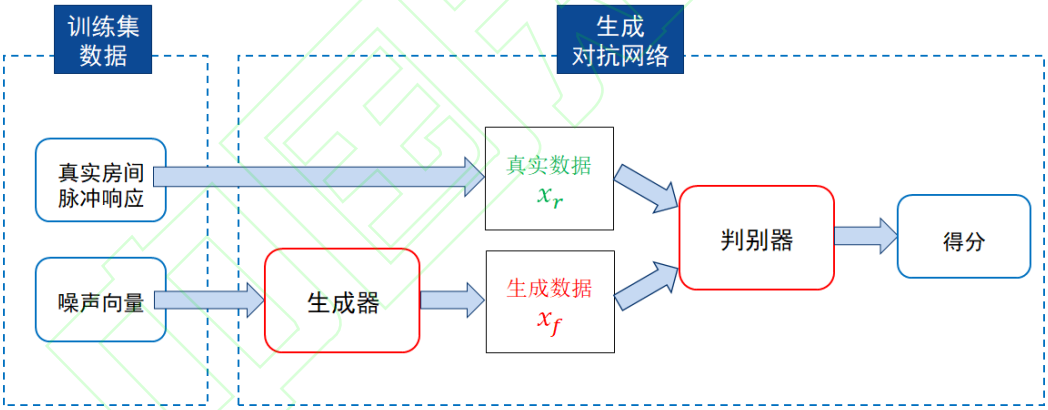


图 1 基于 GAN 的 RIR 模拟方法算法流程图

Fig. 1 RIR simulation framework based on GAN

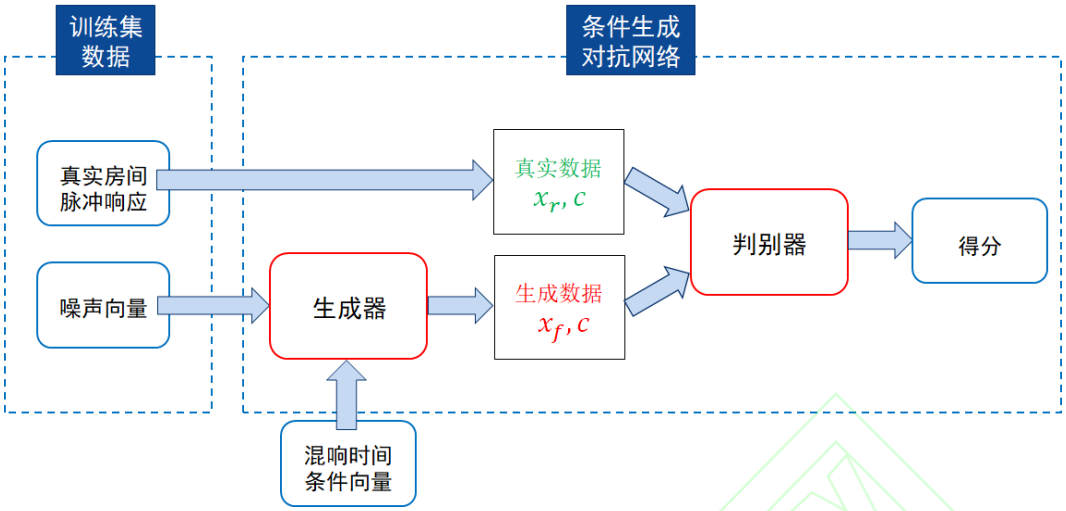


图 2 基于条件生成对抗网络的 RIR 生成模拟算法流程图

Fig. 2 RIR simulation framework based on conditional GAN

1.2 网络架构

为了保证模拟高质量的 RIR，本方法基于 WaveGAN 网络架构[19]进行拓展。 WaveGAN 网络架构针对声频信号的周期性等特点，通过叠加大步长卷积层增加了卷积神经网络的感受野。该网络架构被广泛应用于语声模拟、音效模拟等声频模拟任务。然而，WaveGAN 属于一般 GAN，无法额外输入条件向量对生成器进行限制。因此，本文基于 WaveGAN 网络架构，结合条件模拟 RIR 的任务，进行如下改进：首先，将条件向量加入生成器和判别器的输入中。对于每一个真实 RIR 训练样本，计算其混响时间作为条件向量。在生成器将噪声向量映射到生成样本的过程中，随机选取 0.3~1.5 s 的混响时间作为条件向量。其次，对生成器的输出进行归一化。由于 RIR 具有尺度不变性，因此对生成器的输出的 RIR 进行归一化使其最大值为 1，可以增加训练过程的稳定性。

