从英语走向更多语言：用于跨语言语音识别的参数高效模型重编程

Chao-Han Huck Yang[*1, 2] *Bo Li*[1] *Yu Zhang*[1] *Nanxin Chen*[1] *Rohit Prabhavalkar*[1] *Tara N. Sainath*[1] *Trevor Strohman*[1]

[1]Google，美国[2]佐治亚理工学院，美国

arXiv:2301.07851v1 [cs.SD] 19 Jan 2023

摘要

本文提出一种基于神经模型重编程的跨语言语音识别参数高效学习框架，可将训练充分的英语自动语音识别（ASR）模型**重新用于**识别其他语言。我们设计了多种辅助神经架构，着重于可学习的预训练特征增强，首次使模型重编程能够应用于自动语音识别。具体而言，我们研究了如何选择基于 Conformer 的 RNN-Transducer 中可训练的组件（即编码器），并将其作为冻结的预训练骨干网络。在涵盖七种语言的多语言 LibriSpeech 语音（MLS）任务上的实验表明，模型重编程仅需使用原始完整自动语音识别模型中 4.2%（270M 中的 11M）至 6.8%（660M 中的 45M）的可训练参数，即可取得具有竞争力的结果；不同语言的平均词错误率（WER）为 11.9% 至 8.1%。此外，我们还发现了多种设置，可使大规模预训练自动语音识别模型在单语和多语种语音识别中均取得成功。与现有的自动语音识别调优架构及其结合自监督损失（例如 w2v-BERT）的扩展方法相比，我们的方法词错误率更低，训练效率也更高。

索引词— 跨语言语音识别、模型重编程、预训练适配和基础语音模型

1. 引言

近年来， [1, 2, 3, 4, 5, 6]大规模自动语音识别架构的发展取得了进展，在英语语音识别任务中展现出良好的前景。此外，采用自监督训练目标的英语自动语音识别模型（如 wav2vec2 [7], w2v-BERT [8], 和 BigSSL [9], ）进一步提升了识别性能，这是对现有基于标注数据的监督式自动语音识别框架的扩展。与此同时，当前神经自动语音识别模型的成功**仍与训练数据的规模有关**；对于非英语和低资源语言，使用中小规模语料库训练大型神经自动语音识别模型，并不总能确保取得有竞争力的结果。当下的自动语音识别数据主要以英语为基础，[9], 如何将训练有素的英语自动语音识别系统（例如 RNN-T [10]）的能力拓展至其他语言 [11]，仍是一个尚待解决的问题，而解决这一问题**将使全球更多终端用户受益**。

此前关于英语自动语音识别模型编码器预训练和微调的研究，已在西日耳曼语（例如英语和德语）、非典型语音识别以及口音语音识别方面取得了一些最新进展。受此前讨论的启发，我们旨在研究如何高效迁移大规模英语自动语音识别模型，以实现单语和多语种语音识别。将大规模自动语音识别应用于移动应用时，一个众所周知的挑战是其内存占用方面的模型复杂度（例如可训练参数数量）。针对新任务或数据集对大规模自动语音识别模型进行调优，往往需要付出高昂的训练成本（例如时间和能耗），这也使得大型预训练语音模型难以部署到移动应用和智能家居语音应用中，尤其是在延迟和能耗方面。

近年来，参数高效学习被视为缓解大型预训练语言模型适配难题的一种潜在解决方案。其目标是通过**仅训练一些小型加性模块**来适配冻结的预训练模型（例如，残差适配器 [16], 神经重编程 [17]和输入提示 [18]）。通过结合参数高效学习，预训练语言模型 [19] (PLM)只需较少的训练时间和计算资源，便能在不同的自然语言处理任务上达到新的最先进性能。

总之，如何利用现有的英语自动语音识别模型推进参数高效学习，是一个有待探索的课题，可助力缺乏充足数据来源（如英语数据）的语音应用。在这项工作中，我们提出三**种自动语音识别模型重编程的具体设计**，采用基于 Conformer [20] 的架构进行跨语言适配。如图 1 所示，我们提出的 **C**onformer-based **A** 自动语音识别 **R**eprogramming（CAR）冻结了大部分可训练神经网络架构（即设为不可训练），仅插入少量可训练模块，以实现参数高效的模型训练。

