一种用于语音去混响的复合式 T60 回归与分类方法

Yuying Li，*IEEE* 学生会员，Yuchen Liu，*IEEE* 学生会员，以及 Donald S. Williamson，*IEEE* 高级会员

arXiv:2302.04932v1 [eess.AS] 9 Feb 2023

摘要——**去混响通常直接作用于混响音频信号，而不考虑声学环境。混响时间，T_{60}，是反映混响可能如何影响信号的一项重要声学因素。在这项工作中，我们提出在执行去混响的同时，利用环境中的关键信息。具体而言，我们开发了一种联合学习方法，采用复合式 T_{60} 模块和独立的去混响模块，同时进行混响时间估计和去混响。微调期间，混响时间模块向去混响模块提供关键特征。我们在模拟环境和真实环境中评估了所提出的方法，并与多种方法进行了比较。结果表明，这种复合式框架能够提升环境中的性能。**

索引词—**去混响、混响时间、深度神经网络、联合学习**

I 引言

混响现象在日常环境中普遍存在，这是由于房间内家具、墙壁和地板等众多表面会反射声音。这会使听者听到直达语音信号与反射声的混合。混响的影响常常会使语音在时间和频率上变得模糊，从而对听力受损人士产生负面影响 [1], [2], ，因为混响会降低感知质量和可懂度。这也给各种语音应用带来了挑战，包括自动语音识别（ASR） [3], [4], 、说话人识别[5], [6] 和说话人定位 [7], [8], 等。

目前的单耳方法通常使用深度神经网络（DNN）去除混响。Han 等人 [9], 提出了一种频谱映射方法，利用全连接深度神经网络将含噪混响信号映射到时频（T-F）域中的无混响语音信号。该方法包含一个后处理阶段，通过迭代相位重建重新合成估计的时域信号。Xiong 等人 [10] 使用多层感知机（MLP），根据伽博滤波器组的特征估计混响时间（T60）。深度神经网络估计复数理想比率掩码（cIRM） [11],，它在虚部域和实部域中处理幅度和相位响应。该方法可同时处理含噪和混响条件。然而，使用深度神经网络存在局限性，因为它们无法捕捉长期上下文信息。为克服这一局限，Santos 等人 提出了一种去混响方法，使用循环神经网络（RNN） [12] 捕捉长期上下文信息，并采用二维卷积编码器提取局部上下文特征。Zhao 等人 [13] 提出了另一种采用长短期记忆网络（LSTM）的循环神经网络模型，用于预测晚期反射，并将其从混响信号中减去，以估计混响的直达分量和早期分量。Zhao等人 随后提出了一种采用带自注意力模块的时间卷积网络（TCN）的去混响模型 [14]。该方法使用自注意力从输入中提取动态特征，并使用时间卷积网络学习从混响语音到无混响语音的非线性映射。上述所有方法均在时频域中运行。

传统信号处理方法也已得到发展。这些方法可分为两类：谱减法和逆滤波。Lebart 等人提出了一种谱减法方法 [15], 该方法利用混响时间和音素的先验知识，通过消除音素能量中的扩散效应来去除混响。具体而言，该方法估计混响的功率谱密度（PSD），再从混响信号中减去估计得到的 PSD 的平方根，从而估计出去混响信号的频谱。Yoshioka等人提出了一种无需声学条件先验知识[16]的通用子带域多通道线性预测方法（也称为加权预测误差-WPE）。Tomohiro等人 [17] 使用延迟线性预测（DLP）模型，在无需房间脉冲响应（RIR）先验知识的情况下消除晚期混响。WPE 估计一个用于预测混响尾部的滤波器，并将其从混响信号中减去，以获得最大似然估计。该方法已应用于许多领域 [18], [19]。

上述所有方法均直接处理混响信号，并且要么（1）不考虑房间和信号的声学及上下文信息，要么（2）假设已知这些信息而不进行估计。具体而言，这些方法没有利用或估计混响时间 T_{60} 的信息，而混响时间是混响拖尾效应的一个重要指标。例如，可以对这类信息进行估计；Bryan等人 [20] 使用卷积神经网络（CNN）将输入下采样为单个

Y. Li 任职于印第安纳大学 Luddy 信息学、计算与工程学院智能系统工程系，地址为美国印第安纳州布卢明顿 47408（电子邮箱：liyuy@iu.edu）。Y. Liu 任职于印第安纳大学 Luddy 信息学、计算与工程学院计算机科学系，地址为美国印第安纳州布卢明顿 47408（电子邮箱：liu477@iu.edu）。D. S. Williamson 任职于俄亥俄州立大学计算机科学与工程系，地址为美国俄亥俄州哥伦布 43065（电子邮箱：williamson.413@osu.edu）。本文已提交 IEEE，供其考虑发表。版权可能在未另行通知的情况下转让，届时本版本可能不再 доступ。

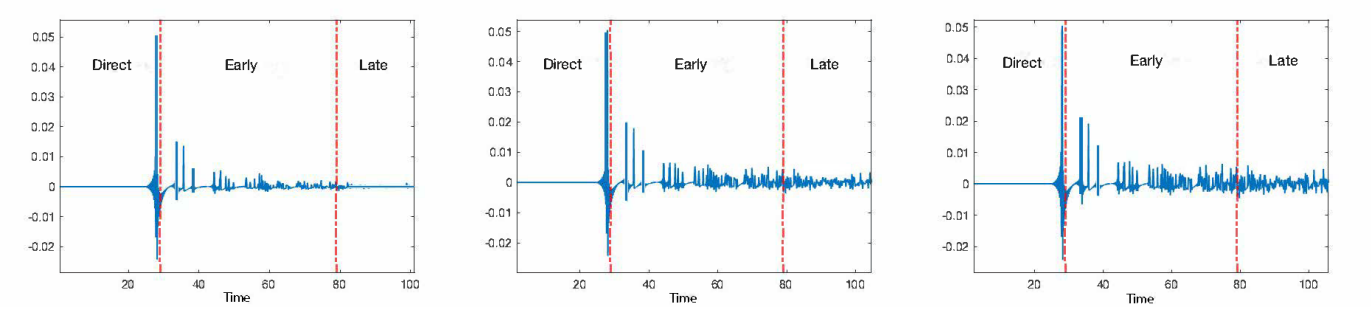


图1。（在线彩色版）将房间脉冲响应分解为直达路径、早期反射和晚期混响。从左至右：$T_{60} = 0.3$s、0.6s、0.9s。

T_{60} 值。鉴于房间环境对去混响至关重要，Wu等人 [21] 研究了不同的上下文信息如何影响混响抑制。该方法使用混响时间感知深度神经网络，通过在特征提取过程中恰当选择帧移和上下文窗口大小来估计混响时间。随后，它将对数功率谱图输入深度神经网络进行去混响。自注意力模块无需根据选定的帧移和窗长手动生成不同的上下文信息，而是自动学习不同的表征。然而，该方法需要针对不同的 T_{60}s手动选择上下文信息，并且需要可靠的 T_{60} 估计器。这种混响时间感知深度神经网络已被用作自动语音识别的前端处理 [22]。Wang等人 [23]提出了另一种时间上下文感知方法。其主要去混响模型利用具有时间感知能力的上下文帧来预测去混响频谱，同时联合优化混响时间。这种环境感知网络联合执行混响时间估计和去混响，但该方法并非总能良好泛化，在真实环境中的性能仍需提升。这些方法表明，利用混响时间进行优化有助于去混响。

本文提出一种用于语音去混响的联合学习方法，可准确估计 T_{60} 和晚期反射。早期反射有助于提高语音可懂度，因此我们决定去除晚期反射 [24],[25]。我们分别训练了一个 T_{60} 估计器，其模型与我们初步研究 [26], 中的模型相匹配；此外，还训练了一个去混响网络 [13]。同时，将来自 T_{60} 估计器的额外特征提供给去混响模块。新的特征连接了这两个网络，随后进行微调，以生成最终的去混响结果。我们确实进行了实验，以确定在 [26] 中最适用于联合方法的损失函数。所采用的去混响模型与[13]中的模型有四个主要区别：1）窗长和快速傅里叶变换的尺寸；2）我们堆叠了三层 LSTM 层，而非两层；3）训练时不再采用权重丢弃方法，改为使用随机失活，并在规模更大、分布更均衡的数据集上训练；最后，4）我们还采用了不同的输入特征。