1.3 损失函数

为了提高训练的稳定性，本文使用 WGAN 损失函数[20]对模型进行优化，使模型生成样本的分布逼近真实样本的分布。公式(1)~(2)表示了 WGAN 生成器的损失函数 $\mathcal{L}_G$ 和判别器的损失函数 $\mathcal{L}_D$：

$$\mathcal{L}_G = -\mathrm{E}_{x_f \sim \mathbf{Q}}\left[C\left(\boldsymbol{x}_f, \boldsymbol{c}\right)\right], \tag{1}$$

$$\mathcal{L}_D = -\mathrm{E}_{x_r \sim \mathbf{P}}\left[C\left(\boldsymbol{x}_r, \boldsymbol{c}\right)\right] + \mathrm{E}_{x_f \sim \mathbf{Q}}\left[C\left(\boldsymbol{x}_f, \boldsymbol{c}\right)\right], \tag{2}$$

其中，$\boldsymbol{x}_r$ 表示真实的数据向量，$\boldsymbol{x}_f$ 表示生成器生成的数据向量，$\boldsymbol{c}$ 表示条件向量，E 表示数学期望，C 表示判别器根据输入向量的真实程度从输入向量到得分的映射函数。输入向量越接近真实输入，得分越高；输入向量越偏离真实输入，得分越低。在生成器的损失函数中，生成器的目标是最大化生成的数据向量经过判别器后的得分，使生成数据的分布 $_{x_f \sim \mathbf{Q}}$ 更接近真实的数据分布 $_{x_r \sim \mathbf{P}}$。在判别器的损失函数中，判别器的目标是最大化生成数据的分布和真实数据的分布之间的距离，使生成数据和真实数据更容易被区分。在训练过程中，生成器和判别器交替训练，最终达到动态平衡。

1.4 训练数据与超参数

本文使用公开的 ACE 数据库[8]对模型进行训练，ACE 数据库中包含 7 个房间中的不同位置的单通道和多通道 RIR 数据，总共可以拆分为 700 条单通道 RIR 数据。该数据集录制设备、录制方法、原始数据信息记录良好且包含从 0.3~1.35 s 不同混响时间的房间，适用于进行模型训练。本文使用非线性拟合方法[21]计算 RIR 对应的混响时间，将混响时间标签和 RIR 数据输入模型进行训练。为了提

升网络的收敛性能，每个 RIR 数据在训练前进行幅度归一化的预处理。

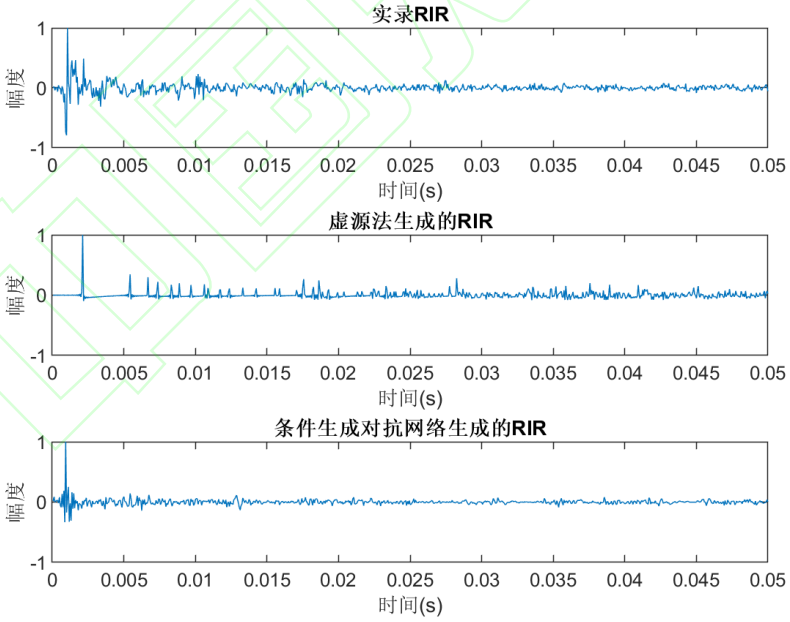
使用 Adam 优化器进行训练，其学习率为 0.0001，并使用学习率衰减策略。总迭代次数为 10×10^4 次，批次大小为 16。噪声向量的维度为 100。为了便于训练，RIR 的长度固定为 16384 点，对过长和过短的 RIR 分别进行尾部裁剪和尾部补零处理。