1.1. 使用冻结的自动语音识别模型进行参数高效学习

我们回顾了采用冻结的自动语音识别模型进行参数高效学习的最新进展，并从神经模型重编程的视角阐明其区别。残差适配器 [16]最初应用于视觉领域 [21]，作为全模型微调的一种计算高效替代方案。Houlsby 等人 [16]进一步改进了残差适配器的设计，在冻结的特征提取器（例如，预训练的 Transformer 层）中对潜在特征采用非线性投影机制。鉴于声学特征编码器是自动语音识别模型的标准组件，近期有多项研究展示了残差适配器 [16]在各种语音任务中的有效性，例如非典型语音 [15], 、多语种语音 [22], 和儿童语音识别 [14]。与此同时，也有相关研究探讨如何基于说话人适配 [23] 和潜在空间对抗重编程领域的潜在特征构建可训练参数。Bapna 和 Firat [24] 表明，自动语音识别中的残差适配器比基于隐藏单元调制的方法有效得多。然而，各种方案之间的联系仍值得深入研究，而模型重编程领域 [17] 已经

*在 Google 研究实习期间完成的工作。

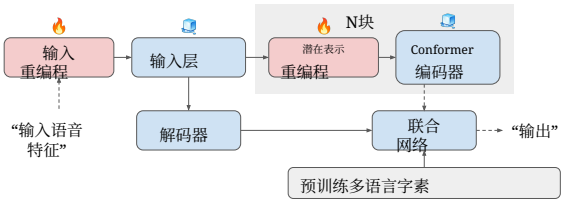


图 1：基于Conformer的自动语音识别重编程的拟议设计流程，其中包括针对声学特征的可训练输入重编程，以及针对潜在表示的可训练特征重编程。

近期首次从总体风险分析的角度，为预训练语音模型适配的成功提供了理论依据。

1.2. 神经重编程：从输入到潜在空间

神经重编程（NR）方法最初于 [25]提出，通过少量可训练输入噪声重新利用冻结的预训练分类器，以进行域外图像预测。最近，NR 已用于序列预测任务的基准测试，例如时间序列信号 [17] 和语音命令 [26]分类。NR 主要在预训练模型的输入层级添加可训练参数，因此在设备端分布式预训练模型中具有联邦优势。与此同时，NR 最近也被引入潜在空间优化 [27]，并展现出最先进的文本分类性能。在下一节中，我们将介绍现有不同参数高效学习方法中的类似组件，并提出一种新的设计，用于总结现有的参数高效学习技术，这些技术能够借助单语模型实现跨语言语音识别。

2. 面向自动语音识别模型的神经重编程

在上述参数高效学习技术取得成功的基础上，我们的神经重编程包含三个主要组成部分：(1) 输入重编程，(2) 潜在空间重编程，以及 (3) 多语言字素预训练。简而言之，(1) 与标准模型重编程或输入提示相关；(2) 关注重编程和残差适配器的有效性；(3) 则旨在解决跨语言学习中现存的字素不匹配挑战。

2.1. 输入层级重编程

给定一个参数 Θ 冻结的预训练神经网络 M，以及一个输入特征 x，我们可以将输入馈入预训练模型 M_Θ，从而获得其预测输出 $y' = \mathrm{M}_\Theta(x)$。输入层级重编程的目标是找到一个可训练的重编程函数 of $\mathcal{R}_\theta$，以最小化 y' 与其真实标签 $\hat{y}$ 之间的预测损失 ($\mathcal{L}_{\text{error}}$)。在此前的语音模型重编程研究 [17, 26]，中，研究者采用了可训练的通用噪声进行跨域适配，这与我们的特征无关重编程目标 w_{θ_2}。然而，在实证研究中，我们发现仅应用通用噪声并不能取得有竞争力的性能，因此我们在式（1）中进一步引入了特征相关的可训练特征提取器（$\mathcal{H}_{\theta_1}$）。

$$\theta^* = \arg\min_\theta \{\mathcal{L}_{\text{error}}(\mathrm{M}_\Theta(\mathcal{R}_\theta(x)), \hat{y})\} \tag{1}$$

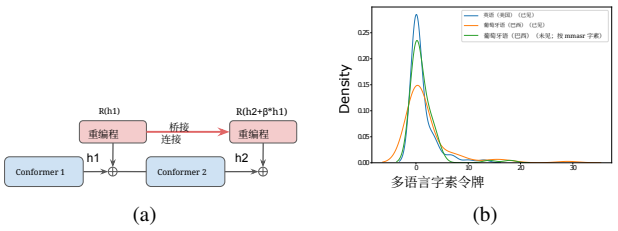
$$\text{where} \quad \mathcal{R}_\theta(x) = \underbrace{x}_{\text{original input}} + \underbrace{w_{\theta_2}}_{\text{feature-independent}} + \underbrace{\mathcal{H}_{\theta_1}(x)}_{\text{feature-dependent}}$$

