基于自监督表征蒸馏的Whisper低资源语音识别优化方法

胡剑[1,2]，董凌[1,2]，王文君[1,2]，相艳[1,2]，高盛祥*[1,2]，余正涛[1,2]
1. 昆明理工大学，信息工程与自动化学院，昆明，650500
2. 云南省人工智能重点实验室，昆明，650500
hujian151@stu.kust.edu.cn,ling.dong@kust.edu.cn,20203104003@stu.kust.edu.cn,
sharonxiang@126.com,gaoshengxiang.yn@foxmail.com,ztyu@hotmail.com

摘要

Whisper是一种强大的多语言语音识别模型，在英语等高资源语言上表现优异，但在缅甸语等部分低资源语言的性能仍受限于预训练数据的不足。为此，本文提出了一种基于自监督表征蒸馏的Whisper低资源语音识别优化方法。通过跨模型表征蒸馏机制，实现自监督模型表征向Whisper编码器的知识迁移，提升对缅甸语等语言的表征建模能力。实验结果表明，该方法在缅甸语、柬埔寨语、乌兹别克语和旁遮普语ASR任务中有效降低了字符错误率，验证了所提方法的有效性。

关键词： Whisper ； 自监督语音预训练模型 ； 语音识别 ； 表征蒸馏

Optimizing Whisper for Low-Resource Speech Recognition via Self-Supervised Representation Distillation

Jian Hu[1,2],Ling Dong[1,2], Wenjun Wang[1,2],
Yan Xiang[1,2],Shengxiang Gao*[1,2],Zhengtao Yu[1,2]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming, 650500, China
hujian151@stu.kust.edu.cn,ling.dong@kust.edu.cn,20203104003@stu.kust.edu.cn,
sharonxiang@126.com,gaoshengxiang.yn@foxmail.com,ztyu@hotmail.com

Abstract

Whisper is a robust multilingual automatic speech recognition (ASR) model that demonstrates exceptional performance on high-resource languages such as English. However, its effectiveness on certain low-resource languages, including Burmese, remains constrained by insufficient pretraining data. To address this limitation, this paper proposes an optimization method based on self-supervised representation distillation for Whisper low-resource speech recognition. By leveraging a cross-model representation distillation mechanism, our method facilitates the transfer of knowledge from self-supervised model representations to the Whisper encoder, thereby enhancing its representational modeling capability for Burmese and other low-resource languages. Experimental results demonstrate that the proposed approach significantly reduces the character error rate (CER) on ASR tasks for Burmese, Khmer, Uzbek, and Punjabi, validating the efficacy of our method.

Keywords: Whisper , Self-Supervised Speech Pretrained Models , Speech Recognition , Representation Distillation

*高盛祥（通信作者）：gaoshengxiang.yn@foxmail.com
基金项目：国家自然科学基金（62466030，U24A20334，62376111）；云南省重点研发计划（202303AP140008，202502AD080014）；云南省人工智能重点实验室开放基金（CB24069D018A）

1 引言

近年来，随着海量语音标注数据的积累与利用，自动语音识别(Automatic Speech Recognition, ASR)(Malik et al., 2021)技术取得了显著进展。由OpenAI提出的Whisper模型(Radford et al., 2023)通过利用68万小时的大规模多语言标注数据，构建了一个统一的多语言、多任务端到端语音识别框架。该模型集成了自动语音识别(ASR)、语音翻译(Speech Translation, ST)和语音活动检测(Voice Activity Detection, VAD)等多项功能，在高资源语言尤其是英语任务中展现了卓越的识别性能。这种性能优势主要得益于其海量训练数据支持以及统一的编解码架构设计。

