arXiv:2304.01448v1 [eess.AS] 4 Apr 2023

TORCHAUDIO-SQUIM：TORCHAUDIO 中的无参考语音质量与可懂度测量

Anurag Kumar[1], *Ke Tan*[1], *Zhaoheng Ni*[2], *Pranay Manocha*[3*], *Xiaohui Zhang*[2],*Ethan Henderson*[1], *Buye Xu*[1]

[1] Meta Reality Labs Research, [2] Meta, [3] 普林斯顿大学

摘要

测量语音信号的质量与可懂度，通常是开发语音处理系统的关键步骤。为此，人们开发了多种指标，用于在不同假设下衡量质量与可懂度。本文介绍了一套工具和模型，可利用深度神经网络估计这些已知指标。这些模型已在成熟的 *TorchAudio* 库中提供；该库是 PyTorch 深度学习框架的核心音频与语音处理库。我们将其称为*TorchAudio-Squim*，即 TorchAudio 语音质量与可懂度测量。具体而言，在当前版本的 TorchAudio-squim 中，我们建立并发布了用于估计客观指标 PESQ 、 STOI 和 SI-SDR，以及主观指标平均意见分（ MOS）的模型。我们提出了一种新颖的 客观指标估计方法，并采用近期提出的方法进行主观指标估计。这些模型以“无参考”方式运行，也就是说，它们不需要相应的干净语音作为语音评估的参考。鉴于现实场景中干净语音可能无法获取，且主观评估过程费时费力，这类易用工具将极大助益语音处理研究与开发。

索引词—— 语音质量、语音可懂度、PESQ、短时客观可懂度、尺度不变信号失真比、平均意见分

1. 引言

语音处理系统往往需要处理退化或受损的语音信号。因此，在设计和开发此类系统时，需要衡量语音信号的质量。多年来，人们开发了各种语音评估方法和指标。尽管这些方法会在一定假设下衡量语音信号的退化，但本文将指标分为两大类：主观指标，通过对语音信号进行人工听音测试获得；以及客观指标，无需人类判断，主要通过将给定的（退化）语音信号与相应的参考干净语音进行比较得出。由于依赖参考干净语音，这些指标被称为“侵入式”，与无需参考干净语音的“非侵入式”方法相对。请注意，某些客观指标可能不需要参考干净语音 [1, 2], , 但本文所说的客观指标指更常见的、需要参考干净语音的情形 [1]。

主观和客观指标各有优缺点。与客观方法相比，主观方法是评估语音信号更可靠的方法，因为它们基于人类感知和判断。然而，基于人工的评估难以扩展，通常也不易开展。在大多数情况下，这类评估需要专业听众参与，且过程可能耗时又繁琐。客观指标可以避免这些问题，但也有自身的局限。它们与主观评估的相关性未必总是很好。更重要的是，客观指标需要参考干净语音来评估质量或可懂度，因此不适用于现实场景，因为通常无法获得参考干净语音信号。

为了解决上述主观和客观指标方面的限制，近年来，人们一直在努力构建基于机器学习的指标估计器。在主观指标方面，近期举办了一项平均意见分（MOS）预测挑战赛 [3]。已有多项研究探索了用于语音 MOS 估计的神经网络训练方法 [4, 5, 6, 7, 8,9, 10, 11]。其中一些方法（如 DNSMOS [12]）基于大规模语音 MOS 评分众包数据进行训练，另一些则利用预训练自监督模型来提升估计效果 [7]。在客观指标方面，也有不少研究致力于对 PESQ、短时客观可懂度、ESTOI、HASQI 等知名客观指标进行无参考估计 [13, 14, 15, 16, 17]。

尽管这些研究推动了指标估计器的发展，但这类无参考估计器在不同语音应用中的广泛使用仍不常见。主要原因是缺乏简单易用的工具和推理模型，无法轻松集成到现有语音系统中。本文旨在解决这一问题。此外，这类开源工具和推理模型也将促进并支持未来关于指标估计的研究。

