arXiv:2305.03837v1 [eess.AS] 5 May 2023

掩蔽偏差：利用内部语言模型估计提升基于 CTC 的自动语音识别的领域自适应泛化能力

Nilaksh Das、*Monica Sunkara*、*Sravan Bodapati*、
Jinglun Cai、*Devang Kulshreshtha*、*Jeff Farris*、*Katrin Kirchhoff*AWS AI Labs，美国

摘要

在大量数据上训练的端到端自动语音识别模型往往会隐式地偏向训练数据中的语言语义。为减轻这种偏差，已有研究提出内部语言模型估计（ILME），用于注意力式编码器-解码器和 RNN-T 等自回归模型。通常，ILME 通过将模型架构中的声学组件与语言组件模块化，并消除声学输入，再与纯文本后验概率进行对数线性插值来实现。然而，对于基于 CTC 的自动语音识别，要将模型解耦为这类声学组件和语言组件并不那么直接，因为 CTC 对数后验概率是以非自回归方式计算的。本文为基于 CTC 的自动语音识别模型提出了一种新颖的 ILME 技术。我们的方法通过迭代地对音频时间步进行掩码，仅累积被掩码时间步的对数后验概率，从而估计内部语言模型的伪对数似然。在多个域外数据集上的广泛评估表明，当仅有目标域的文本数据可用时，与浅层融合相比，所提出的方法可将词错误率最多降低 9.8%，将未登录词 F1 分数最多提升 24.6%。在零样本领域自适应的情况下，即无法获取任何目标域数据时，我们证明，利用 ILME 去除源领域偏差，仍可胜过浅层融合，使词错误率相对降低最多 9.3%。

索引词— 内部语言模型估计、语音识别、CTC

1. 引言

如今，自动语音识别（自动语音识别）已通过数字助理、新闻转录和基于人工智能的交互式语音应答电话系统等常见的现实应用，深度融入日常生活。在这些实际场景中，自动语音识别模型通常基于某个源域的训练数据进行训练，然后部署到其他目标领域，而这些目标领域的训练数据，尤其是配对音频文本数据，可能十分有限。为应对这类领域不匹配问题，研究者提出了外部语言模型（语言模型）集成技术，利用目标领域的纯文本数据 [1, 2, 3, 4, 5, 6]。除了添加外部目标领域语言模型以捕捉目标领域语义之外，研究者还提出了密度比估计技术，通过利用额外的源域语言模型，尽可能减小源域带来的偏差 [7, 8, 9]。在此基础上，近年来越来越多的研究还开始探索内部语言模型估计（ILME），旨在估计模型自身内部学习到的隐式语言语义，从而减轻源域造成的内部模型偏差 [10, 11, 12]。

ILME 旨在通过将模型架构中的声学模型（AM）和语言模型（LM）组件模块化，并完全消除声学输入，来减小源域模型偏差。随后，将整体对数似然分数进行对数线性插值，以减去纯文本对数后验概率 [10]。

表 1：CTC 模型内部偏差示例。尽管输入音频中的词语被屏蔽，模型仍会尝试解码出连贯的输出。

REF	他是……的一名成员共和党主街合作组织
HYP：贪心解码	他是共和党主街合作伙伴关系的成员。
HYP：屏蔽“the”	他是……的成员公共与主街合作伙伴关系
HYP：掩码“street”，他是共和党主流 伙伴关系	

这种内部语言模型估计方法从根本上仅适用于自回归模型，例如基于注意力的编码器-解码器（注意力编码器-解码器，AED）和循环神经网络转换器（RNN-T），因为相对容易将此类模型架构解耦并近似为声学模型和语言模型组件。

最近，基于连接时序分类（CTC）损失的模型凭借较高的推理速度和较低的内存占用，展现出了最先进的自动语音识别性能 [13, 14, 15, 16]。然而，所提出的 ILME 技术无法直接应用于基于 CTC 的自动语音识别模型。鉴于 CTC 的非自回归架构，将 CTC 模型分解为声学模型和语言模型组件存在歧义。本文提出了一种适用于 CTC 模型的新型 ILME 技术。我们的方法是在迭代过程中对输入音频时间步进行掩码，通过累积每次迭代中仅对应掩码时间步的对数后验概率，来估计内部语言模型（ILM）的伪对数似然。这一过程以迭代式非自回归的方式消除声学输入，从而显现内部语言模型偏差，并避免了将模型解耦为声学模型和语言模型的需求。本文的主要贡献如下：