尽管Whisper在英语等高资源语言上表现优异，但其对于部分低资源语言(Magueresse et al., 2020)的适用性仍面临显著挑战。以缅甸语、柬埔寨语为例，其公开可用的标注语料往往不足数十小时，远低于英语、中文等高资源语言的万小时级别。这种数据稀缺性直接导致Whisper预训练数据中存在严重的语言分布不平衡问题——例如，缅甸语仅占其68万小时训练数据的0.1小时。实验数据显示，在此极端数据不平衡条件下，Whisper对缅甸语的词错率(WER)甚至超过100%。在此背景下，研究者们尝试提出改进策略，以缓解其在低资源语言中的性能瓶颈。现有的主要探索方向集中在两个方面：一是基于相似语言的迁移微调，采用"高资源-低资源"的迁移学习范式，将知识从高资源语言迁移到低资源语言。例如Pillai et al. (2024)提出了一种多阶段微调流程：首先在中等资源的高相关语言(如泰米尔语)上进行中间微调，再对目标低资源语言(如马拉萨尔语)进行最终微调，以逐步缩小语言间的差距；Nagasawa et al. (2025)则利用语言之间的家族归属关系，先在同语系的高资源语言上微调Whisper模型，然后将微调后与原始模型参数的差值作为"任务向量"，迁移应用至同语系的低资源语言，以此实现跨语言适应。然而，这类方法的有效性严重依赖于语言间的音系或语法相似性，并且可能会牺牲模型在原始高资源语言上的性能，产生"灾难性遗忘"现象。二是引入外部语言模型，在解码或训练过程中融合语言建模能力，以增强模型对低资源语言的上下文建模与语义一致性能力。现有工作如Li et al. (2024)将语言模型GPT与Whisper模型集成提升Whisper在哈萨克语上的识别性能，并利用未标记的语音数据生成伪标签，根据平均令牌对数概率(ALP)选择样本进行模型微调；de Zuazo et al. (2025)通过将传统和新型语言模型与经过微调的Whisper模型相结合，从而提升其在巴斯克语、加利西亚语和加泰罗尼亚语上的表现。这类方法虽可在推理阶段提升文本质量，但本质上并未增强Whisper编码器对低资源语言的语音建模能力，无法根本缓解表征能力不足的问题。综合而言，现有方法虽在一定程度上提升了低资源语言识别效果，但未能增强模型对低资源语言表征建模能力，同时也可能会损坏模型原本性能。

值得注意的是，Whisper是一种基于监督学习的语音模型，其训练高度依赖大规模标注语音数据。然而，另一种学习范式——自监督学习(Self-supervised Learning, SSL)——近年来在语音表征学习中展现出巨大潜力。与传统监督方法不同，自监督学习通过构建预测性任务(如掩码建模、对比学习等)，可以从无标注语音数据中学习通用的语音特征，显著降低了对标注资源的依赖。在语音识别等下游任务中，这种方法已取得广泛成功。特别地，诸如XLS-R(Babu et al., 2021)和mHuBERT(Boito et al., 2024)等多语言自监督模型，在大规模无标注多语言数据上预训练后，展现出强大的跨语言迁移能力——即使对训练样本较少的语言，也能捕捉到鲁棒的声学模式。这启发我们思考：能否将这些模型在低资源语言中学到的丰富表征知识，通过蒸馏(Distillation)机制迁移至Whisper，从而弥补其在低资源语境下的表征能力不足?

由此，本文提出了一种基于自监督表征蒸馏的方法，利用自监督语音预训练模型作为教师模型，将其强大的表征能力迁移到Whisper编码器。为实现表征蒸馏并保留Whisper原本性能，我们在Whisper编码器末端插入Adapter模块，采用Wasserstein距离(Panaretos and Zemel, 2019)或Soft-DTW(Cuturi and Blondel, 2017)两种对齐方法对教师模型与Whisper表征进行表征蒸馏，实现从自监督模型到Whisper编码器的知识迁移，增强Whisper的表征建模能力。本文选择XLS-R模型作为教师模型，主要因其在大规模无标注多语言数据上的自监督预训练具备更强的泛化能力。以缅甸语、柬埔寨语、旁遮普语和乌兹别克语为例，XLS-R的平均训练时长达36小时，而Whisper仅为0.6小时，这种差距使XLS-R更能捕捉这些语言的声学特征，弥补了Whisper模型对这些语言的表征不足。在训练中，Whisper模型参数保持冻结，仅对Adapter进行微调，保留模型原始性能并降低训练开销。本文的贡献如下：