我们在构建系统时牢记以下标准和原则。***(1)*** 易于在真实应用中使用。由于参考干净语音往往不可用，我们只专注于开发非侵入式方法，以便在没有参考干净语音的情况下估计指标。***(2)*** 深度学习在当前众多语音系统的开发中发挥着关键作用，因此我们的系统也将主要基于深度神经网络。这也确保我们拥有可微估计器，能够轻松用于训练其他基于深度学习的语音系统。***(3)***主观指标和客观指标都得到广泛应用，因此系统将包含用于估计这两类指标的估计器。***(4)*** 工具和模型将持续开发，性能更好的模型也将定期发布。这包括更新模型以提升指标估计效果、降低计算开销，以及扩展模型以估计更多主观和客观指标。

为此，我们在 *TorchAudio-Squim* 中发布了用于估算语音评估指标的工具，该工具属于 *Tor-chAudio* [18] 库。*Tor-chAudio* 是 PyTorch 官方的音频领域库，提供音频与语音处理所需的基础构建模块，并推动各种音频与语音问题研究的发展。通过将语音质量与可懂度评估

*在 Meta 实习期间完成的工作。

组件集成到 TorchAudio 中，我们旨在让语音处理系统的设计与开发更易于使用这些指标。当前版本的 *TorchAudio-Squim* 可估算 3 个客观指标和 1 个主观指标。

2. TORCHAUDIO-SQUIM 概述

我们简要介绍目前通过 *TorchAudio-Squim* 提供的语音质量与可懂度评估工具。如前所述，*TorchAudio-Squim* 同时支持主观和客观指标。

客观指标：在当前版本中，我们发布了一个模型，用于估计语音评估领域 3 种广为人知的客观指标：语音质量感知评估（PESQ） [19, 20], 短时客观可懂度（STOI） [21] 和尺度不变信号失真比（SI-SDR） [22]。请注意，我们使用宽带PESQ [20] ，并且本文中的 PESQ 一词均指宽带感知语音质量评估。本文开发了一个单一网络来估计这 3 种指标。我们用于客观指标估计的方法颇具新颖性，并将在后续章节详细介绍该方法，同时提供全面的实验结果与分析。

主观指标：在主观指标方面，我们提供了一个用于估算平均意见分（MOS）的模型。平均意见分是人类听众给出的 1 到 5 分，用于量化给定语音信号的质量。在此版本中，我们发布了用于估算 MOS 的 NORESQA-MOS [23] 模型。此方法使用数据库中的一个或多个随机干净语音样本作为参考。更多详情请参见第 3 节。

在 *TorchAudio* 中，模型架构定义在 *torchaudio.models* 模块中，预训练模型定义在 *torchaudio.pipelines* 中；后者提供用于语音质量和可懂度评估的端到端解决方案。

以下示例代码展示了如何使用*TorchAudio* 库[1]估算 MOS、STOI、PESQ 和 SI-SDR 分数:

```
from torchaudio.prototype.pipelines import { SQUIM_
```

3. 主观指标：系统描述

我们使用 NORESQA-MOS [23] 来估计给定语音信号的平均意见分。与较常见的方法（例如 DNSMOS [9],NISQA [6]）尝试直接根据给定样本估计平均意见分不同，NORESQA-MOS 依赖于利用非匹配参考进行更有依据的估计这一思路 [24]。具体而言，模型会接收从任意数据库中采样的干净语音信号以及待测语音样本，并预测该样本的平均意见分。请注意，使用非匹配参考（NMR）并不会影响该模型相较于“纯”无参考方法的实用性。任何干净语音信号都可以作为 NMR 输入。

当前版本 TorchAudio-Squim 发布的平均意见分模型与 [23], 中的模型相同，该文对 NORESQA-MOS 进行了全面分析和评估。详情请读者参阅该论文。