- **用于基于 CTC 的自动语音识别的新型 ILME 技术**。虽然内部语言模型估计主要是在注意力编码器-解码器和 RNN-T 等自回归模型的背景下提出的，但我们提出了一种基于掩码的新型内部语言模型估计技术，用于非自回归 CTC解码。我们试验了多种掩码方法，包括在词边界处对音频进行掩码。我们发现，只需将音频划分为等长片段并对每个片段进行掩码，即可获得最佳结果。
- **在多个域外数据集上进行广泛评估**。我们在 4 个具有不同目标域的公开域外数据集上评估了所提出的内部语言模型估计方法。实证评估表明，与浅层融合相比，所提出的方法在未见过的目标领域上可使词错误率相对降低（WERR）最高达 9.8%。
- **零样本领域自适应**。与传统观点不同，后者认为 ILME 需要使用目标域数据训练的目标域语言模型进行推理；我们证明，所提出的 ILME 技术仅使用源域语言模型，相较于浅层融合即可实现最高 9.3% 的相对词错误率降低率，因此无需任何额外的目标域数据。
- **ILME 可提升未登录词 F1**。我们证明，所提出的方法不仅能降低词错误率（WER），还能持续提升目标领域未登录词检测的 F1 分数，相较于浅层融合最高可提升 24.6%。

2. 背景

给定语音特征输入序列 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ 以及对应的输出词元序列 $\mathbf{Y} = \{y_1, \ldots, y_L\}$，自动语音识别模型 θ 的目标是估计后验分布 $P(\mathbf{Y}|\mathbf{X};\, \theta)$。通常会计算对数后验概率以确保数值稳定性。最后，使用自回归束搜索算法根据以下准则，从这些对数后验概率中确定最优转录文本 $\hat{\mathbf{Y}}$：$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \Big[\log\ P(\mathbf{Y}|\mathbf{X};\, \theta^{\mathrm{S}}_{\mathrm{ASR}}) \Big]$。

用于领域自适应的语言模型融合

也称为纯文本 自适应，在语言模型融合中，单独训练的目标域语言模型 $\theta^{\mathrm{T}}_{\mathrm{LM}}$ 与源领域自动语音识别模型 $\theta^{\mathrm{S}}_{\mathrm{ASR}}$ 相结合，使系统倾向于输出具有目标领域语义的文本。浅层融合 [17] 是最常用的语言模型融合方法之一 [18]。在这种方法中，推理期间会将源领域自动语音识别模型的对数后验概率与目标域语言模型的对数概率进行插值。随后，根据以下准则使用加权对数分数，通过束搜索在每一步选择输出词元：$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \Big[\log\ P(\mathbf{Y}|\mathbf{X};\, \theta^{\mathrm{S}}_{\mathrm{ASR}}) + \lambda_T \log\ P(\mathbf{Y};\, \theta^{\mathrm{T}}_{\mathrm{LM}}) \Big]$，其中 $P(\mathbf{Y};\, \theta^{\mathrm{T}}_{\mathrm{LM}})$ 是目标域语言模型赋予词元序列 $\mathbf{Y}$ 的概率，λ_T 是目标语言模型权重。在本文中，我们将提出的内部语言模型估计方法与这种广泛使用的领域自适应浅层融合技术进行比较。

AED 和 RNN-T 模型的内部语言模型估计

为减少从源域隐式学习到的模型偏差，已有研究提出了内部语言模型估计 [10]。为估计内部语言模型，可采用联合 softmax 近似 [19]，将联合模型参数 θ^{S}_{ASR} 分解为独立的声学模型和语言模型部分。假设声学条件具有域不变性，根据贝叶斯定理对后验概率进行重新表述 [10]，可得推理准则：

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \Big[\log P(\mathbf{Y}|\mathbf{X}; \theta^{\mathrm{S}}_{\mathrm{ASR}}) + \lambda_T \log P(\mathbf{Y}; \theta^{\mathrm{T}}_{\mathrm{LM}}) - \lambda_I \log P(\mathbf{Y}; \theta^{\mathrm{S}}_{\mathrm{ASR}}) \Big], \quad (1)$$

其中，λ_I 是内部语言模型权重，而 $\log\ P(\mathbf{Y};\, \theta^{\mathrm{S}}_{\mathrm{ASR}})$ 通过完全消除声学输入，以自回归方式计算。由于 AED 和 RNN-T 模型本质上采用自回归解码，且模型参数可直接拆分为声学模型和语言模型部分，因此该方法仅适用于这两类模型。在本研究中，我们提出一种新方法，将 ILME 算法应用于 CTC 模型的非自回归解码。