(1)提出一种结合Adapter的跨模型蒸馏方法，在保持Whisper原始模型对高资源语言识别能力的同时，有效提升其在低资源语言下的表征能力；

(2)引入基于Wasserstein距离和Soft-DTW的跨模型表征对齐方法，以应对Whisper模型和自监督模型的输出表征在时间维度上的长度差异问题；

(3)在缅甸语、柬埔寨语、乌兹别克语和旁遮普语四种语言上进行的实验表明所提方法有效提升了识别效果，分别实现了4.8%、6.5%、6%和6.3%的字错率降低。

2 相关工作

由OpenAI提出的Whisper模型是当前最具代表性的多语言端到端语音识别系统。该模型采用基于Transformer(Vaswani et al., 2017)的编码器-解码器架构，在68万小时的大规模多语言监督数据(涵盖96种语言)上进行训练。其核心创新在于统一的序列到序列建模框架：编码器通过堆叠卷积下采样层和Transformer层将语音信号转换为隐表征，解码器则通过自回归方式联合完成语音识别(ASR)、语音翻译(ST)和语种检测(LID)等多任务输出，所有任务共享相同的Transformer参数空间。目前，Whisper在英语ASR任务中，在LibriSpeech test-clean数据集上的WER为2.7% (Panayotov et al., 2015)，在CommonVoice英语测试集上的WER为9.0% (Ardila et al., 2019)。然而，Whisper在部分极低资源语言(如缅甸语、柬埔寨语)上的识别性能仍存在明显不足，如Whisper模型对缅甸语，柬埔寨语的WER都超过了100%。这一现象突显了对于部分低资源语言，Whisper模型在语音特征建模上的不足，亟需进一步的优化和提升。

与传统的监督学习方法相比，自监督学习(Self-supervised Learning, SSL)能够通过在大规模未标注语音数据上进行预训练，自动学习语音信号的底层结构和语言无关特征，从而显著减少对人工标注数据的依赖。典型方法包括对比学习(如wav2vec 2.0(Baevski et al., 2020))和掩码语音建模(如HuBERT(Hsu et al., 2021))，通过预测被掩码片段或区分正负样本，使模型获得具有判别性的语音表征。目前，主流的自监督语音模型主要包括：Facebook提出的wav2vec系列模型，通过对比预测编码实现语音特征提取；微软开发的WavLM(Chen et al., 2022)，结合掩码预测和说话人识别任务提升表征的通用性；以及HuBERT模型，利用迭代聚类产生伪标签进行自监督训练。特别值得一提的是XLS-R和mHuBERT等多语言自监督模型，它们在大规模无标注多语言数据上进行预训练，即使对于低资源语言也能学习到鲁棒的语音特征。其中XLS-R是通过对wav2vec 2.0进行多语言无标注数据预训练得到的，具有多种不同参数规模的模型，包括300M、1B和2B参数版本。这些模型在LibriSpeech等基准测试中展现出优越性能，同时在低资源场景下也表现出良好的迁移能力。

3 方法

本文提出一种基于自监督表征蒸馏的跨模型知识迁移框架，旨在提升Whisper模型在低资源语言场景下的建模能力。 如图1所示，该框架由三个关键模块组成：Whisper编码