4. 客观指标：系统描述

本节将详细介绍我们用于估计客观指标的方法。该方法的整体框架如图1左侧面板所示，它使用一个处理时域语音信号的深度神经网络，一次性估计全部三个客观指标。我们提出了一种基于双路径循环神经网络（DPRNN）的新型架构 [25] ，用于对输入的时域信号进行序列建模；随后，多个基于 Transformer 的分支利用所学习到的表示，分别估计各项指标。此外，我们还提出了一种新型多任务训练策略，以提高指标估计的效果。

4.1. 使用双路径循环神经网络模块进行序列建模

对于待评估的长度为T 的信号波形 $y \in \mathbb{R}^T$ ，我们使用一个带步长的一维卷积层，后接修正线性单元（ReLU）函数，对其进行分段和编码，从而得到 L 个相互重叠的表示帧。该卷积层有 N 个输出通道，卷积核大小为 P，步长为 $\lfloor P/2 \rfloor$，因此帧长和帧移分别为 P和 $\lfloor P/2 \rfloor$。随后，将这一时间帧序列划分为 S 个相互重叠的块，块大小为 R，步长为 $\lfloor R/2 \rfloor$。

随后，采用由四个DPRNN模块组成的堆栈来处理块序列$\mathbf{U} = [U_1, \ldots, U_S] \in \mathbb{R}^{N \times S \times R}$，其中$U_1, .., U_S \in \mathbb{R}^{N \times R}$。这种块处理过程可写为

$$\mathbf{U}_b = f_b(\mathbf{U}_{b-1}), b \in \{1, 2, 3, 4\}, \mathbf{U}_0 = \mathbf{U}, \quad (1)$$

其中，下标 b 表示第 b 个 DPRNN 块， $f(\cdot)$ 表示相应块所定义的映射函数。我们采用双向长短期记忆网络（BLSTM）对块内和块间依赖关系进行建模。如图1右侧所示，我们使用两个子块分别进行块内和块间处理。每个子块由一个 BLSTM 层和一个线性投影层组成，线性投影层后接层归一化。块内子块作用于三维表示的第三个维度，块间子块则作用于第二个维度。此外，每个子块都使用残差连接，将输入直接传递至输出。

最后一个 DPRNN 块的输出 $\mathbf{U}_4 \in \mathbb{R}^{N \times S \times R}$ 进一步经过一个含有 N 个单元的线性投影层处理，随后接一个参数化修正线性单元（PReLU）函数。我们对所得三维表示在分块级别执行重叠相加操作，从而得到二维表示 $Z \in \mathbb{R}^{N \times L}$。

4.2. 使用 Transformer模块进行多目标学习

块序列经 DPRNN 模块建模后，学习到的二维表示 Z 将作为不同指标对应的网络分支的输入，每个分支都会生成相应指标得分的估计值。如图 2 所示，每个分支由一个 Transformer 模块组成。具体而言，二维表示 Z 首先输入 Transformer，其中主要包含

[1]https://pytorch.org/audio/main/prototype.pipelines.html

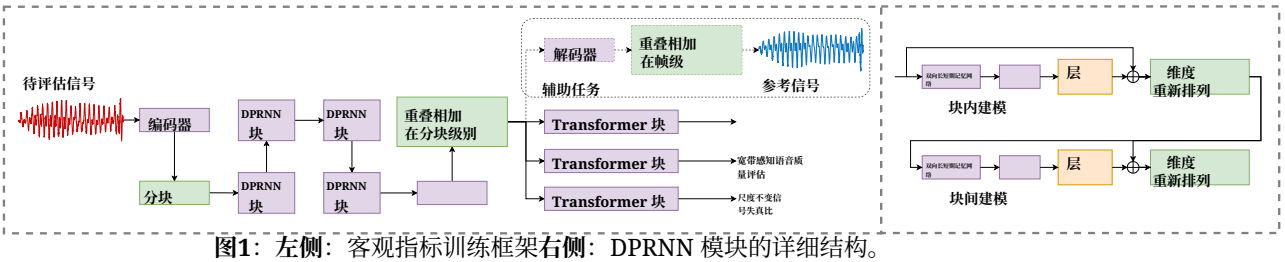

图1：**左侧**：客观指标训练框架**右侧**：DPRNN 模块的详细结构。

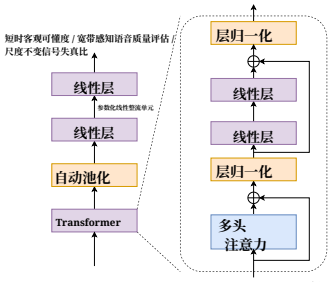

图2：Transformer 块的详细结构。

一个多头注意力模块和两个线性层，如图 2 所示。 [26], 多头注意力的计算公式如下：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, \ldots, H_h)W^O. \quad (2)$$