我们进行了消融研究，并选用简单的 1D 轻量级卷积 [28]，接着采用空间注意力 [29] 编码，作为跨语言自动语音识别的最佳评估提取器配置。

2.2. 带桥接连接的潜在空间重编程

为了进一步提升使用冻结的 Conformer-ASR 模型时的性能，我们在各编码器之间的潜在空间中引入额外的可训练特征。我们将此基线方法称为潜在空间重编程，它可以被视为词级重编程 [27] 和残差适配器 [16] 的一种灵感来源，并在潜在空间中进行优化。我们进一步为潜在空间重编程设计了一种新的“桥接连接”，以增强冻结自动语音识别模型上的加性特征学习。图 2a 展示了桥接连接的重编程模块如何在冻结的 Conformer 编码器之间插入可训练特征。给定第 i 个冻结的 Conformer 编码器作为函数 $\mathcal{F}^i_\Theta$，我们从输入 x中提取第 i 个 Conformer 层的潜在特征 h^i。桥接连接机制已应用于后续的第 $(i+1)$ 个 Conformer 层 $\mathcal{F}^{i+1}_\Theta$，其计算方式对应于公式（2）的第三项，并采用确定性的丢弃率（$\hat{\beta} = 0.15$）。在本研究中，我们使用相同的重编程生成器 $\mathcal{R}_\theta$，为跨语言适配生成加性特征。

$$\underbrace{\mathcal{F}^{i+1}_\Theta\ (h^i)}_{\text{冻结的编码器}} \rightarrow \underbrace{\mathcal{F}^{i+1}_\Theta\ (\mathcal{R}_\theta(h^i))}_{\text{潜在表示重编程}} \rightarrow \underbrace{\mathcal{F}^{i+1}_\Theta\ (\mathcal{R}_\theta(h^i + \hat{\beta} h^{i-1}))}_{\text{桥接连接重编程}} \tag{2}$$



(a) (b)

图 2：（a）桥接连接的重编程机制。（b）多语言字素分布，其中通过英语预训练（绿色）学习判别性信息，以提升泛化能力，使其分布与真实值（橙色）相似。

2.3. 面向多语言的英语字素预训练 数据

为有效地将大规模预训练英语自动语音识别模型适配到不同的语言识别任务（例如，从英语适配到法语），以往的研究工作 [30] 提出通过替换自动语音识别模型的最后一个预测层来解决这一问题。然而，我们发现，采用“多语言字素”甚至比直接替换最终预测头更有效。如图 2b 所示，我们发现，经过英语预训练的多语言字素也能学习一些判别信息，从而处理未见过的语音（例如葡萄牙语）。表 1 展示了对多语言字素在降低分类错误率方面重要性的研究。我们选取了包含 80 个词元的统一多语言字素集合 [4] 作为自动语音识别系统的输出词汇表。值得注意的是，我们发现，使用单语字素（例如仅英语字素）与使用多语言字素之间的性能差距相对较小（下降 ±0.11%）。使用额外全连接层调整 Conformer 模型（F0b 和 F1b），也称为线性探测，其表现不如微调基线（F0 和 F1），且多使用了 4k 个可训练参数。

表 1：使用预训练的 Conformer RNN-T 时，多语言字素预训练对英语（记作 gh_{en}^{multi}）的重要性。

设置（基于英语-自动语音识别）	WER
F0：**微调**所有参数（使用 gh_{en}^{multi}）	**10.5**
F0a：F0，不加载 gh_{en}^{multi}	13.4
F0b：F0，增加全连接层	12.6
F1：**微调**最后一个 Conformer（使用 gh_{en}^{multi}）	**18.2**
F1a：不加载 gh_{en}^{multi} 时的 F1	49.2
F1b：使用额外全连接层时的 F1	21.1