本文其余部分的组织如下。第II节介绍背景信息。第III节详细描述算法。第IV节解释实验。实验的评估及结果见第V节。第VI节和第VII节讨论相关问题并给出结论。

II. 背景

混响信号 $x(t)$ 可建模为无混响语音信号 $s(t)$ 与房间脉冲响应 $h(t)$ [27], $x(t) = s(t) * h(t)$ 的卷积，其中 $*$ 表示卷积，t 表示时间索引。房间脉冲响应可分解为三个部分：直达房间脉冲响应、早期房间脉冲响应和晚期房间脉冲响应：$h(t) = h_d(t) + h_e(t) + h_l(t)$，其中 $h_d(t)$、$h_e(t)$、$h_l(t)$ 分别对应直达分量、早期分量和晚期分量。分解示例见图1。直达分量 $h_d(t)$ 从房间脉冲响应的起始处开始，约在第一个峰值后 1 ms 结束；而早期分量则延伸至直达声后约 50 ms [24]。晚期分量从 $h_e(t)$ 的末尾开始，一直延续到房间脉冲响应的末尾。早期和晚期反射之间的过渡时间应在 50 至 150 ms 之间 [28], ，该范围内的值常被采用 [25], [29], [30]。本文将早期和晚期反射之间的过渡时间固定为 50 ms。每个房间脉冲响应分量的长度都与 $h(t)$ 相同，但在上述区间之外的值均为零。

根据分配律，混响信号可建模为三个房间脉冲响应分量与无混响语音信号卷积后的总和，分别得到直达声、$x_d(t)$、早期反射、$x_e(t)$ 和晚期反射、$x_l(t)$，如下式所示。

$$\begin{aligned} x(t) &= s(t) * (h_d(t) + h_e(t) + h_l(t)) \\ &= s(t) * h_d(t) + s(t) * h_e(t) + s(t) * h_l(t) \\ &= x_d(t) + x_e(t) + x_l(t) \\ &= x_{de}(t) + x_l(t) \end{aligned} \tag{1}$$

$x_{de}(t)$ 表示直接早期分量，其定义为直达声与早期反射之和。

混响时间 T_{60} 是房间内的声音衰减 60 dB 所需的时间。它可以用萨宾公式 [31]建模：

$$T_{60} = 0.16V/\alpha S \tag{2}$$

其中，V 表示房间的体积，S 表示其表面的面积，α 表示吸声系数。较长的混响时间意味着会产生更多反射声，从而导致时间和频率上的模糊更严重。因此，

混响时间与房间脉冲响应之间存在函数关系。图1也展示了这一点，其中给出了使用不同混响时间生成的房间脉冲响应。

III. 算法描述

我们提出使用深度神经网络联合估计混响时间 T_{60}，以及混响语音的直达早期分量。我们所提出的方法包含三个阶段：(1) T_{60} 估计；(2) 直达早期分量估计；然后 (3) 联合执行 T_{60} 估计和去混响。图2展示了我们的方法。

A. T_{60} 估计与分类

我们将 T_{60} 估计视为一个多任务问题，同时执行：(a) 直接从混响信号估计混响时间；(b) 对其进行分类。我们采用这种方法，是因为多任务学习已被证明有利于语音增强 [32], [33], [34], 以及感知质量指标 [35], [36]的估计。

1) 特征： 我们对混响信号进行短时傅里叶变换（ STFT），然后将对数幅度响应与相位的 sin 和 cos在所有时频单元上拼接起来， θ。这样做是因为以往研究表明，加入相位信息有助于提升性能 [37]。我们对拼接后的特征进行归一化，使其在每个频率单元上、跨所有训练样本的均值为零、方差为一。

2) 网络架构：归一化输入被送入特征提取模块，如图 2(B) 所示。该模块由六个卷积层（ Conv ）组成，每个卷积层在训练期间均采用批归一化（BN）。每两个卷积层后使用最大池化对潜在表示进行下采样；但对于最后两层，最大池化则插入到最后两个卷积层之间。该架构在我们的初步实验中表现良好 [26] ，当时我们评估了不同的代价函数；此外，该架构在相关语音处理任务中也表现良好 [38], [39]。

图 2(C) 展示了联合执行 T_{60} 回归和分类的网络架构。由特征提取器生成的潜在表示用作复合 T_{60} 估计阶段的输入。复合T_{60} 估计阶段包含两个分支，两个分支接收相同的输入：(1) 仅回归分支（例如，左侧分支），以及 (2) 分类与基于分类的回归组合分支（例如，右侧分支）。 [40], [41]中提出了类似的架构。对于仅回归分支，共享输入被送入带有线性整流（ReLU）激活函数的卷积层。接下来是平均池化层、带批归一化和带泄漏的 ReLU 激活的全连接层，以及用于生成估计 T_{60}值的全连接层。图 2(C)（右）展示了复合估计方法中 T_{60}分类部分的网络细节。该部分包含两个作为隐藏层的全连接层，每个全连接层后接一个批归一化层。两个全连接层均使用 Leaky ReLU 激活函数。隐藏层之后，线性层用作输出层，其输出进一步分为两个子部分： T_{60} 分类和 T_{60}基于分类的回归。前一个子部分使用 softmax 激活函数，并由图 2(C) 中的分类输出表示。后一个子部分基于以下回归损失函数：

$$C_{Reg_T_{60}} = \sum_{i=1}^{H} C_{\text{out}}^{i} \times T_i \tag{3}$$

其中， $\times$ 表示逐元素乘法，H 表示类别数， T_i 表示第 i 类的 T_{60} 时间， C_{out}^{i} 表示第 i 类的估计概率。

B. 直达早期分量估计

图 3 展示了我们所提出的去混响过程。已有研究表明，可以在时频域中通过谱减法去混响，以去除晚期反射声 [42], [13], ，因此我们选择在深度神经网络估计之后进行谱减法。

1) 特征和训练目标：我们首先计算短时傅里叶变换，并将立方根压缩幅度用作输入特征，将直达早期分量的立方根幅度用作去混响模块的训练目标，这与 [13] 类似。

2) 网络架构：将压缩幅度特征输入单向 LSTM，因为它能够学习长期依赖关系。图 3 中的去混响模块展示了该 LSTM 网络。它由 3 层 LSTM 层组成。随后接入一个全连接层，将隐藏状态映射为输出，在此我们预测压缩幅度响应中的晚期反射。然后，我们通过从相应混响信号的立方根幅度中减去估计的晚期反射来执行谱减法。由此得到直达早期分量的估计值。

C. 联合网络

图 2 和图 3 所示的两个网络最初分别进行训练。最后，我们提出将这两个网络结合起来并进行微调训练，形成一个新的联合网络，如图 3 所示。在联合训练期间， T_{60} 分类模块倒数第二层的输出被用作去混响模块输入的一部分，并与压缩幅度拼接。由此形成一种更新的输入特征，同时包含音频特征和混响时间特征。我们尝试了连接这两个模块的不同输入：1）仅回归分支的标量输出，2）分类分支的独热向量，以及 3）倒数第二层的输出。我们使用 T_{60} 估计器倒数第二层的输出，因为它更有可能包含与混响时间 [43] 相关的信息，并且在实证测试中表现更好。