每个注意力头都

$$H_i = \text{Att}(QW_i^{(Q)}, KW_i^{(K)}, VW_i^{(V)}), i \in \{1, \ldots, h\}, \quad (3)$$

其中 $W_i^{(Q)} \in \mathbb{R}^{d_q \times d}$ $W_i^{(K)} \in \mathbb{R}^{d_k \times d}$ $W_i^{(V} \in \mathbb{R}^{d_v \times d}$ $W_i^{(O)} \in \mathbb{R}^{hd_v \times d}$,,) 表示可训练的投影权重矩阵。我们采用缩放点积注意力函数，即

$$\text{Att}(Q', K', V') = \text{Softmax}(\frac{Q'(K')^\top}{\sqrt{d}})V', \quad (4)$$

其中 $Q' = QW_i^{(Q)}$, $K' = KW_{(K)i}$和 $V' = VW_{(V)i}$。给定单个输入 Z $Q = K = V = Z^\top$，执行自注意力，其中和 $d_q = d_k = d_v = N$。Transformer 中的两个线性层分别有 d_1 和 d 个单元，第一线性层后接 ReLU 函数。

我们对 Transformer 生成的二维表示 $X \in \mathbb{R}^{L \times d}$执行自动池化，得到一个一维表示 $x \in \mathbb{R}^d$。如 [27], 所示，自动池化算子可通过一个可学习参数 $\alpha \in \mathbb{R}$自动适应表示的特征：$x = \sum_{i=1}^{L} x_i \left(\frac{\exp(\alpha \cdot x_i)}{\sum_{j=1}^{L} \exp(\alpha \cdot x_j)} \right)$, 其中 $x_1, \ldots, x_L$ 是一个特征序列。随后，将得到的一维表示输入两个连续的线性层，其单元数分别为 d 和 1，最终得到标量输出 t。请注意，第一个线性层后接 PReLU 函数。对于估计的标量输出，我们应用非线性函数来处理短时客观可懂度和宽带感知语音质量评估，其取值范围大致分别为 $[0, 1]$ 和 $[1, 4.64]$。具体而言，我们对这两个指标都采用 Sigmoid 函数，并额外对宽带感知语音质量评估进行仿射变换，以适配其取值范围，如下所示：

$$s = 1 + \sigma(t) \cdot (4.64 - 1), \quad (5)$$

其中 σ 表示 Sigmoid 函数。

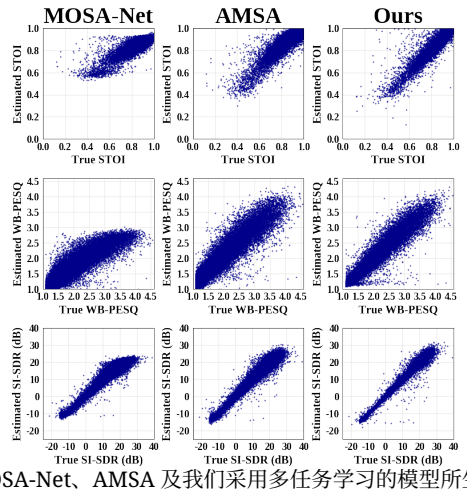

图3：MOSA-Net、AMSA 及我们采用多任务学习的模型所生成的语音评估估计值散点图。

4.3. 通过多任务学习促进语音评估

除指标估计这一主要任务外，我们还通过引入额外的输出分支，将参考信号估计设为辅助任务，类似于 [16]。这种多任务学习（MTL）框架有两方面的益处。首先，侵入式指标依赖参考信号，因此将底层参考信号作为监督信号，可能会促使共享层学习参考信号的潜在表征，从而有助于提升指标估计效果。其次，多任务学习会对共享层的训练施加正则化，我们预计这能提升泛化能力。