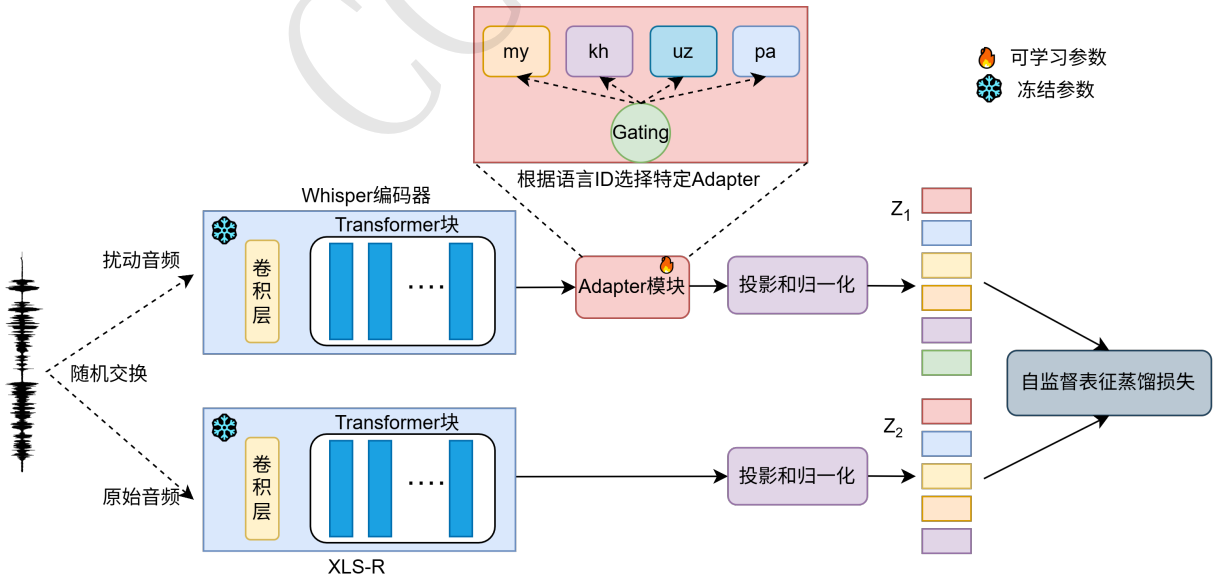


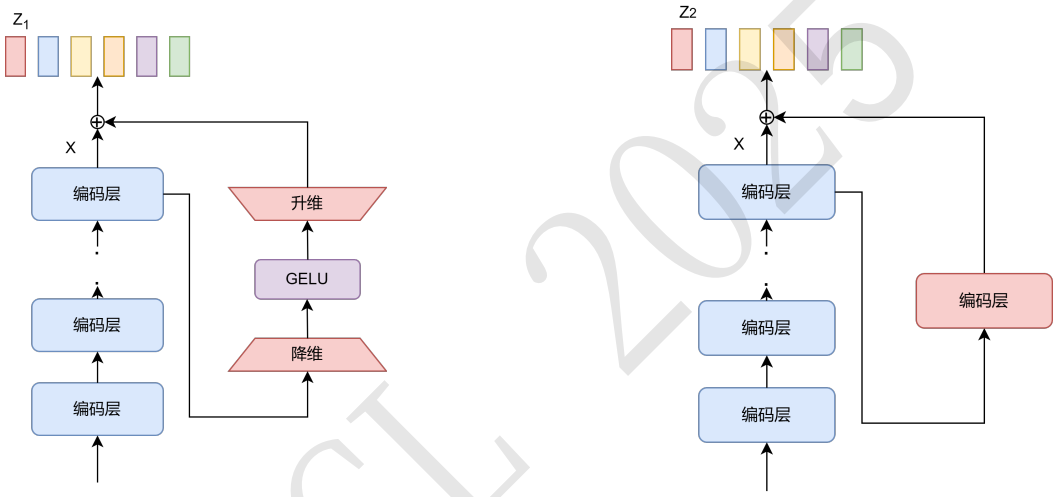
Figure 1: 面向Whisper模型的跨模型表征蒸馏方法结构图

器作为目标模型，用于提取语音表征；XLS-R作为教师模型，提供更强大的语音表征；以及接入Whisper输出端的Adapter模块，用于实现语言特化建模。Adapter模块内部包含基于语言ID的动态门控机制，用于选择性激活对应语言的子模块，以提升适配效率与扩展性。训练阶段，原始语音与其扰动版本分别输入Whisper编码器与XLS-R模型，通过随机交换(Swapping)操作将原始语音和加噪样本分别输入Whisper和XLS-R模型。Whisper编码器输出的表征记为Z_1，XLS-R 输出的表征记为Z_2。本文使用自监督表征蒸馏损失(具体方法见第3.3节)对Z_1和Z_2进行帧级对齐学习，引导Whisper编码器中的Adapter更好地拟合XLS-R在低资源语言下的特征分布，从而增强Whisper对低资源语言的表征能力。

3.1 Adapter模块设计

直接微调Whisper编码器在实验中可以带来一定的性能提升，但这种方式存在两个问题：
迁移干扰风险：编码器层的参数更新可能破坏模型原本对其他语言(尤其是高资源语言)学习到的表示，导致整体性能退化。
训练效率问题：直接微调涉及参数规模大、训练开销高，且在数据稀缺条件下易发生过拟合。