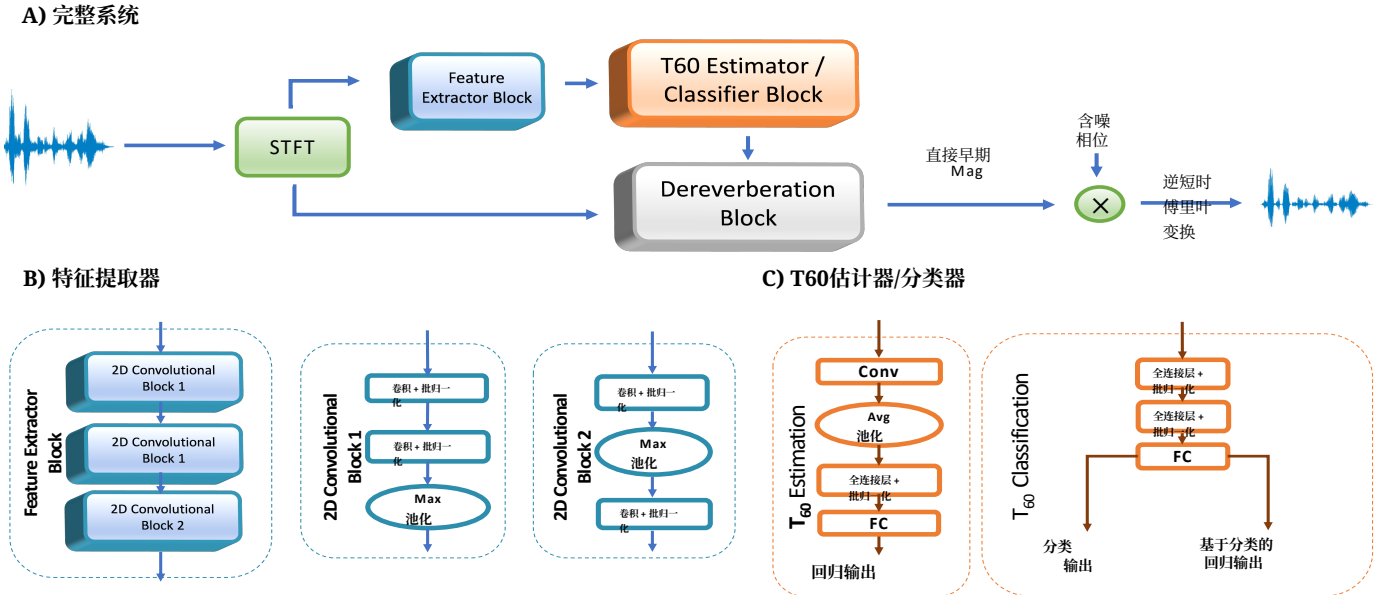


图2。（在线彩色版）所提出的联合网络如图所示，其中（A）展示了完整方法，（B）展示了特征提取模块，（C）展示了 T_{60} 分类器和估计器的网络架构。去混响模块如图3所示。

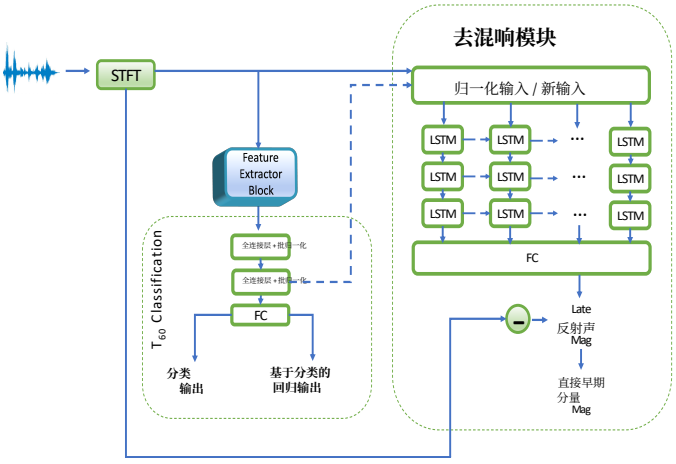


图3。所提出的去混响模块示意图。虚线箭头仅在微调期间出现。

D. 代价函数

用于对 T_{60} 分支进行预训练的所提出的代价函数如下所示，其基于我们先前的工作 [26]:

$$L^A = \beta \times (\alpha \times L_{\text{cls}} + (1-\alpha) \times L_{\text{creg}}) + (1-\beta) \times L_{\text{reg}} - |\rho_{\text{reg}}| - |\eta_{\text{reg}}| - |\rho_{\text{cls}}| - |\eta_{\text{cls}}| \quad (4)$$

其中， L^A 是分类分支和回归分支中交叉熵损失项 L_{cls}与均方误差（MSE）损失项 L_{creg} 和 L_{reg}的组合损失。β 控制两个分支损失之间的权重。α 平衡分类分支中交叉熵损失与基于分类的回归任务的均方误差。L_{creg} 使估计的基于分类的 T_{60} 与真实值 T_{60}之间的均方误差最小化。该组合损失还包含基于皮尔逊相关系数（PCC，ρ）和斯皮尔曼等级相关系数（SRCC，η）的项，这两个系数通常用作评估指标。$|\cdot|$表示绝对值。

联合网络所提出的代价函数为:

$$L^B_{\text{joint}} = \gamma * [\alpha \times L_{cls} + (1-\alpha) \times L_{\text{creg}}] + (1-\gamma) L_{\text{derev}} \quad (5)$$

其中，$\gamma \in [0,1]$ 用于控制网络两部分的权重，α 对 T_{60} 估计分支进行加权，而 L_{derev}表示在去混响分支中计算的均方误差损失。根据我们此前关于 T_{60} 估计的研究，分类分支的表现优于回归分支，因此我们使用分类分支的损失和去混响损失，在微调期间更新联合网络。

IV. 实验

A. 数据

我们使用 TIMIT 语料库 [44], 进行训练，该语料库已用于以往的研究 [45], [46], , 其中包含来自美国八个地区的 630 名英语母语者。我们随机选取 5000、500 和 500 条信号，分别构建模拟环境中的训练、验证和测试数据集。将全部 6000 条信号下采样至 8 千赫。我们使用镜像法 [47]模拟 14 种不同房间尺寸的房间脉冲响应。各房间的尺寸列于表 I。房间 1 至 10 用于生成训练集和开发集，房间 11 至 14 用于生成测试集。为使各种条件下的直达声与混响声比（DRR）保持稳定，所有情况下麦克风与扬声器之间的距离均设为 1 m。我们选取 13 种混响时间：从 0.3 s 到 1.5 s，间隔为 0.1 s，以评估不同程度的混响。我们使用全部 13 种混响时间对 T_{60} 模块进行预训练，同时采用以往研究设置中的三种混响时间（0.3 s、0.6 s 和 0.9 s）对去混响模块进行预训练 [9], [13]。对联合网络进行微调时，我们使用相同的三种混响时间。增加用于训练 T_{60} 模块的混响时间

表 I 矩形模拟房间脉冲响应的房间尺寸

组	Name	尺寸（米）
已见房间	房间1	9 × 8 × 7
	房间2	10 × 7 × 3
	房间3	6 × 6 × 10
	房间4	8 × 10 × 4
	房间5	7 × 7 × 8
	房间6	7 × 9 × 5
	房间7	8 × 8 × 10
	房间8	10 × 10 × 8
	房间 9	8 × 8 × 6
	房间 10	7 × 8 × 6
未见房间	房间 11	9 × 9 × 10
	房间 12	9 × 7 × 9
	房间 13	9 × 10 × 5
	房间 14	10 × 10 × 7

表 II 真实房间脉冲响应的房间特性：ACE 语料库

Name	尺寸（米）	T_{60}（秒）
办公室 1	4.8 × 3.3 × 3.0	0.34
办公室 2	5.1 × 3.2 × 2.9	0.39
会议室 1	6.6 × 4.7 × 3.0	0.44
会议室 2	10.3 × 9.2 × 2.6	0.37
讲堂 1	6.9 × 9.7 × 3.0	0.64
讲堂 2	13.4 × 9.2 × 2.9	1.25
大楼大厅	5.1 × 4.5 × 3.2	0.65

有助于提升泛化能力。我们针对前 10 种房间设置中的每种 T_{60} 模拟 500 个不同的房间脉冲响应，用于生成训练数据集。另外，针对每种 T_{60} 模拟 50 个房间脉冲响应用于验证。因此，验证集中每种 T_{60} 总计有 10 × 50 =500 个房间脉冲响应。对于测试房间，每种 T_{60} 和每种房间设置均使用 500 个房间脉冲响应。因此，测试集中每种T_{60} 有 4 × 500 = 2000 个房间脉冲响应。我们将每个房间脉冲响应与每种 T_{60}对应的一条独特语音信号进行卷积。因此，在 T_{60} 估计的预训练阶段，训练集中有 5000 × 13 = 65000 条混响信号，验证集中有 500 × 13 = 6500 条混响信号。在去混响模块的预训练阶段和联合学习阶段，训练集中有5000 × 3 = 15000 条混响信号，验证集中有500 × 3 = 1500 条混响信号，测试集中有2000× 3 = 6000。我们还使用 Pyroomacoustics [48] 镜像声源模型/射线追踪（ISM/RT）模拟器 [49], [50]生成未见过的非矩形房间。我们模拟了 3 个 L形房间，其尺寸分别为：8.5 m × 3m × 6 m 和 2 m × 4 m × 6 m；10 m × 5 m × 10 m 和 3 m × 1.5 m × 10m；以及 7 m × 4 m × 10 m 和 1.5 m × 2 m× 10 m。声源与扬声器之间的距离约为 1.88 m。每个房间模拟 200 个房间脉冲响应，共计 $200 \times 3 = 600$ 个房间脉冲响应。我们将每个房间脉冲响应与一条来自 TIMIT 测试集的信号进行卷积，共得到 600 条信号。 T_{60}的范围为 0.5 s 至 1 s。我们对房间脉冲响应进行零填充，使其长度与数据集中最长的房间脉冲响应相匹配 [26], ，并在卷积前将所有干净信号截取为 6 秒。短时傅里叶变换采用 480 个采样点的汉明窗、512 点快速傅里叶变换，以及连续帧之间 75% 的重叠率计算。特征的维度为 771×442。我们使用 ACE 挑战赛 [51]评估真实环境。