3. CTC 自动语音识别的内部语言模型估计

原始 ILME 技术的核心原理是以自回归方式估计模型的内部语言模型，$P(\mathbf{Y};\, \theta^{\mathrm{S}}_{\mathrm{ASR}})$：$P(\mathbf{Y};\, \theta^{\mathrm{S}}_{\mathrm{ASR}}) = \prod_{l=1}^{L} P(y_l|\mathbf{Y}_{1:l-1};\, \theta^{\mathrm{S}}_{\mathrm{ASR}})$，这使其天然适用于 AED 和 RNN-T 模型。由于这些模型采用自回归解码，它们包含一个参数化组件，可视为学习词元序列上的概率——对于 AED 模型，这一组件是解码器网络；对于 RNN-T 模型，则是预测网络。相较之下，CTC 模型以非自回归方式计算对数后验概率，即每个输入时间步 $\mathbf{x}_t$ 都对应 $\mathbf{Y} = \{y_1, \ldots, y_T\}$ 个输出词元（包括 *blank* 词元）。因此，不存在专门用于词元的参数化组件。给定输入 $\mathbf{X}$，CTC 损失还使输出词元 $\mathbf{Y}$ 之间具有固有的条件独立性。然而，Transformer 和基于 Conformer 的神经网络架构中的自注意力与卷积机制会隐式地放宽这种条件独立性，这是因为注意力和卷积产生的中间特征在序列层面相互依赖。因此，CTC 模型仍会无意中学习到内部语言模型。表 1 展示了一个例子。此外，CTC 模型的非自回归架构使得人们难以在结构上将模型分解为相应的声学组件和语言组件，而这种分解是通过常规方式估计内部语言模型时消除声学输入的先决条件。

从根本上说，内部语言模型是在缺少任何声学上下文时模型所体现的偏差。因此，我们提出一种迭代式输入掩码方法，以克服上述在 CTC 模型中进行 ILME 所面临的挑战。概括而言，我们多次对输入音频进行掩码，并迭代执行前向传播，以确定相应掩码时间步的 CTC 对数后验概率。随后，我们累加不同掩码时间步组的对数后验概率，并在后验域中对其进行归一化，从而计算出最终的内部语言模型分布。下面将进一步详细介绍我们的方法。

迭代式输入掩码

由于 CTC 模型的声学组件与语言组件并不容易解耦，因此我们无法简单地将整个声学输入置零来估计 ILM。因此，我们不完全消除声学输入，而是提出通过迭代掩蔽输入音频来执行 ILME。给定一个输入语音序列$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$，我们将输入时间步划分为 K 个相等且互不重叠的分区，并分别消除每个分区中的输入，从而得到一组 K 个经过掩蔽的序列 $\{\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K\}$，使得

$$\tilde{\mathbf{x}}_t^{(k)} = \begin{cases} 0, & \text{if } t_{k-1} \le t < t_k \\ \mathbf{x}_t & \text{otherwise} \end{cases} \quad (2)$$

其中，$\tilde{\mathbf{x}}_t^{(k)}$ 是 $\tilde{\mathbf{X}}_k$在 t 处的输入，同时也是 $t_0 = 1$ 和 $t_k = kT/K$。

这里可以采用多种掩码策略，例如在词边界处进行掩码，但通过实验我们发现，将音频简单地划分为若干等长片段能够取得最佳效果。接下来，我们将讨论如何利用这些掩码序列进行 ILME。

为 CTC 计算内部语言模型

给定一个包含 N 个词元 $\{\mu_1, \ldots, \mu_N\}$ 的词汇表，我们首先将 CTC 模型在时间步 t 计算得到的对数后验概率表示为：

$$\mathbf{\Psi}_t(\mathbf{X}) = \begin{bmatrix} \log P(y_t = \mu_1|\mathbf{X}; \theta^{\mathrm{S}}_{\mathrm{CTC}}) \\ \vdots \\ \log P(y_t = \mu_N|\mathbf{X}; \theta^{\mathrm{S}}_{\mathrm{CTC}}) \end{bmatrix} \quad (3)$$