2 混响时间盲估计实验

为了评估条件生成对抗网络的效果，本文使用不同的 RIR 模拟方法模拟不同的 RIR 数据库，并且使用这些数据库构建不同的混响语声数据集分别训练不同的混响时间盲估计模型。模型训练完成后，在真实房间测试不同模型的性能，从而间接地评判不同的 RIR 模拟方法。

2.1 对比方法

本文的对比方法是 Schroder 模型和先前研究中常用的虚源法。分别模拟 5000 个混响时间范围为 0.3~1.5 s 的 RIR，使 RIR 数据库的混响时间标签均衡。此外，本文还采用公开的 2432 个真实单通道 RIR 对模型进行训练，其中包括 OpenAir 数据集[22]、REVERB 数据集[23]和 RWCP 数据集[24]。图 3 表示不同 RIR 的时域对比图，通过对比虚源法模拟的 RIR，可以发现使用条件生成对抗网络模拟的 RIR 在时域波形上与真实 RIR 更加接近。由于真实 RIR 数据集由于大部分房间的混响时间都在 0.8 s 以下，该真实 RIR 数据集的混响时间标签不均衡。图 4 和图 5 分别表示真实的 RIR 数据库和使用条件生成对抗网络模拟的 RIR 数据库的混响时间分布。