3. 端到端基于 Conformer 的自动语音识别系统

本节介绍我们的端到端自动语音识别系统，并说明模型重编程如何优于其他参数高效学习方案。我们研究了两种端到端自动语音识别系统中的跨语言适配任务：（1）仅使用监督训练损失进行训练；（2）使用联合监督与无监督损失（例如结合 w2v-BERT）。我们的多语言自动语音识别模型采用基于 Conformer 的 RNN-T 架构。我们主要在 RNN-T 上开展参数高效学习实验，其研究发现也可推广至其他编码器-解码器自动语音识别预训练方案，供未来研究参考。

3.1. 系 第 1 部分：基于 Conformer 的监督训练 ASR

我们根据先前的工作 [4], 选择预训练 Conformer RNN-T 系统。该系统仅采用监督训练目标，便取得了具有竞争力的多语言识别性能。Conformer RNN-T 包含一个编码器网络、一个解码器网络和一个联合预测网络。对于编码器，我们采用全上下文 Conformer 层，其中包括一个输入投影层、一个相对位置嵌入层，后接由 17 个 Conformer 层组成的堆栈。

与 [4] 类似，堆叠的 Conformer 层可分为三个编码器块。第一个块由 4 个 Conformer 层和一个用于计算时间缩减的时间堆叠层组成；第二个块由一个 Conformer 层（第五个 Conformer 层）和一个将当前特征维度映射回原始维度的投影层组成。其余 12 个 Conformer 层构成第三个编码器块。我们使用 Lingvo [31] 工具包中现有的卷积模块，以支持每个 Conformer 层中的相对位置信息和组归一化。解码器采用了一个 2 层单向 LSTM 网络。该监督式 Conformer 自动语音识别模型总共包含 270M 个可训练参数。

为实现参数高效学习，我们首先使用 MLS [32]中的 44.6k 小时英语数据训练模型。使用第 2.3 节讨论的上述多语言字素后，Conformer 自动语音识别模型取得了具有竞争力的 5.5% 词错误率。随后，我们将大部分参数设为不可训练（例如，如图 1 所示将其冻结），仅训练从预训练英语 Conformer RNN-T 中加载的重编程层。我们借鉴现有的残差适配器研究开展了消融研究，以论证我们的设计。

3.2. 系统 2：采用自监督损失的监督式自动语音识别

我们进一步探究了这样一种方案：通过无监督预训练增强的先进监督式自动语音识别模型，即使可训练参数预算很小，也能训练出较高的跨语言识别性能。我们采用先前的无监督与监督联合训练（JUST）自动语音识别模型 [33]，将监督 RNN-T 损失与自监督学习（SSL）的（i）对比学习损失和（ii）掩码语言建模（MLM）损失结合起来。在 JUST 自动语音识别模型中，对比网络和 MLM 网络这两个自监督模块分别由 8 层和 16 层标准 Conformer 层堆叠而成。按照 [33], 中的基本设置，我们冻结这些 Conformer 层，并在模型训练期间于层间插入所提出的带桥接连接的重编程层。预训练 JUST 自动语音识别模型的总损失定义为 $\mathcal{L}_{\text{JUST}} = \mathcal{L}_{\text{rnn-t}} + \gamma(\mathcal{L}_{\text{c}} + \mathcal{L}_{\text{mlm}} + \alpha\mathcal{L}_{\text{div}})$，其中 $\mathcal{L}_{\text{rnn-t}}$ 是监督损失，如第 3.1 节所述；γ 是自监督学习损失的联合系数（设为 0.01）；$\mathcal{L}_{\text{c}}$ 是对比网络的损失；$\mathcal{L}_{\text{mlm}}$ 是 MLM 网络的损失；基于熵的多样性损失（$\mathcal{L}_{\text{div}}$）用于码本相关的优化，其中 $\alpha = 0.1$ 沿用了先前 wav2vec2 研究 [7] 中类似的实现。

4. 实验与结果

本节介绍三种跨语言自动语音识别任务中的基本设置和所提出的参数高效重编程方法。需要注意的是，表 2 和表 4 中的 Para. 表示可训练模型参数。

4.1. 设置与参数高效架构

数据集： 我们在广受欢迎的多语言 LibriSpeech [32] (MLS) 基准上开展实验，涵盖八种语言，包括：(1) 日耳曼语族：英语 (en)，44.6k 小时；德语 (de)，1.96k 小时；荷兰语 (nl)，1.55k 小时；(2) 罗曼语族：法语 (fr)，1.1k 小时；西班牙语 (es)，0.9k 小时；意大利语 (it)，0.2k 小时；葡萄牙语 (pt)，0.16k 小时；以及 (3) 西斯拉夫语族：波兰语 (pl)，0.1k 小时。MLS 是一套得到广泛研究的标准评测集。我们使用 MLS 官方测试集报告各项设置下的测试词错误率 (WER) 性能。