如图1左侧面板所示，我们在次级分支中使用一个含有 N 个单元的线性层作为解码器。随后，通过帧级的重叠相加操作，将输出的二维表示转换为时域信号。所有分支共同训练，以最小化不同损失的加权和：

$$\mathcal{L} = \sum_{i=1}^{3} w_i \cdot \mathcal{L}_s(s_i, \hat{s}_i) + w_0 \cdot \mathcal{L}_z(z, \hat{z}), \quad (6)$$

其中，$\hat{s}_i$ 和 s_i 分别表示估计的指标分数及其对应的真实指标分数。$\hat{z}$ 和 z 分别表示估计的参考信号及其真实值。下标 $i = 1$、2、3分别表示短时客观可懂度、宽带感知语音质量评估和尺度不变信号失真比。 $\mathcal{L}_s$ 和 $\mathcal{L}_z$ 分别表示用于指标估计和参考信号估计的训练损失函数， w_i、 $i \in \{0, 1, 2, 3\}$ 表示加权因子。

4.4. 实验

4.4.1. 数据与设置

我们在实验中使用 DNS Challenge 2020 [29] 数据集。退化语音主要通过两种方法获得。首先，

表 1：不同损失函数和加权因子的对比研究。粗体数字标出了每种情况下的最佳结果。

$\mathcal{L}_s$	加权因子	短时客观可懂度 (%)			WB-PESQ			SI-信号失真比 (dB)			With MTL?
		平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	
MSE	$w_1 = 1, w_2 = 1, w_3 = 1$	2.606	0.929	0.919	0.193	0.925	0.939	1.269	0.973	0.969	No
MAE	$w_1 = 1, w_2 = 1, w_3 = 1$	2.442	0.934	0.928	0.175	0.938	0.950	1.163	0.977	0.974	No
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5$	2.324	0.939	0.935	0.168	0.942	0.951	1.158	0.977	0.973	No
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 0.01$	2.310	0.936	0.934	0.165	0.944	0.954	1.129	0.978	0.975	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 0.1$	2.182	0.942	0.943	0.157	0.949	0.956	1.010	0.982	0.980	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 1$	2.039	0.947	0.947	0.143	0.956	0.962	0.843	**0.986**	**0.985**	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 1.5$	2.018	0.949	0.947	0.143	0.957	0.962	0.841	**0.986**	0.984	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 2$	**1.994**	**0.950**	0.950	**0.142**	**0.958**	**0.963**	**0.838**	0.985	**0.985**	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 2.5$	2.035	**0.950**	**0.951**	0.149	**0.958**	**0.963**	0.841	0.985	0.984	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 3$	2.078	0.949	0.949	0.149	0.956	**0.963**	0.849	**0.986**	**0.985**	Yes
MAE	$w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 5$	2.001	0.949	0.950	**0.142**	0.957	0.961	0.845	0.985	0.984	Yes

表 2：不同方法在平均绝对误差、皮尔逊相关系数和斯皮尔曼等级相关系数方面的比较。此处给出了处理 5 秒音频样本所需的乘加运算次数（MACs）。

方法	STOI（%）			WB-PESQ			SI-信号失真比（dB）			参数数量	每 5 秒的乘加运算次数
	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑		
Quality-Net [28]	-	-		0.396	0.845	0.849	-	-	-	0.30 百万	297.30 千
MOSA-Net [17]	5.254	0.900	0.864	0.335	0.904	0.914	1.990	0.965	0.958	317.19 M	94.86 G
AMSA [13]	3.498	0.913	0.826	0.207	0.932	0.938	1.562	0.968	0.964	2.96 M	687.61 M
MetricNet [16]	-	-	-	0.182	0.938	0.947	-	-	-	6.61 M	2.08 G
不使用多任务学习的方法	2.324	0.939	0.935	0.168	0.942	0.951	1.158	0.977	0.973	7.39 M	40.27 G
使用多任务学习的方法	**1.994**	**0.950**	**0.950**	**0.142**	**0.958**	**0.963**	**0.838**	**0.985**	**0.985**	7.39 M	40.27 G