(a) 实录 RIR　　(b) 虚源法生成的 RIR　　(c) 条件生成对抗网络生成的 RIR

图 3 不同 RIR 的时域图

Fig. 3 Time-domain diagrams of different RIRs

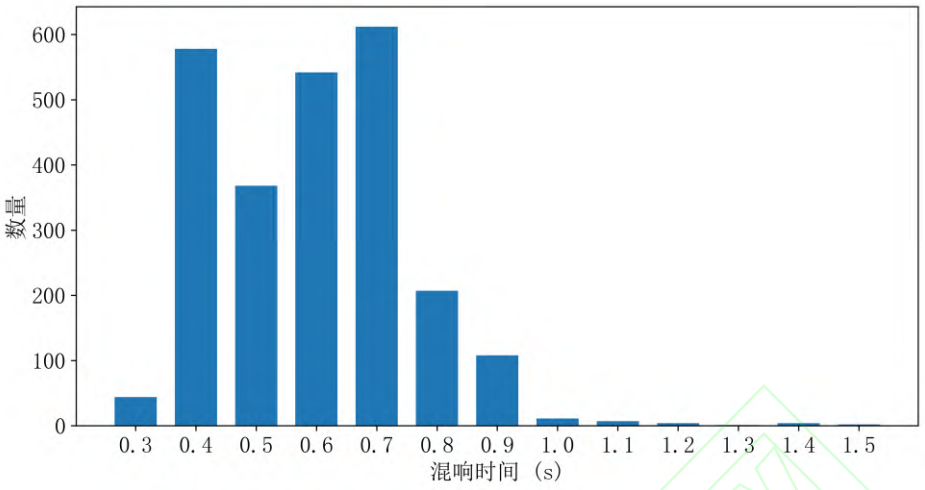


图 4 真实的 RIR 数据库的混响时间分布

Fig. 4 Reverberation time distribution of real RIRs

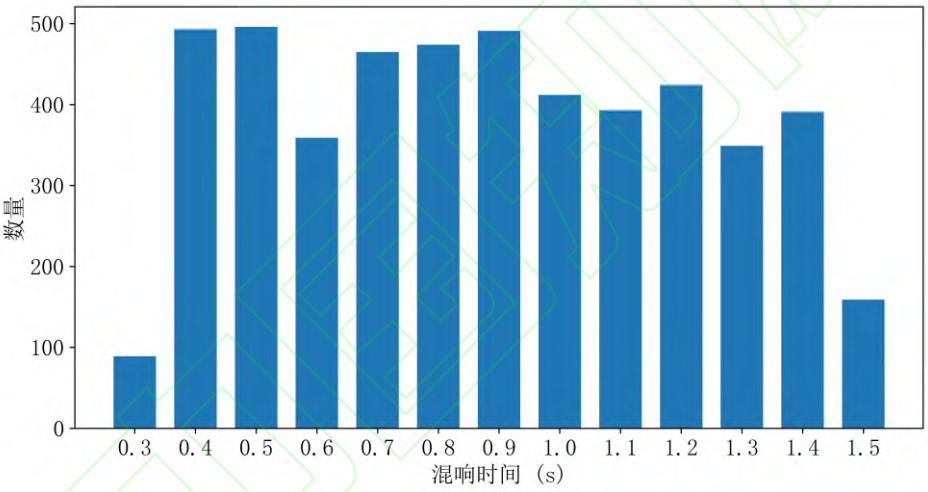


图 5 使用条件生成对抗网络生成的 RIR 数据库的混响时间分布

Fig. 5 Reverberation time distribution of RIRs generated by conditional GAN

2.2 混响语声训练集构建

基于房间系统的 LTI 假设，混响语声在时域可以表示为纯净语声和 RIR 的卷积。由于不可避免地存在噪声，因此在构建数据集使通常也考虑加性噪声。混响语声 $y(t)$ 的信号模型如公式(3)表示：

$$y(t)=s(t)*h(t)+n(t), \tag{3}$$

其中，$s(t)$ 表示纯净语声信号，$h(t)$ 表示 RIR，$n(t)$ 表示噪声信号，$*$ 表示卷积操作。

为了提高模型在噪声混响环境中的鲁棒性，本文在训练时考虑了 ACE 挑战赛中的 3 种加性噪声。信噪比从 0 dB、10 dB 和 20 dB 中随机选取，根据信噪比将噪声按不同比例加入混响语声中。语声被切分为每句 4 s，采样率为 16 kHz。分别使用 3 种 RIR 模拟方法构建的 3 个 RIR 数据库以及真实 RIR 数据库进行混响语声训练集模拟。每个语声训练集总共包含 3×10^4 句语声，总时长约为 33 h。

2.3 混响时间盲估计模型

本文测试使用不同训练集训练的混响时间盲估计模型的性能，从而间接判断模拟的 RIR 的质量。

本文选用的混响时间盲估计模型是文献[25]中提出的单步估计网络。该模型的计算复杂度适中，且网络性能优良。在预处理阶段，使用窗长为 20 ms、间隔为 10 ms 的汉宁窗对每句语声进行分帧和短时傅里叶变换。将短时傅里叶变换后的幅度谱输入混响时间盲估计模型，由模型输出每帧的混响时间估计。本文的估计模型在原有模型的基础上做了如下两点改进：(1) 使用非因果卷积，使得网络能够更有效利用上下文信息。(2) 由连续多帧输出的平均值决定最终估计结果，减少估计结果的方差。

2.4 估计性能评价指标

2.4.1 估计误差

估计误差定义为估计值与真实值的差值，可表示为

$$e = T_{60} - \hat{T}_{60}, \tag{4}$$

其中，T_{60} 表示混响时间真实值，$\hat{T}_{60}$ 表示混响时间估计值。对于 n 个样本，定义均方根误差(Root mean square error, RMSE)为

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_{60,i} - \hat{T}_{60,i})^2}, \tag{5}$$

其中，i 表示样本的下标。

2.4.2 皮尔森相关系数(Pearson correlation coefficient)

由于估计误差和 RMSE 没有被混响时间的真实值归一化，所以它们不能完全表征估计模型的性能。因此，使用皮尔森相关系数 ρ 作为另一个评价指标。估计结果越准确，ρ 就越接近 1。

对于 n 个样本，皮尔森相关系数 ρ 的公式可表示为

$$\rho = \frac{\sum_{i=1}^{n}(T_{60,i} - \bar{T}_{60})(\hat{T}_{60,i} - \bar{\hat{T}}_{60})}{\sqrt{\sum_{i=1}^{n}(T_{60,i} - \bar{T}_{60})^2}\sqrt{\sum_{i=1}^{n}(\hat{T}_{60,i} - \bar{\hat{T}}_{60})^2}}, \tag{6}$$

其中，$\bar{\hat{T}}_{60}$ 表示估计的混响时间的平均值，$\bar{T}_{60}$ 表示混响时间真实值的平均值。

3 实验与结果讨论

3.1 声学实验设置

在模拟训练集时，基于房间 LTI 系统假设，可以使用纯净语声和 RIR 卷积模拟混响语声。因此，大多数之前的研究使用这一方法模拟混响语声用于评估估计模型的性能。然而，声音在真实场景中传播的过程往往不符合 LTI 系统假设[1]。因此，本文在 4 个具有不同尺寸、声学性质和混响时间的房间中进行了现场声学实验和录制，以验证各个模型在真实环境下的实用性。4 个房间的尺寸和混响时间如表 1 所示。