特征和自动语音识别系统： 采用从 MLS 声学语句中提取的对数梅尔滤波器组特征（80 维）作为输入（例如，时长为 10–20 秒）。第 3.1 节介绍了有监督 Conformer-ASR 的编码器层（研究 1）， [4]中也有相关细节。对于基于 JUST 的 Conformer 自动语音识别，Contrastive 网络和 MLM 网络均有 1024 个隐藏单元和 8 个注意力头。MLM 的样本掩码比例为 6.5%，码本大小为 1024。JUST 的更多相关细节请参阅 [33]。

可训练参数预算和超参数： 我们谨慎控制重编程和其他辅助调优基线，使其参数量大致相当，并报告表现最佳的设置。近期一项内存高效自动语音识别研究 [34] 探讨了相同规模的 ∼10M 至 30M 可训练参数。我们的最终目标是找到最适合适配大规模冻结预训练自动语音识别模型的架构。我们使用 64 个张量处理单元、批量大小为 1024 来训练有监督 Conformer-ASR；使用 256 个张量处理单元、批量大小为 1024 来训练 JUST-ASR。对于梯度训练，建议采用 Adam 优化器 $\beta_1 = 0.9$, $\beta = 0.98$。对于 JUST 训练，我们使用 [33]中所述的全局学习率调度器。对于残差适配器，我们在消融实验后采用 [16, 15] 中的基准设计，并将潜在表示维度设为 256。

4.2. 跨语言语音识别结果

我们使用冻结的有监督Conformer-ASR（第 3.1 节中的系统 1）评估不同架构及其在跨语言识别任务中的表现。找到一种最佳的参数高效学习方案后，我们进一步评估该方案能否使用自监督学习损失继续训练有监督 Conformer 模型（第 3.2 节中的系统 2）。

研究1：**基于英语预训练的单语语音识别** 我们首先考察单语语音识别的结果。为进行评估，我们分别从头开始使用七种非英语数据（西班牙语、意大利语、葡萄牙语、法语、德语、荷兰语、波兰语）进行训练，并报告10次运行后的平均词错误率。根据

表2：**C**onformer-**A**SR **R**eprogramming（CAR），用于研究1。

设置	平均词错误率	参数$_{train}$
B0：**B**aseline 冻结的英语 Conformer	92.1	0
B1：从头开始训练 Conformer	10.7	270M
B2：Wav2Letter [32]	11.8	100M
B3：XLSR-53（使用外部数据） [35]	10.6	300M
F0：**F**从英语微调	10.5	270M
F1：微调最后一个 Conformer	18.2	13M
F2：添加一个额外的 Conformer	34.9	13M
F3：来自英语的残差适配器	13.5 ± 0.2	11M
F4：微调解码器	20.9	20M
F5：偏置项微调（BitFit [36]）	33.0	0.2M
CA**R**1：注意力-**重编程**	12.6 ± 0.4	11M
CAR2：卷积重编程	13.0 ± 0.2	11M
CAR3：CAR1 + 桥接连接	**11.9** ± 0.3	11M

结果，直接使用冻结的、基于英语预训练的 Conformer-ASR，词错误率应会高于90%。如表2第5行（F0）至第9行（F4）所示，基于残差适配器的方法在七种语言上的平均词错误率较低，为13.5%。无论是（F1）直接微调 Conformer 编码器的最后一层，还是（F2）在编码器后额外训练一层 Conformer，调整 Conformer 层本身的效果都不如残差适配器。值得注意的是，我们还进行了逐层微调实验，但微调最后一个 Conformer 层的表现优于微调其他单个 Conformer 层。同时，我们还发现，冻结自动语音识别模型时，性能提升**主要来自对编码器的微调**；直接微调 RNN-T 的解码器 （F4）则会使词错误率超过20%。类似地，仅微调偏置项 （BitFit）[36]（F5）的词错误率为33%。根据我们在 MLS 上的实证评估，这种方法在基于 RNN-T 的自动语音识别建模中仍未展现出竞争力。

