arXiv:2603.11378v1 [cs.SD] 11 Mar 2026

低资源斯瓦希里语自动语音识别的持续预训练：仅用少量标注数据实现最先进水平

Hillary Mutisya JohnMugane
Thiomi-LughaNLP 哈佛大学

摘要

我们研究了持续预训练（CPT），用于将 wav2vec2-bert-2.0 适配到斯瓦希里语自动语音识别（ASR）。我们的方法结合未标注音频与少量标注数据，先进行伪标注持续预训练，再进行监督微调。使用 20,000 个标注样本，我们在 Common Voice 斯瓦希里语上取得了 3.24% 的词错误率（WER），相较基线相对提升 82%。这一结果超过了此前报告的最佳学术系统（基于 XLS-R，WER 为 8.3%），相对提升达 61%。我们提供了具体的数据需求和可复现的方法论，该方法也适用于其他低资源语言。

关键词： 自动语音识别　、低资源语言、持续预训练、斯瓦希里语、非洲语言

1 引言

1.1 数据挑战

斯瓦希里语是非洲使用最广泛的语言之一，使用者超过 1 亿，但它面临着几乎所有非洲语言共同面临的根本性挑战：高质量标注语音数据匮乏。英语等高资源语言拥有数万小时的专业转写语音，而斯瓦希里语和其他非洲语言必须依靠少得多的标注数据，才能实现具有竞争力的自动语音识别。数据匮乏严重阻碍了服务非洲语言社群的语音技术发展。

wav2vec 2.0 和 wav2vec2-bert-2.0 等自监督基础模型通过支持从未标注音频中学习，改变了这一局面。然而，一个重要问题仍未解决：如何最有效地结合易于获取的未标注音频和有限的标注数据，为低资源语言实现高质量的自动语音识别？

1.2 持续预训练假设

持续预训练（CPT）提供了一种潜在的解决方案：在使用标注数据进行监督微调之前，先利用未标注音频使基础模型适应目标语言。其假设很简单：如果能充分利用丰富的未标注音频，我们或许只需远少于传统监督方法所需的标注数据，就能取得出色的性能。

然而，近期研究表明，持续预训练（CPT）的效果并不一致，尤其是在低资源场景中。一些研究报告了显著的改进，另一些则观察到性能下降或收益甚微。具体而言，对于斯瓦希里语——一种资源适中且在基础模型预训练中有所体现的语言——持续预训练（CPT）的价值仍未得到实证验证。

1.3 研究问题

本研究围绕三个基本问题展开：

Q1：将持续预训练应用于伪标注的无标注数据，并结合有限的标注数据，是否能够提升斯瓦希里语自动语音识别的性能？我们系统比较使用和未使用持续预训练训练的模型，以单独评估其贡献。

Q2：标注数据量如何影响持续预训练的效果？我们分别测试极少量（5K 个样本）和适量（20K 个样本）的标注数据规模，以了解持续预训练的收益是否会随监督水平而变化。

Q3：通过最优结合无标注数据和标注数据，能够达到怎样的性能水平？我们建立具体的词错误率基准，证明在资源有限的情况下实现高质量斯瓦希里语自动语音识别的可行性。

1.4 主要贡献

我们的研究作出了以下贡献：

1. **首次系统评估斯瓦希里语伪标注持续预训练**：我们证明，使用基线模型为未标注音频生成标签后进行持续预训练，再进行监督微调，可取得最先进水平的结果。

2. **斯瓦希里语自动语音识别达到最先进水平**：仅使用 20K 个标注样本，我们在 Common Voice 斯瓦希里语数据集上实现了 3.24% 的词错误率，较已发表研究中最佳学术基线（XLS-R 的 8.3% [1]）相对提升了 61%。

3. **部署所需的具体数据量**：我们证实，将 ∼20K 个标注样本（∼11 小时）与未标注音频结合，即可满足许多实际应用对高质量斯瓦希里语自动语音识别的需求。

4. **低资源语言的实用方法论**：我们提供了一套可复现的训练流程，将易于获取的未标注音频与规模适中的标注数据集相结合，为其他资源匮乏的语言提供了可行路径。

1.5 文章结构

本文其余部分安排如下：第2节回顾持续预训练和斯瓦希里语自动语音识别的相关工作。第3节介绍我们的实验方法论，包括伪标签方法。第4节报告斯瓦希里语在两种数据规模下的结果。第5节讨论其影响，并提供实用指南。第6节总结全文并展望未来方向。