以及 BUT Speech a◯FIT Reverb 语料库 [52]。ACE 使用在真实环境中采集的房间脉冲响应，并通过将房间脉冲响应与干净语音进行卷积生成混响信号。BUT语料库包含来自 LibriSpeech语料库的重新传输信号 [53]。这两个语料库将展示我们的方法在未见过的真实环境中的泛化能力。ACE 数据集（表 II）包含七种设置，其中麦克风与扬声器之间的距离较远，混响时间范围为 0.34 秒至 1.25 秒。BUT语料库包含五种房间设置，其尺寸分别为：17.2 米 × 22.8 米 × 6.9 米、4.6 米× 6.9 米 × 3.1 米、7.5 米 × 4.6 米 × 3.1 米、6.2 米 × 2.6 米 ×14.2 米、10.7 米 × 6.9 米 × 2.6 米，混响时间范围为 0.61 秒至 1.85 秒。对于 BUT语料库，麦克风与扬声器之间的平均距离为 1.41 米至 7.93 米，因此模拟数据集和真实环境数据集中的信号都与 1 米距离下的信号有所不同。ACE语料库中的平均距离为 1.35 米至 2.14 米。该测试集中总共有 $1022 + 2620 \times 5 = 14122$ 个混响信号。

B. 设置

在图2(B)中，前两个卷积层使用16个卷积核滤波器，中间两个卷积层使用32个卷积核滤波器，最后两个卷积层使用64个卷积核滤波器。所有卷积核大小均设为 3 × 3，所有最大池化层的卷积核大小设为 2 × 2 。图2(C)中的平均池化层使用大小为3 × 3 的卷积核。在 T_{60}分类模块中使用了三个全连接层。前两个全连接层后接一个批归一化层，并使用斜率设为0.1的Leaky ReLU激活函数。最后一个全连接层也使用ReLU。

去混响模块的输入特征维度为 257×442，即压缩幅度的维度。在微调阶段，该维度与 T_{60} 特征向量一同使用。对于去混响模块中的长短期记忆网络，隐藏层大小设为512，丢弃率设为0.5，以避免过拟合。我们将 β 设为0.9，并将 α 设为0.1（例如，式(4)中的L^A），因为这些值优于其他选项。

所有预训练模型均使用批量大小 50 和 0.001 的学习率。不过，两项任务采用了不同的优化器： T_{60} 估计使用 RMSprop优化器，去混响则使用 Adam优化器。对于联合学习网络，两个子网络使用相同的优化器和学习率，批量大小则调整为 64。所有模型均使用标准反向传播算法进行训练，在预训练阶段训练 100 个周期，在微调阶段训练 60 个周期。我们针对式（5）代价函数中的不同 γ 值进行了实验。我们展示了以下情况的结果： $\gamma = 0.2$、$\gamma = 1$ （仅根据 T_{60} 估计网络的损失更新权重矩阵），以及$\gamma = 0.7$ （在两个子网络之间略微平衡损失）。我们还提供了仅进行去混响的网络的结果，该网络使用随机初始化的LSTM状态。

V. 评估

A. 基线方法与评价指标

我们通过两项主要任务评估了所提出的方法：（1）去混响和（2） T_{60} 估计。以下首先

表 III 去混响：四个模拟房间。 ⋆ 表示所提出的方法显著优于四种基线方法（t 检验）。(·)表示标准差。

	感知语音质量评估⋆				短时客观可懂度⋆				SDR			
	0.3	0.6	0.9	AVG	0.3	0.6	0.9	AVG	0.3	0.6	0.9	AVG
未处理	4.08	2.74	2.25	3.02	0.9970	0.9381	0.8647	0.9333	31.36	12.94	8.16	17.49
加权预测误差 [16], [17]	3.80	3.02	2.42	3.08 (0.31)	0.9818	0.9502	0.8954	0.9425 (0.04)	28.46	**16.71**	10.58	18.58 (4.07)
长短期记忆网络 [13]	3.70	3.13	2.65	3.15 (0.33)	0.9803	0.9501	0.8999	0.9434 (0.03)	30.51	15.21	10.07	18.60 (4.03)
RTA [21]	4.03	3.15	2.61	3.26 (0.29)	0.9950	0.9503	0.8998	0.9350 (0.03)	30.96	15.04	11.02	19.00 (4.26)
TeCANet [23]	4.11	3.22	2.69	3.34 (0.28)	0.9973	0.9571	0.9103	0.9549 (0.03)	31.43	15.45	11.32	19.40 (3.84)
仅去混响	4.20	3.19	2.66	3.35 (0.31)	0.9972	0.9553	0.9062	0.9529 (0.03)	31.04	15.13	11.10	19.09 (4.64)
所提出的方法 (γ=0.2)	**4.26**	3.44	2.91	3.54 (0.27)	**0.9975**	0.9644	0.9276	0.9631 (0.02)	**31.88**	16.31	12.49	20.23 (3.71)
所提出的方法 (γ=0.7)	4.25	**3.47**	**2.97**	**3.56** (0.27)	0.9974	**0.9650**	**0.9304**	**0.9643** (0.02)	31.80	16.37	**12.63**	**20.27** (3.69)
所提出的方法 (γ=1)	4.24	3.45	**2.97**	3.55 (0.27)	0.9974	0.9644	0.9295	0.9638 (0.02)	31.75	16.27	12.55	20.19 (3.70)

表IV 所提出的方法在四个模拟房间中的 T_{60} 估计性能比较

	MSE		MAE		ρ		η	
	Reg	Cls	Reg	Cls	Reg	Cls	Reg	Cls
多层感知机 [10]	0.253	–	0.334	–	0.683	–	0.703	–
卷积神经网络 [20]	0.271	–	0.291	–	0.792	–	0.824	–
$L^B_{joint}(\gamma = 0.2)$	1.544	0.032	1.219	0.121	0.459	0.844	0.032	0.852
$L^B_{joint}(\gamma = 0.7)$	0.903	0.035	0.938	0.126	**0.889**	0.799	**0.905**	0.751
$L^B_{joint}(\gamma = 1)$	0.417	**0.021**	0.613	**0.094**	0.602	0.844	0.631	0.840

介绍去混响任务的基线方法，然后介绍 T_{60} 估计任务的基线方法。

我们实现了五种基线对比方法。第一种方法采用基于统计模型的方法，即加权预测误差（WPE），通过估计逆滤波器，从给定的混响信号中去除晚期混响 [16]，[17]。第二种方法采用长短期记忆网络来估计晚期混响 [13]，并将其从原始信号中减去，以生成直达早期分量。第三种方法采用混响时间感知（RTA）深度神经网络[21] 和上下文窗口，直接估计无混响语音的幅度。类似地，第四种方法 [23],称为 TeCANet，采用上下文感知输入和基于全频带的时间注意力（FTA）方法，其中当前帧作为键，上下文帧作为查询和值。TeCANet 直接预测无混响幅度，并利用混响相位重建波形。我们对 RTA 和 TeCANet 进行修改，使其预测直达早期分量，从而确保所有框架的一致性，便于比较。最后，我们将其与第 III-B 节中的仅去混响模型进行比较。该模型与我们所提出的模型的唯一区别在于，它仅将幅度信息作为输入，并训练模型学习直达早期分量的幅度，而不提供任何房间声学信息。

我们使用语音质量感知评估（PESQ） [54],短时客观可懂度（STOI） [55],和信号失真比（SDR） [56]来评估去混响性能。在每种情况下，参考信号均为混响信号的直达早期分量。

我们实现了两种 T_{60} 估计对比方法：(1) 卷积神经网络（CNN） [20] 和 (2) 时频调制滤波与多层感知机（MLP） [10]。请注意，训练卷积神经网络时我们没有进行数据增强。我们使用对数梅尔频谱图作为卷积神经网络的输入，并采用与 [20]中完全相同的架构。对于多层感知机，我们使用二维Gabor滤波器提取输入，并使用一个用于分类的三层多层感知机 [10]。

我们计算真实值与估计值 T_{60}之间的均方误差（MSE）、平均绝对误差（MAE）、皮尔逊相关系数（PCC， ρ）和斯皮尔曼等级相关系数（SRCC， η）。对于 MSE 和 MAE，值越小表示性能越好；而对于 PCC 和 SRCC，分数越接近 1 则越好。