表 3：多目标学习的研究。

方法	加权因子	短时客观可懂度 (%)			WB-PESQ			尺度不变信号失真比 (dB)			With MTL?
		平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	平均绝对误差 ↓	皮尔逊相关系数 ↑	斯皮尔曼等级相关系数 ↑	
仅 STOI	-	2.329	0.935	0.928	-	-	-	-	-	-	No
仅使用 PESQ	-	-	-	-	0.177	0.935	0.947	-	-	-	No
仅使用尺度不变信号失真比	-	-	-	-	-	-	-	1.177	0.976	0.947	No
多目标	$w_1 = 1, w_2 = 1, w_3 = 1$	2.442	0.934	0.928	0.175	0.938	0.950	1.163	**0.977**	**0.974**	No
多目标	$w_1 = 1, w_2 = 2, w_3 = 0.5$	**2.324**	**0.939**	**0.935**	**0.168**	**0.942**	**0.951**	**1.158**	**0.977**	0.973	No

我们将干净语音与加性噪声混合，信噪比（SNR）范围为 -15 至 25 dB。其次，我们随机选取部分噪声混合语音，使用三种语音增强系统之一进行处理。这些语音增强系统基于 GCRN架构 [30], , 由于配置不同，性能也各不相同。训练集、验证集和测试集分别包含约 364500、14600 和 22800 个音频样本。受篇幅限制，我们无法展示这些数据中 PESQ、短时客观可懂度和尺度不变信号失真比的分布，但其取值范围分别为：PESQ 为 1 至 4.6，短时客观可懂度为 0.25 至 1，尺度不变信号失真比为 -18 至 35 dB。所有训练信号均截断为 5 秒。

为确保公平比较，所有模型均在我们的训练集和测试集上进行训练与评估。对于我们的模型，我们采用以下超参数配置：$N = 256$ $P = 64$ $R =$,,,,,。与此 71 $h = 4$ $d = 256$ $d_1 = 1024$ [25], R 类似，的取值经过选择，使得 $R \approx \sqrt{2}L \approx S$ 适用于 5 秒训练信号。

4.4.2. 结果与讨论

我们使用平均绝对误差（MAE）、皮尔逊相关系数（PCC）和斯皮尔曼等级相关系数（SRCC）来衡量指标估计的性能。对于 PCC 和 SRCC，分数越高表示性能越好。

表 1 对比了不同的损失函数和加权因子。我们观察到，将平均绝对误差损失用作 $\mathcal{L}_s$，比使用均方误差（MSE）损失能显著提升性能。此外，我们还比较了不同的加权因子取值，结果表明，采用 $w_1 = 1, w_2 = 2, w_3 = 0.5, w_0 = 2$ 进行多任务学习（MTL）时，在全部三项指标上的表现几乎最佳。

表 2 将我们的模型与近期几种基于深度学习的语音评估模型进行了比较，包括 Quality-Net [28], MOSA-Net [17], AMSA [13] 和 MetricNet [16]。请注意，MOSA- Net 和 AMSA 旨在同时估计多个指标，而 Quality-Net 和 MetricNet 仅估计 PESQ。对于我们的模型，如果采用多任务学习（MTL），则使用 $w_1 = 1$，$w_2 = 2$，$w_3 = 0.5$ 和 $w_0 = 2$。我们观察到，在平均绝对误差、皮尔逊相关系数和斯皮尔曼等级相关系数方面，我们的模型显著优于所有基线方法。在此外，通过采用多任务学习训练，我们的模型性能还能进一步提升。例如，采用多任务学习后，宽带感知语音质量评估（WB-PESQ）估计的平均绝对误差、皮尔逊相关系数和斯皮尔曼等级相关系数分别从 0.168、0.942 和 0.951 提升至 0.142、0.958 和 0.963。我们在图 3 中进一步分析了这一点，可以看到，与 MOSA-Net 和 AMSA 相比，我们模型的数据点更密集地分布在对角线附近，这表明我们的模型估计的分数与真实分数的相关性更高。