表 1 真实房间尺寸和混响时间

Table 1 Sizes and reverberation times of realistic rooms

房间名称	长/m	宽/m	高/m	体积/m^3	混响时间/s
房间1	6.16	4.72	2.80	81.41	0.324
房间2	12.42	6.93	2.67	229.81	0.822
房间3	6.20	4.66	2.79	80.60	0.838
房间4	5.20	4.26	3.65	80.85	1.512

3.1.1 实验硬件

录声硬件包括笔记本电脑、声卡、恒流源、北京声传科技有限公司 CHZ-213+YG-201 型预极化 1.27 cm 传声器、两套 GENELEC 8030B 有源监听扬声器系统(一套用于播放语声信号，一套用于播放噪声信号)、声频线和电源。

3.1.2 录制流程

在房间中固定噪声源和信号源的位置，选取 5 个接收点作为传声器的位置并进行标记。对于每个接收点，进行如下操作：将从 TIMIT[26]数据库选取的长度为 5 min 的纯净语声作为语声信号；将 ACE challenge[8]数据集中的 3 种噪声，以及选自 NOISEX92 噪声数据集的粉色和白色噪声作为噪声信号，每种噪声持续时间为 1 min。使用最大长度序列(Maximum length sequence, MLS)[27]法和指数正弦扫频(Exponential sine sweep, ESS)[28]法测量 RIR，从而计算房间真实的混响时间。MLS 信号长度为 10.92 s，采样频率为 48 kHz。ESS 信号的频率范围为 20 Hz~20 kHz，持续时间为 20 s。通过比较两种不同方法的结果验证了测量的可重复性。由于 ESS 方法对扬声器非线性失真具有鲁棒性，最终采用 ESS 方法测得的 RIR 作为计算房间真实的混响时间依据。所有信号均使用 48 kHz 采样率和 32 位精度进行录制。

3.1.3 测试集构建

将录制的混响语声切割为长度为 4 s 的混响语声片段，并对每段语声随机添加噪声。信噪比从 0 dB、10 dB 和 20 dB 中随机选取，根据选取的信噪比将噪声按不同比例加入混响语声中。总共构建 3000 句带噪混响语声作为测试集。房间混响时间的真实值使用测量的 RIR 计算得出。为了验证使用 RIR 计算出的混响时间的可靠性，本实验还使用了中断声源法测量混响时间，这两种测量方法的平均误差在 0.02 s 内。

3.2 混响时间盲估计实验结果

使用不同混响语声训练集训练的混响时间盲估计模型在不同信噪比的真实测试集下的性能如表 2 所示(加粗表示每项中的最佳结果)。从表 2 中可以得出，模型的估计性能随着信噪比的增加而提升，表明噪声对估计性能有负面影响；使用 GAN 法训练的估计网络在 RMSE 指标上对不同信噪比下的场景均为性能最优，在皮尔森相关系数指标上在 0 dB 场景下最优；结果表明本方法模拟的 RIR 在训练估计模型时相较其他方法存在优势。

表 2 4 种方法训练的混响时间盲估计模型在不同信噪比下的估计性能

Table 2 Experimental results of four methods in real-world noisy reverberant scenarios

评价指标	RMSE/ms						ρ					
信噪比/dB)	0	5	10	15	20	平均	0	5	10	15	20	平均
真实数据	291	287	267	272	258	275	0.826	0.870	0.889	0.89	0.897	0.874
Schroder模型	225	215	224	224	222	222	0.888	**0.946**	0.938	0.950	0.948	0.934
虚源法	206	180	169	169	160	176	0.878	0.925	**0.942**	**0.954**	**0.961**	0.932
GAN	**197**	**165**	**155**	**146**	**139**	**160**	**0.910**	0.938	0.941	0.946	0.950	**0.937**

模型在 4 个房间中的估计误差箱线图如图 6 所示。从图 6 中可以看出，在低混响时间场景下(房间 1)，4 个模型的误差和方差与其他场景下相比最小，4 个模型均存在高估混响时间的趋势；在中等混响时间场景下(房间 2、房间 3)，真实数据训练的模型和方差较小，而所提方法训练的模型误差和偏差较大；在长混响时间场景下(房间 4)，所提方法训练的模型误差和方差最小，4 个模型均存在低估混响时间的趋势。