接下来，我们计算给定掩蔽序列 $\mathbf{\Psi}_t(\tilde{\mathbf{X}}_k)$ 的对数后验分布向量。现在，我们想确定这些对数概率对应的是原始声学输入，还是受到了声学掩蔽的影响。如果是后一种情况，则表明模型在没有接收到声学信息时存在偏差，因此对应于模型的内部语言模型。为此，我们计算对数后验分布之间逐元素的绝对差，记作 $\boldsymbol{\epsilon}_t^k = |\mathbf{\Psi}_t(\tilde{\mathbf{X}}_k) - \mathbf{\Psi}_t(\mathbf{X})|$。直观来说，如果向量 $\boldsymbol{\epsilon}_t^k$的所有元素值都相对较低，例如低于某个阈值 γ，就意味着时间步 t 未受 K^{th} 掩蔽影响，因此在内部语言模型的计算中可以忽略。反之，如果 $\boldsymbol{\epsilon}_t^k$ 包含任何相对较高的值，我们就希望在内部语言模型的计算中保留 $\mathbf{\Psi}_t(\tilde{\mathbf{X}}_k)$，因为这表明受掩蔽影响的模型偏差。

因此，我们取向量 $\boldsymbol{\epsilon}_t^k$ 中 N 值的最大值，记为 δ_t^k，并在所有时间步上对其进行归一化，得到一个标量 $\hat{\delta}_t^k$，再将其与阈值 γ 进行比较，以确定 $\mathbf{\Psi}_t(\tilde{\mathbf{X}}_k)$ 是否会对内部语言模型产生贡献。具体而言：

$$\delta_t^k = \max_n \big(\boldsymbol{\epsilon}_t^k[n]\big); \quad \hat{\delta}_t^k = \frac{\delta_t^k}{\max_{t'}(\delta_{t'}^k)} \quad (4)$$

最后，我们根据相应的 $\hat{\delta}_t^k$ 是否高于阈值，利用所有 K 掩码对应的对数后验分布计算伪对数似然，从而估计时间步 t 的内部语言模型。因此，

$$\log P_{\text{ILM}}(\mathbf{y}_t; \theta^{\text{S}}_{\text{CTC}}) = \text{LogSoftmax}\left(\sum_{k=1}^{K} \mathbf{\Psi}_t(\tilde{\mathbf{X}}_k)\left[\hat{\delta}_t^k > \gamma\right]\right) \quad (5)$$

虽然我们展示的是单个时间步 t 的 ILM 计算，但讨论的所有步骤都很容易沿序列维度向量化，并可使用高效的次线性方法并行计算所有时间步。我们方法的主要开销来自为所有掩码计算对数后验概率所需的多次前向传播，这会引入随 K 线性增长的成本。通过将原始输入和掩码输入放在同一批次中进行一次批量前向传播，可以在一定程度上降低这一开销。

使用 ILM 伪对数似然进行推理

对于使用 ILME 进行推理，我们采用与式 (1) 中所述类似的方法，从原始 CTC 对数后验概率中减去使用式 (5) 计算得到的 ILM 伪对数似然。然而，CTC 模型具有特殊的 空白 标记，它仅具备功能性，与任何语言语义都无关。因此，与 [10], 类似，我们会跳过原本预测为空白标记的时间步，不对这些时间步减去 ILM。不过，鉴于 CTC 模型的尖峰行为 [20], ，我们仅在模型于后验域中为该空白标记分配的似然高于某个阈值 β 时，才跳过这些时间步：

$$\text{if } P_{\text{CTC}}(y_t = \langle \texttt{blank} \rangle | \mathbf{X}) < \beta,$$
$$\text{SCORE}(\mathbf{y_t}) = \log P_{\text{CTC}}(\mathbf{y}_t|\mathbf{X}) - \lambda_I \log P_{\text{ILM}}(\mathbf{y}_t) \quad (6)$$
$$\text{else, SCORE}(\mathbf{y_t}) = \log P_{\text{CTC}}(\mathbf{y}_t|\mathbf{X}) \quad (7)$$

最后，我们将这些修改后的对数后验分数 SCORE($\mathbf{y_t}$) 传递给束搜索，以结合浅层融合执行推理。

4. 实验

模型。 我们在基于 Conformer 的模型上进行所有实验 [21], ，该模型 [22] 采用混合方式训练，并联合使用 CTC 和解码器注意力损失，以实现更稳定的训练。推理时，我们仅使用非自回归 CTC 头，并弃用浅层解码器。Conformer 编码器由 20 层 Conformer 块组成，每层包含 8 头注意力和 2048 个隐藏单元，后接一个 512 维的输出投影。浅层解码器是一个隐藏维度为 2048 的单层基于 Transformer 的层。该模型约有 1.4 亿个参数，并使用 ADAM 优化器以 3×10^{-3} 的初始学习率训练 45 个训练轮次。对于分词，我们训练了一个词汇表大小为 2048 的 SentencePiece 分词器。我们的模型实现和训练借助广泛使用的 ESPnet 框架完成。

-

工作 [23]。此外，我们使用流行的 KenLM 库 [24] 中实现的改进型 Kneser-Ney 平滑算法训练四元语言模型，用于浅层融合实验和 ILME。

.