B. 模拟数据的结果

表 III 中的结果表明，在模拟环境中，我们所提出的方法在三项指标上的整体性能优于所有其他对比方法，包括仅去混响模型。当 $T_{60} = 0.3\,s$ 时，混响较少，听起来在人耳看来几乎干净，这一点也从未处理的混响信号较高的 PESQ 分数（4.08）中得到印证。我们所提出的方法使 PESQ 提高了 0.18，STOI 略有提升，并使 SDR 至少提高了 0.52 dB，增幅最大。大多数对比方法都未能实现去混响，其中 TeCANet 的 PESQ 提高了 0.07。

在轻度混响（$T_{60} = 0.6\,s$）条件下，所提出的方法使感知语音质量评估提高了 0.73，而对比方法中的最佳结果提高了 0.48。所提出的方法在短时客观可懂度方面也取得了最大增益，约为 0.027。然而，在信号失真比方面，加权预测误差优于所有方法，增益约为 3.77，仅略高于我们所提出的方法。加权预测误差表现更好的一个可能原因是，对于这种 T_{60}，无监督学习会预测适用于每个T_{60}的过渡时间，而其他方法则假设所有条件下的过渡时间固定，并针对平均情况进行优化。若没有合适的混合延迟，模型可能会引入更多失真，从而降低信号失真比得分。

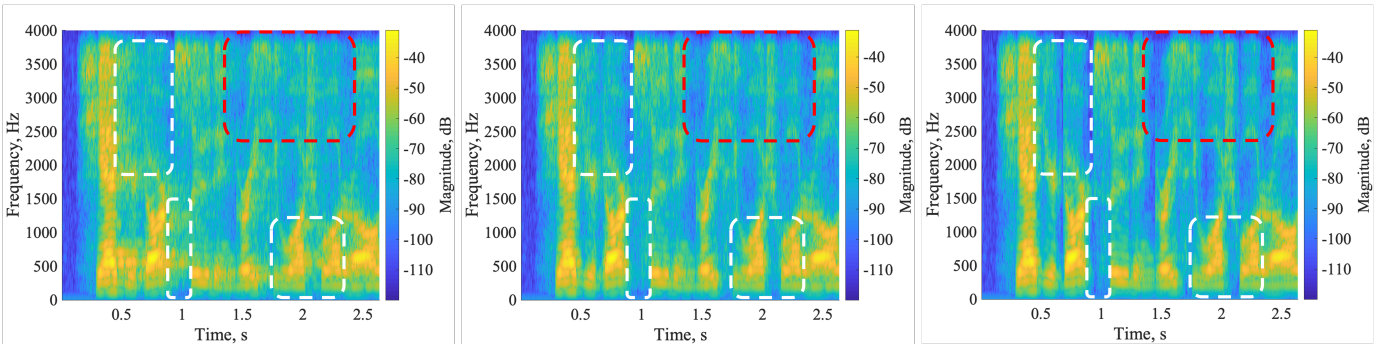


图4。 （左）模拟 的混响信号（$T_{60} = 0.9\,s$）、（中）估计的直达-早期信号和（右）真实值直达-早期信号的频谱图。

在最长混响时间（$T_{60} = 0.9\,s$）下，所提出的方法在全部三项指标上均取得最大提升，感知语音质量评估、短时客观可懂度和信号失真比的提升分数分别为 0.72、0.0657 和 4.47。总体而言，对三种 $T_{60}s$的得分取平均后，所提系统优于所有对比方法，包括仅去混响模型，这表明额外的 T_{60} 信息有助于提升去混响性能。我们给出了所提方法在微调阶段采用较低的T_{60} 估计损失权重（$\gamma = 0.2$）、较高的T_{60} 估计损失权重（$\gamma = 0.7$）以及仅采用 T_{60} 估计损失（$\gamma = 1$）时的结果。即使在 $\gamma = 1$时，感知语音质量评估和短时客观可懂度的平均得分也相当稳定；比较 $\gamma = 0.2$，0.7 和 1 时，信号失真比的平均得分仅有细微差异（0.14）。平均而言，我们所提出的方法在每项评估指标上的标准差均最低，并且三项得分在统计上均显著优于全部四种基线方法（$p < 0.05$，t 检验）以及仅去混响模型。

表 V 展示了非矩形房间中的评估结果，以进一步评估模型泛化能力。我们将所提出的模型与加权预测误差（WPE）以及在其他模拟条件下性能最佳的基线方法 TeCANet 进行了比较。我们的模型在感知语音质量评估、短时客观可懂度和信号失真比方面取得了最大提升，分别提高了 0.43、0.111 和 2.22。

表IV展示了微调阶段后 T_{60} 估计的评估结果。我们对基于分类的回归子任务（例如，式（4）中的 $\alpha = 0.9$ ）赋予更高权重，整体得分反映出联合训练模型在基于分类的估计方面表现更好，尤其是在均方误差和平均绝对误差方面。我们将所提系统与两种基线方法进行比较。Bryan等人 [20] 使用卷积神经网络估计直达混响比（DRR）和T_{60}。Xiong 等人 [10]基于时频调制滤波和深度神经网络估计 T_{60} 。结果表明，我们的模型优于基线方法。不过，对于 $\gamma = 1$ ，分类分支取得了最佳的均方误差和平均绝对误差结果；而对于 $\gamma = 0.7$ ，回归分支在皮尔逊相关系数和斯皮尔曼等级相关系数方面表现最佳。基于分类的回归在均方误差和平均绝对误差方面结果更好，但在皮尔逊相关系数和斯皮尔曼等级相关系数方面得分更差。这可能是因为在联合学习期间，我们只在分类分支中使用均方误差和交叉熵损失项。因此，所提出的方法能够最小化均方误差和平均绝对误差，但这可能会对皮尔逊相关系数和斯皮尔曼等级相关系数产生负面影响。

表 V 非矩形模拟房间中去混响任务所提方法的性能比较。

	PESQ	STOI	SDR
未处理	2.33	0.8790	8.94
加权预测误差 [16], [17]	2.57	0.9095	10.92
TeCANet [23]	2.63	0.9096	11.03
所提出的方法 (γ=0.2)	2.75	0.9096	**11.16**
所提出的方法 (γ=0.7)	**2.76**	**0.9099**	**11.16**
所提出的方法 (γ=1)	**2.76**	**0.9099**	11.13

图4展示了混响信号、所提方法的估计输出信号，以及直达早期参考信号真实值的频谱图。混响信号的幅度相当模糊且失真，尤其是被反射声（特别是晚期反射）扩散的高频段。每个词之间的静音时段难以辨认，便是明证。图4（中）展示了所提系统估计的直达早期分量，其与真实值（右）高度相似，词语之间的混响已被清晰去除（见图4（中）的白色边界框）。不过，部分高频仍受到晚期反射的扩散影响（见红色边界框）。

C. 真实数据结果

尽管我们的方法在模拟测试环境中取得了改进，但仍需使用真实房间脉冲响应和信号进行测试，以确定所提出的方法是否稳健并能推广到真实环境。正如 IV.A 节所述，我们使用 ACE 挑战数据集 [51] 和 BUT Speech@FIT 混响语料库 [52]评估真实环境下的性能。ACE 挑战数据集使用干净语音信号而非直达早期分量作为参考信号。这意味着，退化信号是与干净信号而非直达早期信号进行比较，因此得分会较低。此外，每个真实房间中的物体各不相同，这也可能导致感知语音质量评估分数较低。

表 VI 使用真实混响 ACE 语料库，以感知语音质量评估（PESQ）指标比较所提出的方法。 — 含义与直达声相同

	直达声								直达早期分量							
	大厅	会议室 1	会议室 2	讲堂 1	讲堂 2	Of 1	Of 2	AVG	大厅	会议室 1	会议室 2	讲堂 1	讲堂 2	Of 1	Of 2	AVG
未处理	2.11	2.06	2.12	2.35	1.91	2.10	2.12	2.11	—	—	—	—	—	—	—	—
加权预测误差 [17]	—	—	—	—	—	—	—	—	2.29	2.21	2.27	2.69	2.04	2.15	2.29	2.28
长短期记忆网络 [13]	2.32	2.19	2.20	2.65	2.12	2.17	2.25	2.27	2.12	2.15	2.14	2.57	1.98	2.15	2.19	2.19
RTA [21]	2.44	2.24	2.35	2.70	2.26	2.28	2.34	2.37	2.35	2.20	2.29	2.67	2.13	2.17	2.20	2.29
TeCANet [23]	2.49	2.25	2.37	2.73	2.30	2.31	2.36	2.40	2.39	**2.22**	**2.32**	**2.69**	2.20	2.20	2.24	2.32
仅去混响	2.21	2.10	2.19	2.45	1.97	2.15	2.18	2.18	2.19	2.10	2.17	2.41	1.96	2.15	2.19	2.17
所提出的方法（γ=0.2）	2.55	**2.27**	**2.43**	**2.80**	**2.34**	**2.35**	**2.42**	**2.45**	2.41	2.20	2.31	**2.69**	2.25	**2.23**	2.28	**2.34**
所提出的方法（γ=0.7）	**2.56**	2.26	**2.43**	**2.80**	**2.34**	2.34	2.41	**2.45**	**2.42**	2.20	2.30	2.68	**2.26**	2.22	**2.29**	**2.34**
所提出的方法（γ=1）	2.55	2.25	2.42	2.78	2.30	2.33	2.40	2.43	2.41	2.20	2.30	2.67	**2.26**	2.22	2.28	2.33
所提出的方法（微调）	2.50	·	**2.57**	·	·	·	2.55	·	**2.51**	·	**2.56**	·	·	·	**2.45**	·