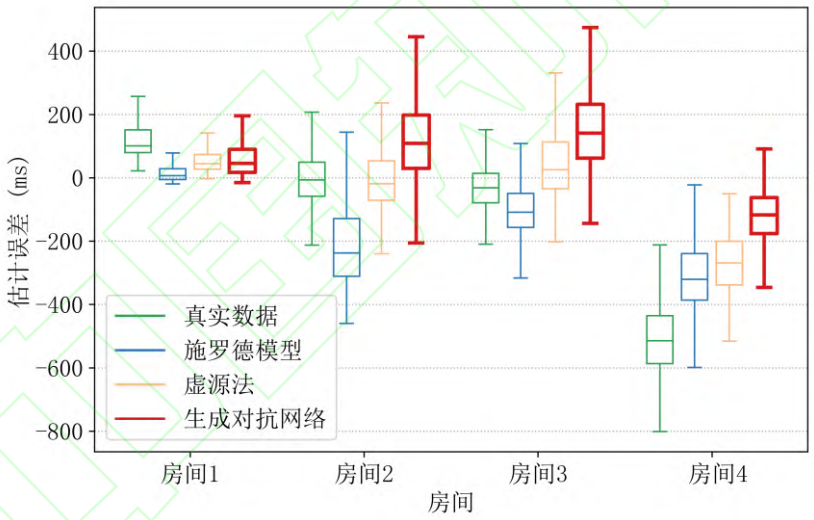


图 6 4 种方法训练的混响时间盲估计模型在不同房间中的估计误差箱线图。房间的尺寸与声学参数见表 1

Fig. 6 Estimation errors of four methods and baselines in different rooms. The details of the room configuration are shown in Table 1.

为了探究在不同混响时间下各个模型的估计性能的影响，将模型和混响时间作为自变量，RMSE 作为因变量进行双因素方差分析(Analysis of variance, ANOVA)。统计结果表明：不同模型[$F(3,11996) = 794.86$; $p < 0.01$]和不同混响时间[$F(2,11997) = 5596.41$; $P < 0.01$]对模型估计的 RMSE 均存在显著性影响；模型和混响时间存在显著的交互作用[$F(6,11988) = 1207.11$; $P < 0.01$]。表 3 展示了在不同混响时间下，两两模型之间的 Fisher LSD 事后检验。统计结果表明：随着混响时间的增加，模型的性能总体上有所下降。在短混响时间的情景下，Schroder 模型具有最小的估计误差和方差，除了虚源法和 GAN 方法之外，其他方法均存在显著的统计学差异；在中等混响时间的情景下，真实数据具有最小的估计误差和方差，所有方法均存在显著的统计学差异；在长混响时间的情景下，本文提出的 GAN

显著优于其他模型。

表 3 不同混响时间下，两两模型之间的 Fisher LSD 事后检验

Table 3 Results of Fisher LSD post hoc test between four methods in different reverberation time.

混响时间/s	方法	RMSE/ms 均值 ± 标准差	*P* 值		
			I	II	III
0.32	真实数据(I)	121 ±63			
	Schroder 模型(II)	32 ±71	**< 0.01**		
	虚源法(III)	67 ±87	**< 0.01**	**< 0.01**	
	GAN(IV)	67 ±76	**< 0.01**	**< 0.01**	0.980
0.82-0.83	真实数据(I)	63 ±49			
	Schroder 模型(II)	173 ±109	**< 0.01**		
	虚源法(III)	90 ±88	**< 0.01**	**< 0.01**	
	GAN(IV)	149 ±113	**< 0.01**	**< 0.01**	**< 0.01**
1.51	真实数据(I)	500 ±117			
	Schroder 模型(II)	308 ±109	**< 0.01**		
	虚源法(III)	269 ±92	**< 0.01**	**< 0.01**	
	GAN(IV)	127 ±82	**< 0.01**	**< 0.01**	**< 0.01**

从以上统计结果中可以看出，随着混响时间的增加，混响时间估计任务的难度更大，在不同的混响时间下不同的模型取得最优性能。Schroder 模型、真实数据、GAN 发分别在短、中、长混响时间下效果最佳；由于在真实数据集中，中等混响时间数据显著多于短混响时间和长混响时间数据，使用真实数据训练的模型的估计集中在中等混响时间，因而模型在中等混响时间下性能最优异。