数据。 该模型使用包含 5 万+ 小时配对音频文本数据的大型英语语料库进行训练。我们从公开和内部的配对音频与文本数据中抽取这些数据，确保涵盖多种口音、说话人、采样率和背景噪声。该数据方案可代表适用于各种语音应用的广泛端到端自动语音识别系统。为评估所提出的内部语言模型估计方法，我们选取了来自不同领域的四个评估数据集：LibriSpeech、VoxPopuli、维基百科和华尔街日报。我们在推理实验中使用各数据集的官方测试划分。**LibriSpeech** [25] 是一个基于 LibriVox 有声读物的英语朗读语音语料库。我们考虑两个官方评估集：*test-clean* 和*test-other*，每个集合均包含 5 小时测试音频。**VoxPopuli** [26] 由公开政治演讲组成，取自 2009 至 2020 年欧洲议会活动录音。为进行评估，我们使用 VoxPopuli 英语数据中一个 5 小时的子集。**维基百科** 评估集 [27] 是一个定制的文本转语音（TTS）数据集，包含各种口语专有名词。具体而言，我们从英语维基百科中抽取 1.4 万个句子，并通过 TTS 服务生成由不同说话人录制的合成音频，总时长为 33 小时。《**华尔街日报**》（**WSJ**）[28] 语料库包含记者的常规口述和即兴口述内容。我们选取 0.7 小时的 *test eval92* 测试划分进行评估。

表 2：域外评估数据集摘要。

数据集	领域	# 小时数	# 未登录词	# 唯一
LibriSpeech（test-clean）	有声读物	5.4	232	8138
LibriSpeech（test-other）	有声读物	5.3	265	7597
VoxPopuli	议会	4.9	52	5168
维基百科	百科全书	32.7	9233	38706
WSJ	News	0.7	0	1834

我们的声学模型训练数据与评估集中的数据完全没有重叠。因此，所有 4 个评估集均被视为域外数据。表 2 总结了这些评估集。此外，在训练目标语言模型时，我们分别仅使用 LibriSpeech、VoxPopuli、维基百科和华尔街日报对应的训练拆分。

实验设置。为展示所提方法的泛化能力，我们在所有数据集的实验中使用同一组超参数，并在内部开发集上进行了全面调优。所有实验均使用束宽为 50 的束搜索算法执行 CTC推理。对于语言模型融合，我们使用权重为 $\lambda_T = 1.0$ 的语言模型。在执行 ILME 时，我们将音频划分为 $K = 5$ 个分区，并使用权重为 $\lambda_I = 0.1$ 的内部语言模型进行推理，同时采用 $\gamma = 0.25$ 的对数后验差阈值和 $\beta = 0.9$ 的空白符过滤阈值。

5. 结果

我们展示了所提出的内部语言模型估计方法在多个域外数据集上的有效性，既采用 CTC 束搜索（基线搜索）作为基线，也采用搭配外部语言模型的更强浅层融合推理。我们通过两种模式进行多维度评估：

1. **纯文本领域适应**，即使用在目标域文本上训练的目标域语言模型。这是以往研究中评估 ILME领域自适应效果的典型做法 [10]。
2. **零样本领域自适应**，即仅使用在源域文本上训练的源域语言模型，也就是说，仅使用训练自动语音识别模型时已有的音频转录文本。

请注意，后一种情况是对泛化能力更严苛的测试，因为推理过程不包含任何有关目标域的信息。为进一步校准泛化性能，我们还从每个域外评估数据集中挖掘未登录词（OOV），并评估检测这些训练期间模型从未遇到过的词汇表外词项的 F1 分数。表 2 还显示了在评估数据集中找到的未登录词参考项数量。

5.1. 纯文本领域适应

转录词错误率。 在表 3 中，我们展示了各推理技术对应的词错误率（WER）和相对词错误率降低率（WERR）。对于使用目标 语言模型进行的纯文本领域适应，可以清楚看到，所提出的内部语言模型估计方法在所有情况下都优于基线，相对 WERR 最高可达 28.5%（维基百科）。对于 LibriSpeech test-{clean、other}、维基百科和华尔街日报数据集，ILME 的表现也优于浅层融合方法，相较于浅层融合，词错误率最高降低了 9.4%