接下来，我们研究所提出的 Conformer-ASR 重编程（CAR）方案中架构差异的影响。我们首先考察加性特征生成器之间的差异。我们发现，基于一维卷积的特征提取器与残差适配器的性能相近。我们使用空间注意力对输入进行编码，并对声学特征应用二维卷积，以减少可训练参数。与基于卷积的消融方案相比，该设置的性能有所提升，相较于残差适配器，词错误率相对降低了 3.7%。我们在表现最佳的基于注意力的重编程设置中加入桥接连接，使词错误率相对降低了 5.6%。由于桥接连接在骨干模型之外提供了额外的梯度路径（位于每个重编程层之间），与残差适配器相比，我们在模型训练期间的计算时间**减少了** 10.1% [22]。我们选择带有桥接连接的基于注意力的重编程作为最佳设置（记为 reprogram$_{CAR3}$），以进一步研究其他适配特性。

研究2**：基于多语言自动语音识别预训练进行调优**

“多语言”预训练能否成为更好的预训练自动语音识别骨干网络形式，是跨语言自动语音识别领域的一个开放问题。在表3中，我们通过仔细控制相同有监督Conformer-ASR的总训练时长（4.4万至4.5万小时），并采用不同的额外混合语言“英语+法语”或“英语+西班牙语”，来探讨这一问题。随后，我们报告该模型在德语、荷兰语、意大利语、葡萄牙语和波兰语这五种未见语言上的平均测试词错误率。有趣的是，与仅使用英语的 Conformer-ASR相比，这两种多语言自动语音识别骨干网络的表现都更好，词错误率绝对值降低了0.5%至1.6%。

Study3**：使用自监督损失进行 ASR 微调** **ses**

表3：**多**对单自动语音识别（研究2），总时长相近。

设置 \ 预训练语言	en	英语+法语	英语+西班牙语
M0：全部微调	11.3	**9.7**	10.1
M1：残差适配器	13.9	**12.5**	12.9
M2：重编程$_{CAR3}$	12.8	**11.8**	12.3
总训练时长	44.6k	45.6k	**45.7k**
已覆盖字素数（共 80 个）	29	**51**	42

由于 JUST-ASR 包含基于 Libri-Light [37] 中 6 万小时未标注英语语音进行的 w2v-BERT 预训练，因此我们按照 [33]中的设置，报告其在英语以及同时识别多语种语音时的七种语言上的词错误率表现。在我们的参数高效研究中，我们仅训练 (i) 所提出的桥接重编程层，以及 (ii) 原始解码器层，在少样本学习设置下使用张量处理单元进行仅 10 万步更新，以更新第 3.2 节讨论的无监督损失。另一个发现是，无论是进行重编程，还是在完全无监督的 w2v-BERT [8] 上使用适配器，词错误率都会高达 70% 以上，这表明监督损失至关重要。如表 4 所示，与微调 JUST (J0) 相比，将微调与重编程模块结合使用 (J1)，能够降低英语（-2.7%）和未见语言（-11.6%）的词错误率。我们已确认，与仅微调解码器 (J4) 相比，参数高效方案 J2 和 J3 的性能提升**主要来自重编程或适配器**。

表 4：面向多语言自动语音识别的 JUST 重编程（研究3）

设置	en	7 种语言	参数
J0：微调（FT）**J**UST [33]	7.5	9.5	660M
J1：微调 JUST +重编程$_{CAR3}$	**7.3**	**8.4**	710M
J2：重编程$_{CAR3}$ +微调解码器	**8.4**	**9.7**	45M
J3：适配器 +微调解码器	8.9	10.2	45M
J4：仅微调解码器	17.4	22.0	20M

局限性与补充讨论尽管当前的 Conformer-ASR 重编程在不同的跨语言评估中表现良好，但我们希望提醒读者，如何适配具有大量字素标记的语言（例如普通话和日语）仍然是一个挑战。一种潜在的解决方案是使用字典学习来组合音素标记，这将留待未来的研究。请注意，当前用于语音的提示调优 [18] 主要涵盖跨任务适配，而对于自动语音识别研究而言，输入重编程可能是这一方向上的一种类似方法。

5. 结论

本文提出了一种面向跨语言自动语音识别的全新参数高效学习方案。我们提出的模型重编程模块通过在输入空间和潜在空间中添加可训练属性来实现其设计，并由此构建出一种轻量级方案，用于适配大规模预训练自动语音识别模型。对于有监督自动语音识别模型，我们仅需 11M 个可训练参数（占完整预训练模型的 4.8%），便可在 MLS 基准的七种语言上实现 11.9% 的词错误率。结合自监督学习设置时，模型重编程展现出竞争力，词错误率分别为 8.4% 和 9.7%；例如，迁移一个 660M 的预训练模型时，仅使用其原始模型参数的 6.8%。我们提出的方法以及关于跨语言识别的新发现，可视为未来研究中设计大型“基础语音模型”的初步途径之一。