我们的方法基于多目标学习，即同时训练一个模型来估计多个指标。我们通过训练三个具有相同架构的模型（图1中每个模型仅有一个输出分支），分别估计短时客观可懂度、PESQ 和尺度不变信号失真比，以考察这一策略的效果。如表3所示，与分别单独估计各项指标的模型相比，训练单个模型同时估计多个指标不仅不会降低性能，反而会略有提升。这是因为不同指标之间存在相关性，因此每个估计分支都会对其他分支的训练起到正则化作用，从而提升模型的泛化能力。此外，由于不同目标之间共享模块，这种多目标学习方法在计算上也更加高效。

5. 结论

我们介绍了*TorchAudio-Squim*，这是一种用于语音质量和可懂度评估的系统。它作为 PyTorch 中*TorchAudio* 的一部分发布，使用户能够轻松、便捷地使用深度学习方法，以非侵入式方式估计语音质量和可懂度。这不仅有助于需要评估语音信号的各种语音系统，也支持对非侵入式语音评估方法的研究。*TorchAudio-Squim*通过新颖的方法支持对主观和客观语音指标进行估计，且已证明其性能优于以往最先进的方法。此外，这些模型还将在*TorchAudio-Squim*的后续版本中持续开发和改进。我们计划将其扩展到其他语音评估指标，并探索开发计算效率更高的模型。

6. 参考文献

[1] P. C. Loizou, 《语音质量评估》，载于多媒体分析与通信。施普林格，2011年。[2]P. Gray、M. Hollier 和 R. Massara, 《使用声道模型进行非侵入式语音质量评估》，*IEE*《视觉、图像与信号处理学报》，第 147 卷，第 6 期，第 493–501 页，2000年。[3] W.-C. Huang、E. Cooper、Y. Tsao、H.-M. Wang、T. Toda 和 J. Yamagishi, 《VoiceMOS 挑战赛 2022》，载于*Interspeech*，2022年，第 4536–4540 页。[4] C.-C. Lo、S.-W. Fu、W.-C. Huang、X. Wang、J. Yamagishi、Y. Tsao 和 H.-M. Wang, 《MOSNet：用于语音转换的基于深度学习的客观评估》，载于*Interspeech*，2019年，第 1541–1545 页。[5] B. Patton、Y. Agiomyrgiannakis、M. Terry、K. Wilson、R. A. Saurous 和 D. Sculley, 《AutoMOS：学习一种非侵入式语音自然度评估器》，载于*NIPS 2016* 语音与音频处理端到端学习研讨会，2016年。[6] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller, 《NISQA：一种用于多维语音质量预测的卷积神经网络自注意力模型及众包数据集》，载于*Interspeech*，2021年，第 2127–2131 页。[7] E. Cooper、W.-C. Huang、T. Toda 和 J. Yamagishi, 《MOS 预测网络的泛化能力》，载于*IEEE* 国际声学、语音与信号处理会议。IEEE，2022年，第 8442–8446 页。[8] W.-C. Huang、E. Cooper、J. Yamagishi 和 T. Toda, 《LD-Net：合成语音 MOS 预测中的统一听者依赖建模》，载于*IEEE* 国际声学、语音与信号处理会议。IEEE，2022年，第 896–900 页。[9] 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*IEEE* 国际声学、语音与信号处理会议。IEEE，2021年，第 6493–6497 页。[10] A. A. Catellier 和 S. D. Voran, 《WaweNets：一种用于估计窄带和宽带语音质量的无参考卷积波形方法》，载于*IEEE* 国际声学、语音与信号处理会议。IEEE，2020年，第 331–335 页。[11] Z. Zhang、P. Vyas、X. Dong 和 D. S. Williamson, 《一种利用多任务框架对真实环境中的语音进行主观和客观评估的端到端非侵入式模型》，载于*IEEE* 国际声学、语音与信号处理会议。IEEE，2021年，第 316–320 页。[12] C. K. Reddy、V. Gopal 和 R. Cutler, 《DNSMOS P. 835：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*IEEE* 国际声学、语音与信号处理会议。IEEE，2022年，第 886–890 页。[13] X. Dong 和 D. S. Williamson, 《一种用于真实环境中客观语音评估的注意力增强多任务模型》，载于*IEEE* 国际声学、语音与信号处理会议。IEEE，2020年，第 911–915 页。[14] X. Dong 和 D. S. Williamson, 《一种结合分类的非侵入式语音质量评估框架》，载于*WAS- PAA*，2019年。[15] R. E. Zezario、S.-W. Fu、C.-S. Fuh、Y. Tsao 和 H.-M. Wang, 《Stoi-net：一种基于深度学习的非侵入式语音可懂度评估模型》，载于*2020* 亚太信号与信息处理协会年度峰会与会议（*APSIPA ASC*）。IEEE，2020年，第 482–486 页。