表 3：词错误率（越低越好）和相对词错误率降低率（括号内所示）。

	方法	LM	LibriSpeech test-clean	LibriSpeech test-other	VoxPopuli	维基百科	WSJ
	BS	无语言模型	4.0 -	8.1 -	9.9 -	13.0 -	3.4 -
零样本	SF	源语言模型	4.1 (-2.5%)	7.9 (+2.5%)	9.3 (+6.1%)	10.8（+16.9%）	3.6 (-5.9%)
领域自适应	ILME	源语言模型	3.7 (+7.5%)	7.6 (+6.2%)	9.3 (+6.1%)	10.7 (+17.7%)	3.0 (+11.8%)
纯文本	SF	目标语言模型	3.2 (+20.0%)	6.6 (+18.5%)	7.9 (+20.2%)	9.5 (+26.9%)	3.2 (+5.9%)
领域自适应	ILME	目标语言模型	**2.9 (+27.5%)**	**6.5 (+19.8%)**	**7.9 (+20.2%)**	**9.3 (+28.5%)**	**2.9 (+14.7%)**

表 4：使用目标 语言模型检测未登录词的 F1值（越高越好）。

数据集	BS	SF	ILME
LibriSpeech（test-clean）	30.7	32.5	**43.1**
LibriSpeech（test-other）	33.3	27.4	**33.3**
VoxPopuli	31.7	34.4	**44.1**
维基百科	39.4	42.4	**47.6**

表 5：使用源语言LM 检测未登录词的 F1值（越高越好）。

数据集	BS	SF	ILME
LibriSpeech（test-clean）	30.7	36.6	**45.8**
LibriSpeech（test-other）	33.3	32.8	**39.4**
VoxPopuli	31.7	26.2	**34.4**
维基百科	39.4	39.5	**45.2**

（LibriSpeech test-clean 和华尔街日报）。对于 VoxPopuli，尽管 ILME 带来的词错误率改善与浅层融合非常相似，但如表 4 所示，从未登录词 F1 分数来看，使用 ILME 显然更有优势。

用于检测未登录词的 F1 分数。 表 4 展示了在域外数据集上使用目标语言模型检测词汇表外词项的 F1 分数。由于在华尔街日报中没有发现词汇表外词项，因此我们将该数据集排除在此次评估之外。我们可以看到，在多个数据集上，ILME 的表现优于基线搜索和浅层融合：LibriSpeech test-clean、VoxPopuli 和维基百科。与基线方法相比，ILME 将 F1 分数相对提高了最高 40.4%（LibriSpeech test-clean）。对于 VoxPopuli，与基线搜索相比，F1 值相对提升了 39.1%；而浅层融合相比基线搜索仅将 F1 值相对提升了 8.5%。这表明，尽管两者的词错误率非常相近，ILME 仍明显优于浅层融合，是 VoxPopuli 更好的选择。对于 LibriSpeech test-other，我们看到 ILME 的表现与基线搜索相同。然而，观察 LibriSpeech test-other 的浅层融合 F1 分数可以明显发现，目标语言模型实际上降低了未登录词检测性能。这可以归因于 LibriSpeech test-other 的难度是刻意设置得更高的 [25], , 并且我们注意到，LibriSpeech 训练数据对词汇表外词项的表示极其稀疏。因此，单独使用目标语言模型并不能保证改善词汇表外词项的检测效果，因为这些罕见词项不一定会出现在目标域文本中。不过，ILME 仍然能够利用目标语言模型恢复浅层融合所造成的 F1 性能下降，同时改善 LibriSpeech test-other 的词错误率。ILME 通过在每个时间步移除模型偏差来实现这一点，从而使未登录词标记能够相较于此前存在偏差的标记更容易显现。

5.2. 零样本领域自适应

在实际的自动语音识别应用中，通常很难预先确定语音输入所属的具体领域。因此，我们评估了在无法获知目标域，或没有目标域数据可用于训练语言模型的情况下，ILME 的泛化能力。为此，我们使用基于自动语音识别训练转录文本训练的源语言模型。若不使用 ILME 而直接采用源语言模型，内部模型偏差与外部语言模型偏差可能会叠加，从而使输出面临双重偏差的风险。因此，ILME 有助于零样本领域自适应：它能够缓解源域的内部偏差，同时外部源语言模型可以确保保留与语言本身相关的高层语义。