6. 参考文献

[1] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, 等，“面向移动设备的流式端到端语音识别”，载于*Proc.ICASSP*。IEEE，2019年，第 6381–6385 页。[2] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, 和 B. Ramabhadran，“用于大规模语音任务的深度卷积神经网络”，神经网络，第 64 卷，48，2015年，第 39 页。[3] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, 等，“基于序列到序列模型的先进语音识别”，载于*Proc. ICASSP*。IEEE，2018年，第 4774–4778 页。[4] B. Li, R. Pang, T. N. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. R. Huang, M. Ma, 和 J. Bai，“扩展端到端模型以支持大规模多语言自动语音识别”，载于*2021*年*IEEE*自动语音识别与理解研讨会 *(ASRU)*。IEEE，2021年，第 1011–1018 页。[5] B. Li, R. Pang, Y. Zhang, T. N. Sainath, T. Strohman, P. Haghani, Y. Zhu, B. Farris, N. Gaur, 和 M. Prasad，“超大规模多语言自动语音识别：一种终身学习方案”，载于*Proc. ICASSP*。IEEE，2022年，第 6397–6401 页。[6] B. Li, S.-y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohman, 和 Y. Wu，“迈向快速准确的流式端到端自动语音识别”，载于*Proc. ICASSP*。IEEE，2020年，第 6069–6073 页。[7] A. Baevski, Y. Zhou, A. Mohamed, 和 M. Auli，“wav2vec 2.0：一种用于语音表征自监督学习的框架”，神经信息处理系统进展，第 33 卷，第 12 449–12 460 页，2020年。[8] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, 和 Y. Wu，“W2v-bert：结合对比学习与掩码语言建模进行自监督语音预训练”，载于*2021*年*IEEE*自动语音识别与理解研讨会 *(ASRU)*。IEEE，2021，第 244–250 页。[9] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, 等，“Bigssl：探索自动语音识别大规模半监督学习的前沿”，*IEEE*信号处理专题期刊，2022年。[10] A. Graves，“使用循环神经网络进行序列转导”，表征研讨会，*ICML*，2012年。[11] K. Hu, A. Bruguier, T. N. Sainath, R. Prabhavalkar, 和 G. Pundak，“端到端模型中面向跨语言语音识别的基于音素的上下文化”，*Proc. Interspeech 2019*，第 2155–2159 页，2019年。[12] S. Tong, P. N. Garner, 和 H. Bourlard，“对多语言语音识别训练与适应中深度神经网络的研究”，载于*Proc. of I NTERSPEECH*，第 CONF 期，2017年。[13] S. Bansal, H. Kamper, K. Livescu, A. Lopez, 和 S. Goldwater，“在高资源语音识别任务上的预训练有助于低资源语音转文本翻译”，载于 *NAACL* 会议论文集，2019年。[14] R. Fan, Y. Zhu, J. Wang, 和 A. Alwan，“探索自监督模型的更优领域适应：儿童自动语音识别案例研究”，*IEEE*信号处理专题期刊，第 1242–1252 页，2022年。[15] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, 和 F. Biadsy，“用于适应非典型语音和带口音语音的参数高效自动语音识别残差适配器”，载于*2021*年自然语言处理实证方法会议论文集，2021年，第 6751–6760 页。[16] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, 和 S. Gelly，“面向自然语言处理的参数高效迁移学习”，载于国际机器学习会议。PMLR，2019年，第 2790–2799 页。[17] C.-H. H. Yang, Y.-Y. Tsai, 和 P.-Y. Chen，“Voice2series：对声学模型进行重编程以实现时间序列分类”，载于国际机器学习会议。PMLR，2021年，第 11 808–11 819 页。[18] K.-W. Chang, W.-C. Tseng, S.-W. Li, 和 H.-y. Lee，“探索生成式口语语言模型的提示调优在语音处理任务中的应用”，载于*Proc. Interspeech 2022*，第 5005–5009 页。