表 VII 所提方法（直达声估计）应用于 BUT SPEECH a◯FIT 语料库真实混响语音去混响任务的性能比较

	PESQ	STOI	SDR
未处理	1.48	0.2735	-5.11
加权预测误差 [16], [17]	**1.50**	**0.2757**	-3.93
所提出的方法（γ=0.2）	1.48	0.2736	**-3.45**
所提出的方法（γ=0.7）	1.47	0.2712	-3.76
所提出的方法（γ=1）	1.48	0.2701	-6.86

表 VIII 所提方法（直达声估计）在 ACE 语料库真实混响语音上的 DNSMOS 性能比较

	直达声		
	大厅	会议室 2	Of 2
未处理	2.94	3.03	2.96
加权预测误差 [17]	3.02	3.11	3.09
TeCANet [23]	2.90	3.05	3.16
所提出的方法（γ=0.2）	3.02	3.10	3.12
所提出的方法（γ=0.7）	3.03	3.11	3.11
所提出的方法（γ=1）	**3.04**	**3.16**	**3.22**

为进行更精确的比较，我们开展了两项实验：(1) 如 III-C 节所述，预测直达早期分量；(2) 将训练目标改为直达声，并使用模拟训练集重新训练模型。对于其他基于深度学习的对比方法，我们也针对直达声训练目标重新训练模型。值得注意的是，直达声相对于干净信号存在时间延迟，因此参考信号未进行时间对齐，这会导致某些分数（例如短时客观可懂度和信号失真比）降低。我们无法修改用于估计直达早期分量的加权预测误差方法，因为它是一种无监督方法。我们采用与 V-B 节相同的网络架构，并使用 15000 个模拟信号进行训练，三个 T_{60}分别为：0.3 s、0.6 s 和 0.9 s。

表 VI 展示了使用 ACE 数据集中的真实房间脉冲响应预测每个房间的直达声和直达早期分量时的平均 PESQ 分数。在直达早期分量估计方面，我们所提出的方法在七个房间中的五个房间取得了最大的增益。我们所提系统的平均 PESQ 分数提升优于所有其他方法。在直达声估计方面，基于深度学习的方法总体得分更高。具体而言，我们所提出的方法、加权预测误差、LSTM RTA 和 TeCANet 均实现了 PESQ 提升，但我们所提出的方法在所有方法中对每个房间都取得了最大的增益（例如，在办公室 1 中为 0.34）。此外，我们从 ACE 语料库的七个房间中随机选取四个房间，使用真实数据进行微调。随后，我们在剩余的三个房间上评估性能。最后一行展示了三个测试房间在微调后的 PESQ 分数。结果表明，与先前模型相比，直达声估计和直达早期分量估计的 PESQ 均有显著提升。在短时客观可懂度（0.83）和信号失真比（12 dB）方面，加权预测误差和我们所提方法表现最佳，且几乎相同。

表 VII 展示了使用 BUT Speech a◯FIT 数据库经真实数据微调后预测直达声时的平均分数。我们随机选取五个房间中的三个用于训练，其余两个用于测试，以此对模型进行微调。所提出的方法在平均信号失真比得分方面表现最佳，提升幅度最大，为 1.66。加权预测误差在感知语音质量评估和短时客观可懂度方面表现最佳，但其结果与所提出的方法相当。请注意，该语料库的总体得分较低，这说明在真实环境中去除混响颇具难度。

此外，我们在表 VIII 中使用一种名为 Deep Noise Suppression Mean Opinion Score（DNSMOS）的多阶段数据驱动感知客观指标，评估所提出的针对真实混响语音的去混响估计方法。DNSMOS 使用机器学习来估计人工评估平均意见分 [57]。

结果显示了 ACE 语料库在使用真实环境数据完成微调阶段后，DNSMOS 得分的情况。所提出的方法在大厅、会议室 2 和办公室 2 中分别取得了 0.1、0.13 和 0.26 的最大提升。这表明，根据人工评估者的评估，我们的方法同样表现最佳。

VI. 讨论

A. 不同权重参数下的结果 γ

从表 III 到表 IV，我们给出了所提方法在模拟未见房间中针对去混响和 T_{60} 估计任务、采用不同 γ 值时的结果。对于去混响任务，当 $\gamma = 0.7$ 时，所提方法在感知语音质量评估和短时客观可懂度得分上均表现最佳；当 $\gamma = 0.2$ 时，所提方法在 $T_{60} = 0.3\,s$ 条件下的信号失真比表现最佳。然而，不同 γ 值对应的整体结果差异不大，其中最大增益为感知语音质量评估的 0.01，以及 $T_{60} = 0.3\,s$ 时信号失真比的 0.09，还有短时客观可懂度的 0.0029

（当 $T_{60} = 0.9\,s$ 时）。三项评估指标上的这些细微差异表明，与微调阶段相比， T_{60} 特征对提升去混响性能更为关键，而无论反向传播过程中是否包含 T_{60} 估计损失。

表 IV 中的 T_{60} 评估得分呈现出不同的模式。对于纯回归任务，在四种不同指标下，得分会随着不同的 γ 值而发生显著变化。然而，基于分类的回归受益于联合学习，因为房间声学估计和去混响这两项相关任务能够实现参数共享。这一点在表 IV 中尤为明显：我们所提出的方法总体得分最佳，并展现出良好的泛化能力。表 IV 的结果也解释了为什么我们在联合学习连接中使用分类分支而非回归分支，因为结果表明，分类分支的输出更加可靠。

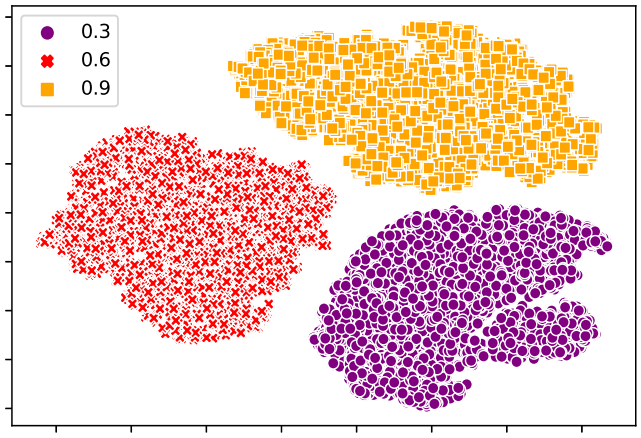


图5.（在线彩色）使用 t-SNE 对模拟训练数据中的 T_{60} 分类特征表示进行可视化。

B. 融入 T_{60} 信息

第 III-B 节介绍了我们的直接早期分量估计（仅去混响）模型。其思路类似于长短期记忆网络[13], , 即随机初始化隐藏状态和单元状态。该仅去混响模型没有房间特性的先验信息，表 III 的结果表明，随着T_{60} 增大，在缺少房间环境信息的情况下，去混响会变得更加困难。融入房间特性的一种方法是在训练阶段（RTA）提供T_{60} 信息: [21], 此时会将一对参数（例如帧移和上下文窗口大小）作为额外输入提供。此类参数会影响输入信号的短时傅里叶变换计算以及由此产生的特征维度。训练期间，网络会根据 T_{60}获得不同的输入特征，额外信息使网络能够更准确地学习无混响语音。至于测试阶段，该方法需要一个外部 T_{60} 估计器和一个查找表，以确定相应的帧移和上下文窗口大小。这种方法表明， T_{60} 对于去混响是必要且有帮助的；然而，外部 T_{60} 估计器并非整体网络的一部分，因此可能不是最优的。