为解决上述问题，本文在Whisper编码器末端插入Adapter模块，作为微调过程中的唯一可训练部分。这种设计既保留了Whisper在高资源语言上的表现，又为特定语言的特征调整提供了灵活性，从而提升模型在多语言环境中的泛化性与可控性。本文设计了两种不同结构的Adapter，其结构如图2所示：



(a) Bottleneck Adapter　　(b) Transformer Block Adapter

Figure 2: 两种Adapter构造

(1)Bottleneck Adapter

该结构采用降维-非线性激活-升维的经典设计，首先通过线性层将输入维度从d降至较低维度$d_{\text{bottleneck}}$，再通过GELU激活，最后通过升维层恢复至原始维度。其计算公式为：

$$\text{Adapter}(x) = W_{\text{up}} \cdot \text{GELU}(W_{\text{down}} \cdot x) \tag{1}$$

其中$x \in \mathbb{R}^{T\times d}$为Whisper编码器最后一层的输出，$T$是帧数，$d$是特征维度，$W_{\text{down}} \in \mathbb{R}^{d_{\text{bottleneck}}\times d}$，$W_{\text{up}} \in \mathbb{R}^{d\times d_{\text{bottleneck}}}$。该结构参数少、效率高，适用于快速适配低资源语言的特征分布。

(2)Transformer Block Adapter

该结构直接复制Whisper编码器最后一层的Transformer Block，并将其作为结构更复杂的Adapter模块，仅对其内部参数进行微调。该层保留原有的多头注意力与前馈模块结构，具备更强的上下文建模与特征变换能力。

无论采用哪种结构，Adapter的输出均通过残差连接与原始编码器输出相加，得到最终表征：

$$Z_1 = \text{Encoder}_{\text{last}}(x) + \text{Adapter}(x) \tag{2}$$

该方式在保证原始表征稳定性的同时，引入了可控的可学习成分，有助于特征空间的柔性调整。

此外，两种Adapter均采用参数继承策略进行初始化，即在训练前直接复制Whisper编码器最后一层的权重与偏置，用于初始化对应的Adapter结构，从而确保初始状态与原模型一致，避免性能退化，并加快模型收敛速度。微调过程中，Adapter的输出将作为后续表征蒸馏模块的输入(详见3.3节)，以增强模型对低资源语言的感知能力。为系统评估不同迁移策略的效果，本文还引入了直接微调Whisper编码器最后一层参数的方式作为对照。一方面用于与Adapter方法在低资源语言上的识别性能进行比较，另一方面也用于考察该方法是否会对原始模型在其他语言(如英语)上的性能产生影响，验证其潜在的迁移干扰问题。

3.2 数据增强

为提升模型对低资源语言中多样语音条件的鲁棒性，同时构造更丰富的跨模态对齐训练样本，本文引入了两种常见的语音扰动方法：语速扰动(Speed Perturbation)与音高偏移(Pitch Shift)。具体如下：

语速扰动：对语音信号进行时间压缩或拉伸，模拟说话速度的自然变化。本文设置了多个速度因子(如0.9, 1.0, 1.1)以生成不同节奏的音频版本。

音高偏移：在保持语速不变的前提下对语音的音高进行平移，模拟说话人差异、情绪变化等因素对语音频谱的影响。本实验中，音高扰动范围设置为±2个半音(semitones)。

原始音频与扰动后的音频通过Swapping操作随机分配给Whisper和XLS-R模型，形成交叉的输入对。这种策略有效提升了训练样本的多样性，使模型在微调过程中能感知更加接近真实场景的语音分布。需要特别指出的是，所引入的扰动仅影响语音信号的音质和节奏层面，不改变语音的语义内容。因此，该方法不会破坏表征对齐所依赖的语义一致性假设，保证了训练过程的有效性与稳定性。

3.3 表征蒸馏

为增强Whisper编码器在低资源语言下的声学建模能力，本文引入一种基于自监督表征蒸馏的方法，通过跨模型特征拟合，引导Whisper学习来自XLS-R模型的泛化表征能力。给定一段语音输入，将其分别输入Whisper编码器(含Adapter 模块)与XLS-R 模型，调整维度后获得两组帧级特征序列：

$$Z_1 = \{z_1^1, z_1^2, \ldots, z_1^n\} \subset \mathbb{R}^d, \quad Z_2 = \{z_2^1, z_2^2, \ldots, z_2^m\} \subset \mathbb{R}^d$$