[19] J. D. M.-W. C. Kenton 和 L. K. Toutanova，“BERT：面向语言理解的深度双向变换器预训练”，发表于 *NAACL-HLT* 论文集，2019年，第 4171–4186 页。[20] A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu，等，“Conformer：用于语音识别的卷积增强型变换器”，*INTERSPEECH 2020* 论文集，第 5036–5040 页，2020年。[21] S.-A. Rebuffi、H. Bilen 和 A. Vedaldi，“利用残差适配器学习多个视觉领域”，神经信息处理系统进展，第 30 卷，2017年。[22] A. Kannan、A. Datta、T. N. Sainath、E. Weinstein、B. Ramabhadran、Y. Wu、A. Bapna、Z. Chen 和 S. Lee，“采用流式端到端模型的大规模多语种语音识别”，*INTERSPEECH 2019* 论文集，第 2130–2134 页，2019年。[23] P. Swietojanski、J. Li 和 S. Renals，“学习隐藏单元对无监督声学模型自适应的贡献”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 24 卷，第 8 期，第 1450–1463 页，2016年。

[24] A. Bapna, 和 O. [illegible]，“用于神经机器翻译的简单、可扩展适配”，载于 *EMNLP-IJCNLP* 会议论文集，2019年，第 1538–1548 页。

[25] G. F. Elsayed、I. Goodfellow 和 J. Sohl-Dickstein，“神经网络的对抗式重编程”，载于国际学习表征会议，2018年。

[26] H. Yen、P.-J. Ku, C.-H. H. Yang, H. Hu, S. M. Siniscalchi, P.-Y. Chen, 和 Y. Tsao，《基于对抗性重编程的低资源语音命令识别研究》，*arXiv* 预印本 *arXiv:2110.03894*，2021。

[27] K. Hambardzumyan, H. Khachatrian, and J. May, “Warp: Word-级对抗性重编程，”载于*ACL* 会议论文集，2021，第 4921–4933 页。

[28] F. Wu, A. Fan, [illegible], D. Grangier, E. Grave, and M. Auli，“通过轻量级动态卷积减少注意力开销”，载于国际学习表征会议，2018年。

[29] L. Chen、H. Zhang、J. Xiao、L. Nie、J. Shao、W. Liu 和 T.-S. Chua，“SCA-CNN：用于图像描述的卷积网络中的空间注意力与通道注意力”，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第5659–5667页。[30] R. Zhao、J. Xue、J. Li、W. Wei、L. He 和 Y. Gong，“关于应对循环神经网络转导器的实际挑战”，载于*2021*年*IEEE*自动语音识别与理解研讨会（*ASRU*）。IEEE，2021，第526–533页。[31] J. Shen、P. Nguyen、Y. Wu、Z. Chen、M. X. Chen、Y. Jia、A. Kannan、T. Sainath、Y. Cao、C.-C. Chiu、等，“Lingvo：用于序列到序列建模的模块化可扩展框架”，*arXiv*预印本 *arXiv:1902.08295*，2019年。[32] V. Pratap、Q. Xu、A. Sriram、G. Synnaeve 和 R. Collobert，“MLS：面向语音研究的大规模多语言数据集”，载于*INTERSPEECH*，2020年。[33] J. Bai、B. Li、Y. Zhang、A. Bapna、N. Siddhartha、K. C. Sim 和 T. N. Sainath，“面向多语言自动语音识别的无监督与监督联合训练”，载于*ICASSP*论文集，2022年，第6402–6406页。[34] G. Venkatesh、A. Valliappan、J. Mahadeokar、Y. Shangguan、C. Fuegen、M. L. Seltzer 和 V. Chandra，“智能设备上的内存高效语音识别”，载于*ICASSP*论文集，2021，第8368–8372页。[35] A. Conneau、A. Baevski、R. Collobert、A. Mohamed 和 M. Auli，“面向语音识别的无监督跨语言表征学习”，载于*Interspeech*论文集，2020年。[36] E. B. Zaken、Y. Goldberg 和 S. Ravfogel，“BitFit：面向基于Transformer的掩码语言模型的简单参数高效微调”，载于*NAACL*论文集，2022年，第1–9页。[37] J. Kahn、M. Rivi`ere、W. Zheng、E. Kharitonov、Q. Xu、P.-E. Mazar´e、J. Karadayi、V. Liptchinsky、R. Collobert、C. Fuegen、等，“Libri-light：面向有限监督或无监督自动语音识别的基准”，载于*ICASSP*论文集。IEEE，2020年，第7669–7673页。