2 相关工作

2.1 自监督模型

自监督学习在语音领域的发展彻底改变了低资源语言的自动语音识别。Wav2vec 2.0 [2] 开创了在未标注音频上进行对比学习的方法，通过区分被掩蔽的语音帧与干扰项来学习表征。XLS-R [1] 利用 436,000 小时的未标注语音，将这一方法扩展至 128 种语言，但非洲语言的表征仍仅涵盖 3 种语言。

Wav2vec2-bert-2.0 [3]代表当前最先进水平，结合了 wav2vec 2.0 与 WavLM 架构，并在涵盖 104 种语言的 450 万小时数据上进行训练。对我们的工作至关重要的是，该模型的预训练数据包含斯瓦希里语，从而提供了一个强基线，供我们的持续预训练方法在此基础上进一步构建。

2.2 持续预训练

持续预训练将自监督学习加以扩展，以使预训练模型适应特定领域或语言。在领域适应方面，Hsu等人 [4] 证明，在领域内未标注数据上进行持续预训练可以改善后续微调，即使微调数据来自不同领域。Attia等人 [5] 通过对课堂音频进行持续预训练，使课堂自动语音识别的词错误率降低了 10%。

在语言适应方面，结果则不尽相同。Nowakowski等人 [6] 发现，持续预训练是将 XLSR-53 适应于阿伊努语（这种语言已濒临灭绝，拥有 <5 小时标注数据）“显然最有效的方法”，词错误率相对降低了 40%。然而，其他研究报告称，持续预训练带来的益处微乎其微，甚至会造成性能下降，尤其是在目标语言具有一定基础模型表征，或持续预训练数据质量较差时。

当纯自监督目标不可用或表现不佳时，面向持续预训练的伪标签已成为一种实用方法。基线基础模型为未标注音频生成转录文本，随后将其用作持续预训练的目标。成功与否关键取决于基线模型的质量——高质量伪标签能够实现有效适应，而噪声标签则可能引入系统性错误。

2.3 斯瓦希里语自动语音识别基准

已发布的斯瓦希里语自动语音识别基准确立了我们旨在提升的性能水平。学术系统包括：在 Common Voice 上达到 8.3% 词错误率的 XLS-R 微调模型 [1], 在 Common Voice 上达到 9–12% 词错误率的 SpeechBrain wav2vec2 模型 [7], 以及报告 8–10% 词错误率的各种开源实现。

我们的 3.24% 词错误率相比这些学术基准有了显著提升，展示了持续预训练对于低资源斯瓦希里语自动语音识别的有效性。

2.4 本文的定位

我们的工作为现有文献作出了以下几项独特贡献：

1. **首次对斯瓦希里语进行系统性的持续预训练评估**：尽管 XLS-R 及其他模型会在斯瓦希里语数据上进行微调，但此前没有研究使用带控制的实验和伪标注数据，系统评估持续预训练所作出的贡献。

2. **易于获取的未标注数据**：我们证明，未标注音频经性能可靠的基线模型进行伪标注后，可作为有效的持续预训练数据——这是一种可扩展的方法，适用于多种语言。

3. **具体的数据需求**：我们不再报告任意数据规模下的性能，而是确定了与持续预训练结合时，实现指定质量水平所需的具体标注数据量（5K 和 20K 个样本）。

3 方法论

3.1 实验设计

我们的实验设计评估了在两种标注数据规模（5K 和 20K 个样本）下持续预训练的有效性。

持续预训练流程：

1. 在可用的已标注数据上训练标注模型　　data

2. 使用标注模型为未标注音频生成伪标签 3. 在伪标注数据上进行持续预训练

4. 在标注数据上进行监督微调

对比基线：为量化持续预训练的贡献，我们单独训练了一个基线模型，即在不进行任何持续预训练的情况下，直接在5万个标注样本上对 wav2vec2-bert-2.0 进行微调。该基线的词错误率为17.71%，并作为评估持续预训练方法的参考点。

该设计确保性能差异可明确归因于持续预训练，而非不同超参数、数据集或评估方案等混杂因素。

3.2 模型架构

我们使用 facebook/w2v-bert-2.0 作为基础模型——这是一个拥有6.07亿参数的架构，融合了 wav2vec 2.0 与 WavLM 的最佳要素。该模型由三个主要组件组成，如图1所示：

1. **卷积编码器**：7层卷积神经网络处理16kHz音频，并生成50Hz潜在表示 2. **Transformer 编码器** [8]：24层通过自注意力机制为每个位置提供约一秒的上下文 3. **输出表示**：768维上下文化表示，同时捕捉声学与语言模式