由于两种模型采用不同的下采样策略与特征抽取结构，导致生成的特征序列长度不一致($n \neq m$)，无法采用逐帧损失进行直接对齐。为此，本文采用Wasserstein距离和Soft-DTW两种方法来分别实现跨模型表征蒸馏。

(1)Wasserstein距离

Wasserstein距离是一种衡量概率分布间差异的有效方法，能够反映分布在整个空间中的几何结构。相比于KL散度等传统指标，它在分布无重叠时仍能提供稳定梯度，具有对称性与三角不等性，并能更细致地刻画语义层面的偏移，因此适用于不同模型间的表征对齐任务。

首先构建代价矩阵$C \in \mathbb{R}^{n\times m}$，其中$C_{i,j}$表示Whisper与XLS-R第$i$帧和第$j$帧特征间的欧氏距离：

$$C_{i,j} = \|z_1^i - z_2^j\|_2^2 \tag{3}$$

随后引入传输矩阵$T \in \mathbb{R}^{n\times m}$，表示从$z_1^i$向$z_2^j$的质量分配权重。最终的Wasserstein损失定义为代价矩阵与传输矩阵的Frobenius内积：

$$\mathcal{L}_{\mathrm{W}} = \langle T, C\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} T_{i,j}\cdot C_{i,j} \tag{4}$$

为实现高效且可微的传输矩阵求解，本文采用Sinkhorn-Knopp算法对T进行近似优化。该方法在最优传输问题中引入熵正则化项，不仅提升了数值稳定性，还加快了收敛速度，适用于神经网络训练过程中的端到端优化。

第二十四届中国计算语言学大会论文集，第563页-第573页，济南，中国，2025年8月11日至14日。

(2)Soft-DTW

Soft Dynamic Time Warping(Soft-DTW)是一种基于时间序列对齐的相似度度量方法，通过软对齐路径捕捉时序局部结构，特别适合建模语音信号中的动态变化。与Wasserstein距离的全局分布匹配不同，Soft-DTW 更侧重帧级局部结构的时间对齐。

与上述方法相同，首先构建局部距离矩阵$C \in \mathbb{R}^{n\times m}$，其中$C_{i,j} = \|z_1^i - z_2^j\|_2^2$。随后递归计算累积代价矩阵$R \in \mathbb{R}^{n\times m}$，其中每个元素$R_{i,j}$表示从起始点到$(i,j)$的最小软对齐代价：

$$R_{i,j} = C_{i,j} + \text{softmin}^{\gamma}\left(R_{i-1,j}, R_{i,j-1}, R_{i-1,j-1}\right) \tag{5}$$

其中softmin^{γ}是由温度参数γ控制的soft minimum函数，定义如下，可平滑近似传统的min运算：

$$\text{softmin}^{\gamma}(a_1,\ldots,a_k) = -\gamma \log \sum_{i=1}^{k} e^{-a_i/\gamma} \tag{6}$$

当$\gamma \to 0$时，Soft-DTW退化为传统DTW；当$\gamma > 0$时，对齐过程保持可微性，适用于神经网络训练。最终的Soft-DTW损失为累积代价矩阵的右下角元素：

$$\mathcal{L}_{\text{Soft-DTW}} = R_{n,m} \tag{7}$$

上述两种跨模型表征蒸馏方法分别从全局分布结构(Wasserstein 距离)与局部时间路径(Soft-DTW)两个维度对Whisper与XLS-R的输出特征进行对齐，旨在促进特征蒸馏效果。在实际训练过程中，Whisper模型的主干参数保持冻结，仅优化插入的Adapter模块，并采用对应的损失函数进行端到端优化：Wasserstein方法以$\mathcal{L}_{\text{W}} = \langle T, C\rangle$为优化目标；Soft-DTW方法以$\mathcal{L}_{\text{Soft-DTW}} = R_{n,m}$为优化目标。针对每种目标低资源语言，本文为其分别训练了一个独立的Adapter模块，并通过语言ID控制在进行ASR任务时激活对应语言的Adapter，从而确保蒸馏优化的表征在特定语言上发挥最大效能。