为了找到最佳 T_{60} 信息，我们还研究了其他实验，在这些实验中，我们输入分类结果或

T_{60} 将估计结果作为去混响模块的输入。这些结果无法与 RTA 相媲美，表明这些信息对去混响性能的帮助不大。另一方面，我们所提出的方法通过学习重要的 T_{60} 特征，并通过跳跃连接将其提供给去混响网络，从而解决了这一问题。对比表 III 中 RTA、所提出的方法和仅去混响模型的结果可知，这三种方法相较于未处理信号，感知语音质量评估平均分别提升了 0.24、0.54 和 0.33，短时客观可懂度分别提升了 0.0017、0.031 和 0.0191，信号失真比分别提升了 1.51、2.78 和 1.6 dB。结果表明，先验信息提供了不同的上下文信息以及一种有利于提升性能的替代输入特征。我们推测，这是因为混响时间估计所得的特征有助于区分不同程度的混响。

的混响。在图 5 中，我们使用 t-SNE 将 T_{60} 分类模块倒数第二层的输出可视化。预训练模块的 T_{60} 特征根据混响时间聚类在一起，这可能是直达早期分量估计模块的重要识别依据。将 t-SNE 可视化用于其他自监督算法时，也能得到类似结果 [43]。

C. 局限性与未来工作

第V节讨论了基于使用模拟数据训练的所提出模型所得结果的性能。尽管该模型在模拟测试数据上的性能显著优于基线方法（$p < 0.05$，t 检验），但仍有一个需要解决的局限性：如何恰当地利用真实环境数据。利用模拟数据，我们可以生成均衡且规模较大的数据集，以满足深度学习的要求。然而，对于真实数据，我们通常没有足够的数据，因为录音往往没有相应的真实信号或声学参数，而这些是监督方法所必需的。此外，我们所提出的方法取决于 T_{60} 模块生成有效 T_{60} 特征的能力。然而，在真实环境中， T_{60}往往是未见过的，这会对 T_{60} 类预测产生负面影响。

我们所提出的方法为去混响模块提供了 T_{60} 特征，以实现更准确的去混响。然而，混响训练数据是通过镜像源法生成的，该方法可以确定纯镜面反射的路径长度和压力，但在模拟晚期反射中可能出现的漫反射方面存在局限。基于模拟数据的这一特性，我们的模型只能学习到不准确的融合反射声中的RIR信息，这对真实环境测试产生了负面影响。

我们可以不直接从混响语音本身估计 T_{60} ，而是利用纯视觉数据 [58], [59] 或音视频数据 [60], [61] 来解决去混响问题。对于 3D 视频，视频中会包含相应的房间环境特征，例如房间尺寸、麦克风和扬声器的位置，以及房间内会造成反射声的物体。这些信息可以提供

声学特性的关键指标。借助准确的房间信息，我们就更有可能估计 T_{60}；通过精心设计的嵌入特征，我们所提出的方法也能从中受益。

VII. 结论

总之，我们提出了一种联合学习网络，将 T_{60} 估计信息整合到去混响模块中，以增强混响信号。具体而言，我们将 T_{60} 估计模块的倒数第二个输出作为混响时间特征，并与压缩幅度一同提供给去混响网络。最重要的是，结果表明，将T_{60} 特征提供给联合网络后，客观语音质量和可懂度均显著提升，并且在模拟和真实环境中都取得了改善。

参考文献

[1] A. K. Nabelek 和 J. M. Pickett，“正常听力和听力受损听者通过助听器在噪声和混响环境中的单耳与双耳语音感知”，《语音与听力研究杂志》，第17卷，第724–739页，1974年。[2] K. L. Payton、R. M. Uchanski 和 L. D. Braida，“正常听力和听力受损听者在噪声和混响环境中对会话语音与清晰语音的可懂度”，《美国声学学会杂志》，第95卷，第1581–1592页，1994年。[3] K. Kinoshita、M. Delcroix、T. Yoshioka、T. Nakatani、E. Habets、R. Haeb- Umbach、V. Leutnant、A. Sehr、W. Kellermann、R. Maas 等，“REVERB挑战：用于混响语音去混响与识别的通用评估框架”，载于*WASPAA*会议论文集，2013年，第1–4页。[4] R. Giri、M. L. Seltzer、J. Droppo 和 D. Yu，“利用房间感知深度神经网络和多任务学习提升混响环境下的语音识别”，载于*ICASSP*会议论文集，2015年，第5014–5018页。[5] X. Zhao、Y. Wang 和 D. Wang，“噪声和混响条件下的鲁棒说话人识别”，*IEEE/ACM* 音频、语音与语言处理汇刊，第22卷，第836–845页，2014年。[6] A. Akula、V. R. Apsingekar 和 P. L. De Leon，“利用高斯混合模型-通用背景模型进行室内混响环境下的说话人识别”，载于*DSP/SPE*会议论文集，2009年，第37–41页。[7] S. Chakrabarty 和 E. A. Habets，“利用噪声信号训练的卷积神经网络进行宽带到达方向估计”，载于*WASPAA*会议论文集，2017年，第136–140页。[8] A. Akula 和 P. L. De Leon，“说话人识别中的房间混响补偿”，载于*EUSIPCO*会议论文集，2008年，第1–5页。[9] K. Han、Y. Wang、D. Wang、W. S. Woods、I. Merks 和 T. Zhang，“学习语音去混响与降噪的频谱映射”，*IEEE/ACM* 音频、语音与语言处理汇刊，第23卷，992，2015年 982页。– 。[10] F. Xiong 、 S. Goetze 和 M. B. T.，“ 基于时频调制滤波的混响时间盲估计”，载于*ICASSP*会议论文集，2013年，第443–447页。[11] D. S. Williamson 和 D. Wang，“用于语音去混响与降噪的复数域时频掩蔽”，*IEEE/ACM* 音频、语音与语言处理汇刊，第25卷，第1492–1501页，2017年。[12] J. F. Santos 和 T. H. Falk，“采用上下文感知循环神经网络的语音去混响”，*IEEE/ACM* 音频、语音与语言处理汇刊，第26卷，第1236–1246页，2018年。[13] Y. Zhao、D. Wang、B. Xu 和 T. Zhang，“利用长短期记忆循环神经网络抑制晚期混响”，载于*ICASSP*会议论文集，2018年，第5434–5438页。[14] ——，“采用带自注意力的时间卷积网络进行单通道语音去混响”，*IEEE/ACM* 音频、语音与语言处理汇刊，第28卷，第1598–1607页，2020年。[15] K. Lebart、J.-M. Boucher 和 P. N. Denbigh，“一种基于谱减法的语音去混响新方法”，《*Acta Acustica united with Acustica*》，第87卷，第359–366页，2001年。[16] T. Yoshioka 和 T. Nakatani，“面向盲多输入多输出脉冲响应缩短的多通道线性预测方法推广”，*IEEE* 音频、语音与语言处理汇刊，第20卷，第2707–2720页，2012年。[17] T. Nakatani、T. Yoshioka、K. Kinoshita、M. Miyoshi 和 B.-H. Juang，“基于方差归一化延迟线性预测的语音去混响”，IEEE 音频、语音与语言处理汇刊，第18卷，第1717–1731页，2010年。[18] M. Delcroix、T. Yoshioka、A. Ogawa、Y. Kubo、M. Fujimoto、N. Ito、K. Kinoshita、M. Espi、T. Hori、T. Nakatani 等，“结合先进语音增强与识别技术的线性预测去混响方法：面向REVERB挑战”，载于混响研讨会，2014年。[19] T. Yoshioka 和 M. J. Gales，“面向深度神经网络声学模型的环境鲁棒自动语音识别前端”，《计算机语音与语言》，第31卷，第65–86页，2015年。[20] N. J. Bryan，“用于盲室内声学参数估计的脉冲响应数据增强与深度神经网络”，载于ICASSP会议论文集，2020年，第1–5页。[21] B. Wu、K. Li、M. Yang 和 C.-H. Lee，“基于深度神经网络的混响时间感知语音去混响方法”，IEEE/ACM 音频、语音与语言处理汇刊，第25卷，111，2016年 102页。– 。[22] B. Wu、K. Li、F. Ge、Z. Huang、M. Yang、S. M. Siniscalchi 和 C.- H. Lee，“采用端到端深度学习同时进行语音去混响与声学建模，以实现鲁棒语音识别”，IEEE 信号处理精选主题期刊，第11卷，第1289–1300页，2017年。[23] H. Wang、B. Wu、L. Chen、M. Yu、J. Yu、Y. Xu、S.-X. Zhang、C. Weng、D. Su 和 D. Yu，“TeCANet：用于环境感知语音去混响的时序上下文注意力网络”，arXiv预印本 arXiv:2103.16849，2021年。[24] J. S. Bradley、H. Sato 和 M. Picard，“早期反射对室内语音的重要性”，《美国声学学会杂志》，第113卷，第3233–3244页，2003年。[25] Y. Hu 和 K. Kokkinakis，“早期和晚期反射对人工耳蜗使用者理解混响语音的影响”，《美国声学学会杂志》，第135卷，EL22–EL28页，2014年。[26] Y. Li、Y. Liu 和 D. S. Williamson，“关于基于深度学习的混响时间估计损失函数”，载于ICASSP会议论文集，2021年，第486–490页。[27] A. V. Oppenheim，《离散时间信号处理》。培生印度，1999年。[28] T. Hidaka、Y. Yamada 和 T. Nakagawa，“房间脉冲响应中早期反射与晚期混响之间边界点的新定义”，《美国声学学会杂志》，第122卷，第326–332页，2007年。[29] E. A. Habets、N. D. Gaubitch 和 P. A. Naylor，“使用单个麦克风对含噪语音进行时域选择性去混响”，载于ICASSP会议论文集，2008年，第4577–4580页。[30] S. Gul、M. S. Khan、S. W. Shah 和 A. Ur-Rehman，“将消声预训练语音分离深度神经网络用于单一声源的双耳去混响”，arXiv预印本 arXiv:2208 04626，2022年。[31] R. W. Young，“萨宾混响方程与声功率计算”，《美国声学学会杂志》，第31卷，第912–921页，1959年。[32] S. E. Eskimez、X. Wang、M. Tang、H. Yang、Z. Zhu、Z. Chen、H. Wang 和 T. Yoshioka，“人工聆听与实时字幕：用于语音增强的多任务训练”，arXiv预印本 arXiv:2106 02896，2021年。[33] C.-J. Peng、Y.-J. Chan、C. Yu、S.-S. Wang、Y. Tsao 和 T.-S. Chi，“面向多说话人对话场景中语音增强与说话人识别的基于注意力的多任务学习”，载于IS- CAS会议论文集，2021年，第1–5页。[34] Z. Chen、S. Watanabe、H. Erdogan 和 J. R. Hershey，“利用长短期记忆循环神经网络的多任务学习进行语音增强与识别”，载于Interspeech会议论文集，2015年。[35] X. Dong 和 D. S. Williamson，“用于非侵入式语音质量评估的分类辅助框架”，载于WASPAA会议论文集，2019年，第100–104页。[36] Z. Zhang、P. Vyas、X. Dong 和 D. S. Williamson，“采用多任务框架进行真实场景语音主客观评估的端到端非侵入式模型”，载于ICASSP会议论文集，2021年，第316–320页。[37] Z.-Q. Wang、J. L. Roux、D. Wang 和 J. R. Hershey，“利用展开式迭代相位重建进行端到端语音分离”，arXiv预印本 arXiv:1804.10204，2018年。[38] Q. Wang、J. Du、L.-R. Dai 和 C.-H. Lee，“采用紧凑型神经网络架构，通过多目标学习与集成方法实现高性能语音增强”，IEEE/ACM 音频、语音与语言处理汇刊，第26卷，第1185–1197页，2018年。[39] C. Xu、W. Rao、E. S. Chng 和 H. Li，“用于多任务学习单通道语音分离的移位差分系数目标函数”。载于Interspeech会议论文集