该模型在涵盖 104 种语言（包括斯瓦希里语）的 450 万小时多语言音频上进行了预训练。预训练采用掩码预测与对比学习，使模型无需转录文本即可学习稳健的语音表征。为使该模型适用于自动语音识别任务，我们添加了一个连接时序分类（CTC） [9] 输出头——一个将 768 维表征映射到斯瓦希里语字符词表的线性投影层。

3.3 数据集

3.3.1 标注数据：CommonVoice

我们的有标注训练和评估数据来自 Mozilla Common Voice 16.0 [10, 11], , 这是一个由社区提供、包含经验证转录文本的数据集。完整数据集包含约 200 小时经验证的音频，我们采用说话人不重叠划分，留出 1,000 个样本用于评估。在持续预训练实验中，我们使用了 5,000 个样本（~3 小时）和 20,000 个样本（~11 小时）。作为对比，我们还使用 50,000 个样本（~28 小时）训练了一个基线模型。该数据集汇集了 1,000 多位不同说话人的贡献，内容来自维基百科及其他多种文本来源。

3.3.2 未标注数据

为进行持续预训练，我们收集了足量的未标注斯瓦希里语音频，以支持 5,000 和 20,000 个样本的实验。我们使用标注模型为这些未标注音频生成伪标签。为确保伪标签质量，我们仅保留模型置信度超过 75% 的片段，并舍弃置信度较低的输出。

未标注音频包含多种领域、说话人和录音条件下丰富多样、自然的斯瓦希里语语音，因此能够代表真实世界部署场景。

3.4 训练流程

我们的训练遵循三阶段流程：

第一阶段：标注模型 我们从预训练的 wav2vec2-bert-2.0 模型开始，添加一个针对斯瓦希里语字符词表配置的 CTC 头，并使用带标签的 Common Voice 样本进行微调。我们采用 1e-4 的学习率、8 的批大小，并训练 15 个训练轮数，同时使用早停。此标注模型用于在第二阶段为未标注音频生成伪标签。

第二阶段：持续预训练 我们使用标注模型，通过贪心CTC解码为未标注的斯瓦希里语音频生成伪标签，并仅保留置信度分数高于 75% 的样本。这种筛选方式可确保伪标签提供有用的训练信号，同时限制噪声。随后，我们使用这些伪标签作为监督目标，继续对 wav2vec2-bert-2.0 进行预训练。

为避免灾难性遗忘，我们采用保守超参数进行训练：学习率 5e-5，仅训练 3 个训练轮数，批大小为 8，预热比例为 10%，并使用 AdamW优化器 [12]，权重衰减为 0.01。

阶段 3：监督微调 我们从 CPT检查点 开始，使用带标签的 Common Voice 数据（5K 或 20K 个样本）进行微调。我们采用 1e-4 的学习率、8 的批大小，并根据配置训练 10–15 个训练轮数。我们采用多种正则化技术，包括标签平滑（0.1）、dropout 和梯度裁剪，并通过早停监控验证词错误率（耐心值为 3）。