转录词错误率。 表 3 显示，在使用源 语言模型进行零样本领域自适应时，ILME 在所有情况下均优于基线搜索，词错误率降低率最高可达 17.7%（维基百科）。由于此处使用的是源 语言模型，简单的浅层融合未必能取得比基线搜索更好的结果，LibriSpeech test-clean（-2.5%）和华尔街日报（-5.9%）的结果便是如此。这是因为内部语言模型和外部语言模型都可能使推理受到源域数据的双重偏置。在这类情况下，ILME 可以通过消除源域的内部偏置发挥作用。从表 3 可以明显看出，在多个数据集上，仅使用源 语言模型时，ILME 的表现优于浅层融合：LibriSpeech test-{clean、other}、维基百科和华尔街日报。对于 VoxPopuli，ILME 的表现同样与浅层融合相近；但在考虑未登录词 F1 分数时，ILME 再次明显成为最佳选择（表 5）。

用于检测未登录词的 F1 分数。在表 5 中，我们展示了使用源语言模型检测词汇表外词项的 F1 分数。我们看到，在这一零样本领域自适应场景的所有情况下，ILME 都明确优于其他技术，相对提升幅度最高可达 49.2%（LibriSpeech test-clean）。对于 LibriSpeech 数据集，我们看到，使用源语言模型的 ILME 以及浅层融合实际上比目标语言模型更擅长检测词汇表外词项。这是因为无法保证词汇表外词项会出现在目标语言模型的训练数据中，因为 LibriSpeech 从不同的说话人和有声读物分区中抽取训练数据 [25]。对于 VoxPopuli，尽管 ILME 在词错误率方面表现得与浅层融合相近，但 ILME 显然是更好的选择，因为它将 F1 值相对提升了 8.5%，而浅层融合实际上使 F1 值下降了 17.4%。巧合的是，对于 VoxPopuli，使用源语言模型的 ILME 与使用目标语言模型的浅层融合表现同样出色。

5.3. 未来工作

我们观察到，之前针对 AED 和 RNN-T 模型的 ILME 方法仅通过消除编码器输入来执行解码器侧去偏。相比之下，我们基于 CTC 的 ILME 方法利用编码器本身，适用于非自回归解码。因此，我们假设，我们的内部语言模型估计方法也可以扩展到 AED 和 RNN-T 模型，以实现互补的编码器侧去偏。此外，还可以制定后训练技术，用于训练学生模型来估计 CTC 模型的内部语言模型，从而加速 ILME 推理。我们计划在未来的工作中继续探索这些方向。

6. 结论

本文提出了一种针对基于 CTC 的自动语音识别模型的新型内部语言模型估计方法。我们全面评估了该方法适配多个未见目标域的效果。结果表明，在仅有目标领域文本数据可用时，与浅层融合相比，所提出的内部语言模型估计技术可将词错误率最多降低 9.8%，并将未登录词 F1 分数最多提高 24.6%。我们还证明，即使完全无法获取目标域数据，将该方法应用于任何新的目标域也有益处，因为它能够消除编码器的偏差，并直接带来性能提升。

7. 参考文献

[1] S. Toshniwal、A. Kannan、C.-C. Chiu、Y. Wu、T. N. Sainath 和 K. Livescu，“编码器-解码器语音识别中语言模型集成技术的比较”，载于*IEEE*口语技术研讨会（*SLT*），2018年。[2] C. Shan、C. Weng、G. Wang、D. Su、M. Luo、D. Yu 和 L. Xie，“组件融合：为端到端语音识别系统学习可替换的语言模型组件”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2019年。[3] A. Shenoy、S. Bodapati、M. Sunkara、S. Ronanki 和 K. Kirchhoff，“为对话代理中的自动语音识别重打分适配长上下文神经语言模型”，载于 *Interspeech 2021* 论文集，2021年。[4] A. Shenoy、S. Bodapati 和 K. Kirchhoff，“利用跨话语上下文和多任务语言建模为电子商务聊天机器人适配自动语音识别”，载于第四届电子商务与自然语言处理研讨会（*ACL*）论文集，2021年8月。[5] L. Liu、Y. Gu、A. Gourav、A. Gandhe、S. Kalmane、D. Filimonov、A. Rastrow 和 I. Bulyko，“用于语音识别的领域感知神经语言模型”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年。[6] S. Dingliwal、A. Shenoy、S. Bodapati、A. Gandhe、R. T. Gadde 和 K. Kirchhoff，“领域提示：实现自动语音识别系统内存与计算高效的领域适配”，载于*Interspeech 2022* 论文集，2022年。[7] E. McDermott、H. Sak 和 E. Variani，“端到端自动语音识别中的语言模型融合密度比方法”，载于*IEEE*自动语音识别与理解研讨会（*ASRU*），2019年。[8] C. Choudhury、A. Gandhe、X. Ding 和 I. Bulyko，“一种基于似然比的端到端模型领域自适应方法”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年。[9] W. Zhou、Z. Zheng、R. Schl¨uter 和 H. Ney，“关于基于循环神经网络转导器的语音识别中的语言模型集成”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年。[10] Z. Meng、S. Parthasarathy、E. Sun、Y. Gaur、N. Kanda、L. Lu、X. Chen、R. Zhao、J. Li 和 Y. Gong，“面向领域自适应端到端语音识别的内部语言模型估计”，载于*IEEE*口语技术研讨会（*SLT*），2021年。[11] M. Zeineldeen、A. Glushko、W. Michel、A. Zeyer、R. Schl¨uter 和 H. Ney，“探究改进基于注意力的编码器-解码器自动语音识别模型语言模型集成的方法”，*arXiv* 预印本 *arXiv:2104.05544*，2021年。[12] Y. Liu、R. Ma、H. Xu、Y. He、Z. Ma 和 W. Zhang，“通过显式上下文向量学习实现基于注意力的编码器-解码器自动语音识别的内部语言模型估计”，*arXiv* 预印本 *arXiv:2201.11627*，2022年。[13] J. Lee 和 S. Watanabe，“基于 CTC 的语音识别的中间损失正则化”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年。[14] R. Fan、W. Chu、P. Chang 和 J. Xiao，“CASS-NAT：用于语音识别的基于CTC对齐的单步非自回归变换器”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年。