为了探究使用本方法所生成的 RIR 对真实数据增广的效果，使用 GAN 模拟混响时间大于 0.8 s 的 RIR 对真实数据集进行增广，使真实数据集的长混响数据与中短混响数据数量接近。使用增广后的混合 RIR 数据集作为训练集训练盲混响估计模型，并在相同的测试集下测试盲混响估计模型的性能。混合数据、真实数据与 GAN 模拟数据训练的混响时间盲估计模型在不同信噪比的真实测试集下的性能如表 4 所示。可以看出，混合模型在高信噪比条件(10 dB、15 dB、20 dB)下具有最小的均方根误差和最大的皮尔森相关系数，表明了混合模型在高信噪比下具有优势，而基于 GAN 的模型在低信噪比下更具有优势。为了探究 3 个模型在具有不同混响时间的房间中的性能，模型在 4 个房间中的估计误差箱线图如图 7 所示。

表 4 混合数据、真实数据与 GAN 模拟数据训练的混响时间盲估计模型在不同信噪比下的估计性能

Table 4 Experimental results of three estimation models trained by mix data, real data and simulated data in noisy reverberant scenarios

评价指标	RMSE/ms						*P*					
信噪比/dB	0	5	10	15	20	Avg.	0	5	10	15	20	Avg.
真实数据	291	287	267	272	258	275	0.826	0.870	0.889	0.89	0.897	0.874
混合数据	281	200	**148**	**138**	**123**	185	0.827	0.917	**0.953**	**0.963**	**0.972**	0.918
GAN	**197**	**165**	155	146	139	**160**	**0.910**	**0.938**	0.941	0.946	0.950	**0.937**

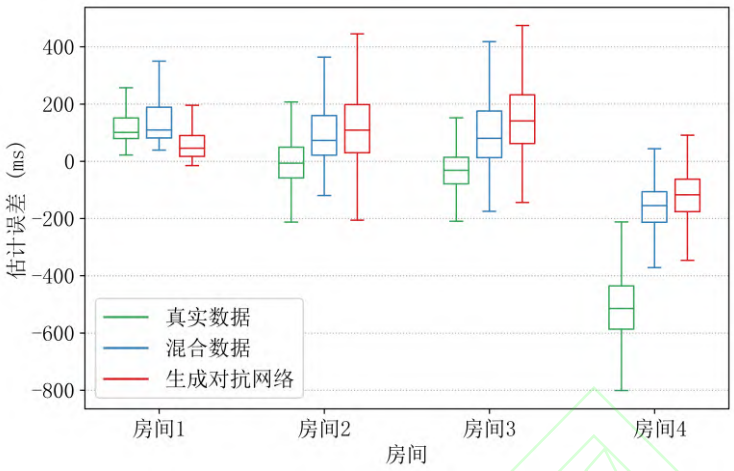


图 7 3 种方法训练的混响时间盲估计模型在不同房间中的估计误差箱线图。房间的尺寸与声学参数见表 1

Fig. 7 Estimation errors of three methods and baselines in different rooms. The details of the room configuration are shown in Table 1.

从图 7 中可以看出，使用真实数据训练的估计模型由于缺少长混响数据，在长混响情况下(房间 4)性能不佳；而通过 GAN 对真实数据进行增广后的混合数据训练的估计模型在长混响情况下相较未增广时性能大幅度提升。同时，由于混合数据中存在真实数据，混合数据在中等混响情况下(房间 2、房间 3)性能比全部使用 GAN 模拟的方法具有更小的偏差和方差；全部使用 GAN 模拟的数据在短混响(房间 1)和长混响情况下具有更小的偏差和方差。通过在不同信噪比和房间下的性能对比，可以发现在高信噪比和中等混响条件下，使用混合数据进行训练的网络相比全部使用 GAN 模拟的网络具有更优的性能。在各种信噪比和长混响条件下，使用混合数据进行训练的网络相比使用真实数据的网络有明显的性能提升。

4 结论

在构建混响语声数据集时，由于真实的 RIR 缺乏长混响数据，且模拟的 RIR 与真实存在差距，因而导致数据驱动的混响时间盲估计模型性能下降。本文提出基于条件生成对抗网络的 RIR 模拟方法，使网络能够根据输入的混响时间模拟更真实的 RIR。实验结果表明，采用本方法模拟的 RIR 训练的盲混响时间估计模型在不同信噪比场景下均具有最小的均方根估计误差，且在长混响场景下显著优于其他模型。该方法可以用于 RIR 增广，以扩展混响语声数据集。