4 结果

4.1 主要发现

表 1 展示了我们在所有训练阶段和配置下的完整结果。

这些结果呈现出几个重要发现：

- 两种配置均显著优于对比基线，其中 20K+持续预训练的词错误率达到 3.24%

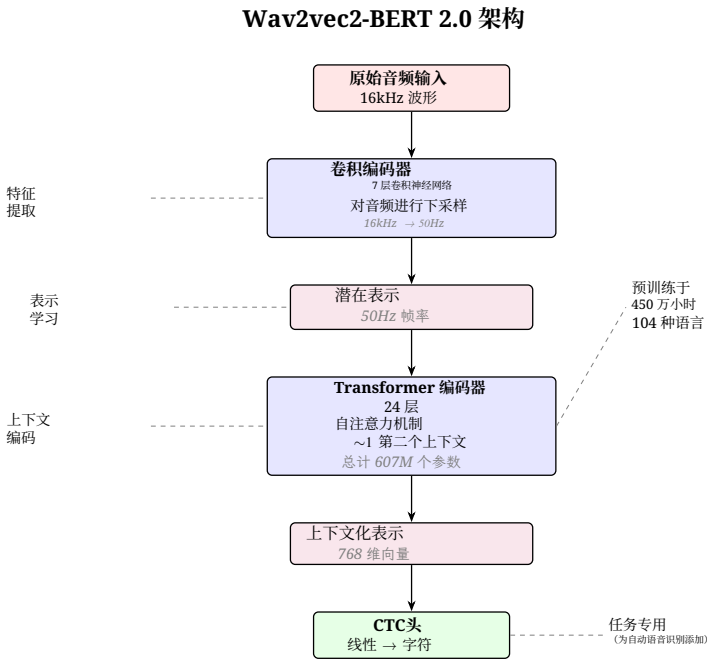


图 1：Wav2vec2-BERT 2.0 架构，展示了从原始音频输入到卷积特征提取、基于 Transformer 的上下文编码，再到用于自动语音识别的任务专用 CTC 头的处理流程。该模型在涵盖斯瓦希里语的 104 种语言、总计 450 万小时的数据上进行了预训练。

表 1：不同训练配置下的斯瓦希里语自动语音识别性能。基线指未经持续预训练训练的 5 万对比模型。 Δ 表示相较于基线的相对提升。

配置	基线	最终词错误率	Δ	收益
5K+持续预训练	17.71%	**10.89%**	-38.5%	✓
20K+持续预训练	17.71%	**3.24%**	-81.7%	✓

- 20K+持续预训练配置显著优于 50K 对比基线（3.24% 对 17.71%），这表明，与仅采用更多标注数据进行监督训练相比，使用较少标注数据进行持续预训练能取得明显更好的结果
- 3.24% 的词错误率超越了此前所有已发表的斯瓦希里语学术基准
- 恰当的训练策略比海量标注数据集更重要——仅用 20K 个样本便达到最先进水平

4.2 详细分析

4.2.1 5K 配置

对比基线（50K 个样本，未进行持续预训练）的词错误率为 17.71%。采用持续预训练方法且仅使用 5K 个标注样本时，模型的词错误率降至 10.89%，提升了 6.82 个百分点，相对降幅为 38.5%。仅使用 3 小时的标注数据，我们的持续预训练方法便能实现具有竞争力的性能，适用于草稿转录、语音搜索和语言学习工具等各种应用。

4.2.2 20K 配置

采用持续预训练方法并使用 20K 个标注样本时，模型的词错误率达到 3.24%，较对比基线降低 14.47 个百分点，相对降低 81.7%。使用 11 小时的标注数据，我们的持续预训练方法在斯瓦希里语自动语音识别方面达到了最先进水平，较最佳学术模型（XLS-R 的 8.3%）相对提升 61%。

4.3 基准对比

表2 展示了我们的结果在已发表的斯瓦希里语自动语音识别研究中的位置。

表 2：斯瓦希里语自动语音识别基准，对比我们的方法与已发表的学术系统

系统	Type	WER
微调后的 XLS-R [1]	学术	8.3%
SpeechBrain [7]	学术	9-12%
我们的方法 (5K+持续预训练)	**研究**	**10.89%**
我们的方法（20K+持续预训练）	**研究**	**3.24%**

意义：

• **新的最先进水平**：在学术系统中，Common Voice 斯瓦希里语上的最佳已报告词错误率

• **超越学术基线**：相较于 XLS-R，相对提升 61%

• **方法易于获取**：使用公开可用的模型和数据集

•**证明了可行性**：表明只需适度资源即可实现高质量的斯瓦希里语自动语音识别

5 讨论

我们的结果表明，利用伪标注的未标注数据进行持续预训练，即使标注数据需求极少，也能实现最先进水平的斯瓦希里语自动语音识别。仅使用 20,000 个标注样本，词错误率便达到 3.24%，相较于最佳学术系统（XLS-R 的 8.3%）提升了 61%。

5.1 持续预训练为何有效?

若干因素促成了持续预训练的成功:

5.1.1 强大的标注模型实现高质量伪标签

词错误率低于25%的标注模型能够提供足够准确的伪标签，从而产生有用的训练信号，而不会让噪声淹没模型。这与标注模型词错误率为 >40%的情形形成对比，在后者中伪标签噪声可能占据主导。

5.1.2 来自未标注音频的领域多样性

未标注音频使模型接触到多样的条件：多样的说话人（年龄、性别、地域差异）；多个领域（新闻、教育、娱乐）；多样的声学条件（录音室、户外、移动设备）；自然语音模式（相对于Common Voice中的朗读语音）。