[15] X. Zhang, F. Zhang, C. Liu, K. Schubert, J. Chan, P. Prakash, J. Liu, C.-F. Yeh, F. Peng, Y. Saraf 等，“流式自动语音识别中 LF-MMI、CTC 和 RNN-T 准则的基准评测”，载于 *IEEE*口语技术研讨会（*SLT*），2021年。[16] S. Dingliwal, M. Sunkara, S. Ronanki, J. Farris, K. Kirchhoff 和 S. Bodapati，“CTC语音识别模型的个性化”，载于 *2022*年 *IEEE*口语技术研讨会（*SLT*），2023年。[17] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates等，“Deep Speech：扩展端到端语音识别”，*arXiv*预印本 *arXiv:1412.5567*，2014年。[18] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen 和 M. L. Seltzer，“用于 RNN-T 个性化的深度浅层融合”，载于 *IEEE*口语技术研讨会（*SLT*），2021年。[19] E. Variani, D. Rybach, C. Allauzen 和 M. Riley，“混合自回归转换器（HAT）”，载于 *IEEE*国际声学、语音与信号处理会议（*ICASSP*），2020年。[20] A. Zeyer, R. Schl¨uter 和 H. Ney，“为什么 CTC 会产生尖峰行为？”*arXiv*预印本 *arXiv:2105.14849*，2021年。[21] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu 和 R. Pang，“Conformer：用于语音识别的卷积增强型Transformer”，载于 *Interspeech 2020*论文集，2020年。[22] S. Watanabe, T. Hori, S. Kim, J. R. Hershey 和 T. Hayashi，“用于端到端语音识别的 CTC/注意力混合架构”，*IEEE*信号处理精选主题期刊，2017年。[23] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala 和 T. Ochiai，“ESPnet：端到端语音处理工具包”，载于 *Interspeech 2018*论文集，2018年。[24] K. Heafield，“KenLM：更快、更精简的语言模型查询”，载于第六届统计机器翻译研讨会论文集，2011年。[25] V. Panayotov, G. Chen, D. Povey 和 S. Khudanpur，“LibriSpeech：基于公有领域有声书的自动语音识别语料库”，载于 *IEEE*国际声学、语音与信号处理会议，2015年。[26] C. Wang, M. Rivi`ere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino 和 E. Dupoux，“VoxPopuli：用于表征学习、半监督学习和解释的大规模多语言语音语料库”，载于计算语言学协会第 *59* 届年会暨自然语言处理第 *11* 届国际联合会议论文集，2021年。[27] N. Das, M. Sunkara, D. Bekal, D. H. Chau, S. Bodapati 和 K. Kirchhoff，“聆听、理解与拼写：用于提升未登录词命名实体自动语音识别性能的知识增强型子词建模”，载于 *IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年。[28] J. S. Garofolo, D. Graff, D. Paul 和 D. Pallett，“CSR-I（WSJ0）完整数据集LDC93S6A”，语言数据联盟，1993年。[在线]。网址：https://catalog.ldc.upenn.edu/LDC93S6A