参 考 文 献

[1] Kuttruff H. Room acoustics[M]. 第 6 版. Boca Raton: CRC Press, 2016.
[2] Bradley J S. Speech intelligibility studies in classrooms[J]. The Journal of the Acoustical Society of America, , 1986, 80(3): 846–854.
[3] Schroeder M R. New method of measuring reverberation time[J]. The Journal of the Acoustical Society of America, , 1965, 37(6): 1187–1188.
[4] Cox T J, Li F, Darlington P. Extracting room reverberation time from speech using artificial neural networks[J]. Journal of the Audio Engineering Society, 2001, 49(4): 219–230.
[5] Jones D L, Wheeler B C, O'Brien Jr W D, et al. Blind estimation of reverberation time[J]. The Journal of the Acoustical Society of America, 2003, 114(5): 2877–2892.
[6] Wen J Y, Habets E A, Naylor P A. Blind estimation of reverberation time based on the distribution of signal decay

rates[C]//2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: 329–332.
[7] de Prego T M, de Lima A A, Zambrano-López R, et al. Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition[C]//2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), 2015: 1–5.
[8] Eaton J, Gaubitch N D, Moore A H, et al. The ACE challenge—Corpus description and performance evaluation[C]//2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)., 2015: 1–5.
[9] Xiong F, Goetze S, Meyer B T. Joint estimation of reverberation time and direct-to-reverberation ratio from speech using auditory-inspired features[J]. arXiv Preprint, arXiv: 1510.04620, 2015.
[10] Gamper H, Tashev I J. Blind reverberation time estimation using a convolutional neural network[C]//2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), 2018: 136–140.
[11] Shuku T, Ishihara K. The analysis of the acoustic field in irregularly shaped rooms by the finite element method[J]. Journal of Sound and Vibration, 1973, 29(1): 67-IN1.
[12] Kirkup S. The boundary element method in acoustics: a survey[J]. Applied Sciences, Multidisciplinary Digital Publishing Institute, 2019, 9(8): 1642.
[13] Allen J B, Berkley D A. Image method for efficiently simulating small-room acoustics[J]. The Journal of the Acoustical Society of America, 1979, 65(4): 943–950.
[14] Krokstad A, Strom S, Sørsdal S. Calculating the acoustical room response by the use of a ray tracing technique[J]. Journal of Sound and Vibration, 1968, 8(1): 118–125.
[15] Ratnarajah A, Tang Z, Manocha D. IR-GAN: room impulse response generator for far-field speech recognition[C]. Interspeech 2021, 2021.
[16] Ratnarajah A, Zhang S X, Yu M, et al. FAST-RIR: Fast neural diffuse room impulse response generator[J]. arXiv Preprint, arXiv: 2110.04057, 2021.
[17] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. Advances in Neural Information Processing Systems, 2014: 27.
[18] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv Preprint, arXiv: 1411.1784, 2014.
[19] Donahue C, McAuley J, Puckette M. Adversarial audio synthesis[J]. arXiv Preprint, arXiv: 1802.04208, 2018.
[20] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International Conference on Machine Learning, 2017: 214–223.
[21] Antsalo P, Makivirta A, Valimaki V, et al. Estimation of modal decay parameters from noisy response measurements[C]//Audio Engineering Society Convention 110, 2001.
[22] Murphy D T, Shelley S. Openair: an interactive auralization web resource and database[C]//Audio Engineering Society Convention 129, 2010.
[23] Kinoshita K, Delcroix M, Yoshioka T, et al. The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech[C]//2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013: 1–4.
[24] Nakamura S, Hiyane K, Asano F, et al. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition[J]. 2000.
[25] Zheng K, Zheng C, Sang J, et al. Noise-robust blind reverberation time estimation using noise-aware time-frequency masking[J]. arXiv Preprint, arXiv: 2112.04726, 2021.
[26] Garofolo J S. Timit acoustic phonetic continuous speech corpus[J]. Linguistic Data Consortium, 1993.
[27] Schroeder M R. Integrated-impulse method measuring sound decay without using impulses[J]. The Journal of the Acoustical Society of America, , 1979, 66(2): 497–500.
[28] Farina A. Simultaneous measurement of impulse response and distortion with a swept-sine technique[C]//Audio Engineering Society Convention 108, 2000.