5.1.3 目标对齐

持续预训练（伪标注）与监督微调均使用带字符级预测的CTC损失。这种目标对齐有助于各阶段之间平稳过渡。

5.1.4 保守超参数

我们采用保守的方法（学习率为 5e-5，仅训练 3 轮），以限制持续预训练对预训练表征的干扰，成功平衡了适应与遗忘。

5.2 实用指南

基于我们的研究发现，我们为从业者提出以下建议：

CPT要求：

• 至少有2万个标注样本可用于训练初始基线模型

• 基线模型的词错误率为 <25%，足以生成高质量伪标签

• 很容易获得未标注的目标语言音频

数据收集策略：

阶段 1：使用标注数据（≥20K 个样本）训练基线模型，使词错误率达到 <25%。

阶段 2：收集未标注音频，并使用基线模型生成伪标签，筛选出高置信度输出。

阶段 3：基于伪标注数据继续预训练，然后使用标注数据进行微调。

总投入：与传统监督式自动语音识别系统通常所需的数百小时数据相比，所需的标注数据较少。

6 结论

我们已证明，在伪标注的未标注数据上进行持续预训练，只需少量标注数据即可实现最先进水平的斯瓦希里语自动语音识别。我们的主要贡献包括：

1. **首次对斯瓦希里语伪标注持续预训练进行系统评估**

2. **斯瓦希里语自动语音识别达到新的最先进水平**（词错误率为 3.24%）

3. **明确的数据需求**：～20K 个标注样本与未标注音频相结合，即可实现高质量部署

4. **实用且可复现的方法论**，适用于其他低资源语言

我们的词错误率为 3.24%，比最好的学术系统提升了 61%。这是学术系统在 Common Voice 斯瓦希里语数据集上报告的最佳词错误率，仅使用了 20K 个标注样本便取得这一成绩。

更广泛的影响惠及超过 1 亿名斯瓦希里语使用者，他们可以受益于高质量的自动语音识别技术，并推动以母语开展教育技术、无障碍工具、语音界面以及口述传统的记录工作。我们的研究结果表明，只要采用恰当的训练方法、拥有适量的标注数据资源，并能获取现成的未标注音频，低资源语言也能实现高质量的自动语音识别。

参考文献

[1] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino等，“Xls-r：大规模自监督跨语言语音表征学习”，载于 *ICASSP* 会议论文集，2022年，第 6543–6547 页。[2]A. Baevski, Y. Zhou, A. Mohamed 和 M. Auli，“wav2vec 2.0：语音表征自监督学习框架”，载于神经信息处理系统进展，第 33 卷，2020年，第 12449–12460 页。[3]Seamless Communication Team, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoff-man, C. Hwang, N. Kalbassi, A. Kohn, M. Lee, Q. Licht, J. Ma, P. Nguyen, R. Pan, E. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Sagar, H. Schwenk, A. S. Tsai, C. Wang, G. Wenzek, B. Wood, Y. Wood, D. Yu 等，“Seamless：多语言表现力语音翻译与流式语音翻译”，*arXiv* 预印本 *arXiv:2312.05187*，2023年。[4] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Syn-naeve 等，“Robust wav2vec 2.0：分析自监督预训练中的领域偏移”，载于*Interspeech*，2021年。[5] A. A. Attia, S. Shehata 和 H. M. Al-Barhamtoshy，“用于自动语音识别中 wav2vec2.0 领域适应的持续预训练”，*arXiv* 预印本 *arXiv:2405.13018*，2024年。

[6] K. Nowakowski、Y. Shekofteh、J. Choi 等，“通过多语言微调和持续预训练，使多语言语音表征模型适应一种新的低资源语言”，*InformationProcessing& Management*，第 60 卷，第 2 期，第 103148 页，2023 年。[7] M. Ravanelli、T. Parcollet、P. Plantinga、A. Rouhe、S. Cornell、L. Lugosch、C. Subakan、N. Dawalatabad、A. Heba、J.-C. Chou 等，“Speechbrain：通用语音工具包”，2021 年。[8] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力即所需的一切”，载于神经信息处理系统进展，2017 年，第 5998–6008 页。[9] A. Graves、S. Fernández、F. Gomez 和 J. Schmidhuber，“连接时序分类：使用循环神经网络标注未分段的序列数据”，载于第 *23* 届国际机器学习会议论文集，2006 年，第 369–376 页。

[10] R. Ardila、M. Branson、K. Davis、M. Kohler、J. Meyer、M. Henretty、R. Morais、L. Saunders、F. Tyers 和 G. Weber，《Common voice：一个大规模多语种语音语料库》，载于 *LREC* 论文集，2020年。

[11]Mozilla，《Common Voice 语料库 16.0》，https://commonvoice.mozilla.org/，2024年。

[12] I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，载于 国际学习表征会议，2019年。