[16] M. Yu, C. Zhang, Y. Xu, S. Zhang, and D. Yu, “MetricNet：迈向改进的非侵入式语音质量评估建模”，发表于*Interspeech*，2021年，第 2142–2146 页。[17] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, “基于深度学习的非侵入式多目标语音评估模型及其跨域特征”，发表于*IEEE/ACM* 音频、语音与语言处理汇刊，2022年。[18] Y.-Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang 等, “Torchaudio：构建音频与语音处理模块”，发表于 *IEEE* 国际声学、语音与信号处理会议。IEEE，2022年，第 6982–6986 页。[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “语音质量感知评估（PESQ）：一种用于电话网络和编解码器语音质量评估的新方法”，发表于*IEEE* 国际声学、语音与信号处理会议。IEEE，2001年，第 2 卷，第 749–752 页。[20] I. Rec, “P.862.2：P.862 建议书的宽带扩展，用于评估宽带电话网络和语音编解码器”，国际电信联盟，瑞士日内瓦，2005年。[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “一种用于预测时频加权噪声语音可懂度的算法”，发表于*IEEE* 音频、语音与语言处理汇刊，第 19 卷，第 7 期，第 2125–2136 页，2011年。[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR——半生不熟还是恰到好处？”，发表于*IEEE* 国际声学、语音与信号处理会议。IEEE，2019年，第 626–630 页。[23] P. Manocha and A. Kumar, “使用非匹配参考的 MOS 语音质量评估”，发表于*Interspeech*，2022年，第 654–658 页。[24] P. Manocha, B. Xu, and A. Kumar, “NORESQA：一种使用非匹配参考进行语音质量评估的框架”，发表于神经信息处理系统大会，第 34 卷，2021年。[25] Y. Luo, Z. Chen, and 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”，发表于*IEEE* 国际声学、语音与信号处理会议。IEEE，2020年，第 46–50 页。[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All You Need”，发表于神经信息处理系统进展，第 30 卷，2017年。[27] B. McFee, J. Salamon, and J. P. Bello, “用于弱标注声音事件检测的自适应池化算子”，发表于*IEEE/ACM* 音频、语音与语言处理汇刊，第 26 卷，第 11 期，第 2180–2193 页，2018年。[28] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, “Quality-Net：基于双向长短期记忆网络的端到端非侵入式语音质量评估模型”，发表于*Interspeech*，2018年，第 1873–1877 页。[29] C. K. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun 等, “INTERSPEECH 2020 深度噪声抑制挑战赛：数据集、主观测试框架与挑战赛结果”，发表于*Interspeech*，2020年，第 2492–2496 页。[30] K. Tan and D. Wang, “使用门控卷积循环网络学习复杂频谱映射以实现单耳语音增强”，发表于*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 380–390 页，2019年。