，2018，第3479–3483页。

[40] J. Wang、S. I. Ng、D. Tao、W. Y. Ng 和 T. Lee，“基于多任务学习的儿童语音声学建模研究”，载于 ISCSLP 会议论文集，2018年，第389–393页。[41] Y. Li、T. Zhao 和 T. Kawahara，“使用自注意力机制和多任务学习改进端到端语音情感识别”，载于 Interspeech 会议论文集，2019年，第2803–2807页。[42] S. Boll，“使用谱减法抑制语音中的声学噪声”，IEEE Trans. Acoust. Speech Signal Process.，第27卷，第120页，1979年，第113页。[43] P.-H. Chi、P.-H. Chung、T.-H. Wu、C.-C. Hsieh、Y.-H. Chen、S.-W. Li 和 H.-y. Lee，“Audio albert：用于音频表示自监督学习的轻量级 BERT”，载于 SLT 会议论文集，2021年，第344–350页。[44] J. S. ” Garofolo、L. F. Lamel、W. M. Fisher、J. G. Fiscus、D. S. Pallett 和 N. L. . Dahlgren，“DARPA TIMIT 声学语音连续语音语料库”，载于 Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html，1993年。[45] M. Togami，“用于多通道去混响和语音源分离的深度神经网络联合训练”，载于 ICASSP 会议论文集，2020年，第3032–3036页。[46] X. Feng、N. Li、Z. He、Y. Zhang 和 W. Zhang，“基于DNN的线性预测残差增强用于语音去混响”，载于 APSIPA ASC 会议论文集，2021年，第541–545页。[47] E. Habets，“房间脉冲响应生成器（http://home.tiscali.nl/ehabets/rir generator.html）”，2010年。[48] R. Scheibler、E. Bezzam 和 I. Dokmani´c，“Pyroomacoustics：用于音频房间仿真和阵列处理算法的 Python 软件包”，载于 ICASSP 会议论文集，2018年，第351–355页。[49] M. Vorl¨ander，《可听化》。Springer，2008年。[50] D. Schr¨oder，《交互式虚拟环境的基于物理的实时可听化》。Logos Verlag Berlin GmbH，2011年，第11卷。[51] J. Eaton、N. D. Gaubitch、A. H. Moore 和 P. A. Naylor，“ACE 挑战赛——语料库描述与性能评估”，载于 WASPAA 会议论文集，2015年，第1–5页。[52] I. Sz¨oke、M. Sk´acel、L. Mo˘sner、J. Paliesek 和 J. ˇCernock`y，“真实房间脉冲响应数据集的构建与评估”，IEEE Journal of Selected Topics in Signal Processing，第13卷，第863–876页，2019年。[53] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“LibriSpeech：基于公版有声读物的自动语音识别语料库 ”，载于 ICASSP 会议论文集，2015年，第5206–5210页。[54] ITU-R，“语音质量感知评估（PESQ）：用于窄带电话网络和语音编解码器端到端语音质量评估的客观方法”，2001年，第862页。[55] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，“一种用于预测时频加权带噪语音可懂度的算法”，IEEE 音频、语音与语言处理汇刊，第19卷，第2125–2136页，2011年。[56] C. F´evotte、R. Gribonval 和 E. Vincent，“BSS Eval工具箱用户指南——第2.0版修订版”，技术报告1706，2005年。[57] C. K. Reddy、V. Gopal 和 R. Cutler，“DNSMOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标”，载于 ICASSP 会议论文集，2021年，第6493–6497页。[58] M. Alawadh、Y. Wu、Y. Heng、L. Remaggi、M. Niranjan 和 H. Kim，“从单张360°照片估计房间声学属性”，载于 EUSIPCO 会议论文集，2022年，第857–861页。[59] H. Kim、L. Remaggi、S. Fowler、P. J. Jackson 和 A. Hilton，“使用360度立体相机进行声学房间建模”，IEEE Trans. on Multimedia，第4117–4130页，2020年。[60] D. Li、T. R. Langlois 和 C. Zheng，“面向360度视频的场景感知音频”，ACM Trans. on Graphics (TOG)，第1–12页，2018年。[61] L. Remaggi、H. Kim、P. J. Jackson 和 A. Hilton，“用于房间边界估计和材料识别的视听方法”，载于 AVSU 会议论文集，2018年，第3–9页。



Yuying Li 于2015年获得印第安纳大学（印第安纳州布卢明顿）计算机科学理学学士学位和信息与流程管理工商学士学位，并于2016年获得该校计算机科学理学硕士学位。她目前正在印第安纳大学布卢明顿分校攻读智能系统工程博士学位。她的研究兴趣包括语音去混响、机器学习和视听处理。



Yuchen Liu 于2016年获得肯塔基州丹维尔森特学院计算机科学与金融经济学理学学士学位，并于2018年获得印第安纳大学（印第安纳州布卢明顿）数据科学理学硕士学位。他目前正在印第安纳大学布卢明顿分校攻读计算机科学博士学位。他的研究兴趣聚焦于多种语音和音频相关的深度学习任务，包括音频表征学习、音频隐私、语音识别和语音评估。



Donald S. Williamson 目前是美国俄亥俄州立大学计算机科学与工程系副教授。此前，他曾在印第安纳大学 Luddy信息学、计算与工程学院的计算机科学系担任助理教授和副教授。他的研究兴趣包括语音增强、语音评估和语音隐私。