4 实验与分析

4.1 数据集

为验证所提出方法在低资源语言场景下的有效性，本文选取了Fleurs(Conneau et al., 2023)多语言语音数据集中的四种低资源语言：缅甸语(约12h)、柬埔寨语(约7h)、乌兹别克语(约10h)和旁遮普语(约6.3h)作为实验目标。这四种语言在下游任务ASR中，使用Fleurs划分的train/test/dev三个数据集，以确保评估的公平性与可比性。此外对于缅甸语和柬埔寨语本文还补充了OpenSLR[2]项目中的SLR80(缅甸语约4h)，SLR42(柬埔寨语约4h)作为微调数据。

4.2 评价指标

本文采用字符错误率(Character Error Rate, CER)作为主要评估指标，用于衡量语音识别模型在低资源语言下的性能。相比于词错误率(Word Error Rate, WER)，CER更适用于本研究所涉及的四种目标语言：缅甸语、柬埔寨语、乌兹别克语和旁遮普语。这些语言由于缺乏有效的分词工具或词边界规则模糊，导致直接计算WER的效果不具备可比性，容易产生偏高的错误率，从而不能真实反映模型性能。因此，本文统一采用CER作为评价标准，以保证跨语言评估的一致性和公正性。

4.3 实验设置

预训练模型：本文选用Whisper-base的编码器作为学生模型，并采用XLS-R(300M)作为教师模型，分别代表有监督与自监督的多语言表征体系。其中Whisper模型保持全冻结，仅微调其编码器末端的Adapter模块。

随机音频拼接：Whisper编码器的输入窗口大小固定为30秒，对于较短的语音输入，会在末尾填充零向量(zero-padding)以匹配窗口大小。然而，XLS-R模型不进行填充，这导致Whisper和XLS-R之间的特征对齐出现问题：Whisper的填充值会被误认为是有效语音特征，而XLS-R没有对应的填充部分，导致计算Wasserstein距离时Whisper的填充区域可能会匹

[2]https://www.openslr.org/resources.php

配XLS-R的真实语音特征，从而影响模型的训练效果。为了解决这一问题，本研究引入了随机音频拼接(Random Audio Concatenation)策略，该方法与(Lin et al., 2022)中的拼接方法类似：在微调过程中，对于每个语音样本，随机从训练集中选取若干个额外的语音片段进行拼接。拼接过程持续进行，直到拼接后的音频长度达到预设的30秒上限。

模型微调策略：从Whisper编码器(带Adapter)和XLS-R模型的最后一层Transformer(第12层)中获得的表征分别是512维和1024维的向量。这些向量通过不同的线性投影层被转换为256维的向量，并进行L2归一化。微调过程共进行3.6k次更新。优化器采用AdamW(Loshchilov and Hutter, 2017)，学习率为2.0e-5，并进行1k次预热更新。本文实验是在单个RTX3090 GPU上进行的，微调过程大约需要3个小时。最后，本文将微调第3600步中的模型用于S3PRL中的下游任务ASR进行训练。

SUPERB Benchmark：S3PRL框架可用于所有SUPERB基准任务(Yang et al., 2021)。在ASR任务中，来自所有层的特征会通过可学习的权重进行聚合，这些聚合后的特征随后被送入下游任务的预测头进行微调。具体而言，ASR任务的预测头由一个包含两层1024单元的双向LSTM网络组成，并使用基于字符的CTC损失进行训练。ASR模型的评估过程中未使用外部语言模型。Adam优化器用于ASR任务，学习率设置为1.0e-4。

4.4 实验结果

为了全面评估本文提出的基于自监督表征蒸馏的Whisper低资源语音识别优化方法在低资源语音识别任务中的效果，本文在四种语言上进行了实验，分别为缅甸语、柬埔寨语、乌兹别克语和旁遮普语。选取原始Whisper-base编码器作为基线，并以XLS-R(300M)作为教师模型，通过引入Adapter模块与最小化Wasserstein距离进行帧级表征对齐训练，结合输入拼接与语音扰动等策略，探索其对低资源语言建模能力的增强效果。最终得到的结果如表1所示。

Table 1: 模型在四种语言上的CER(%)对比。Adapter1代表Bottleneck Adapter，Adapter2代表Transformer Block Adapter，Whisper代表Whisper-base编码器。

模型	缅甸语	柬埔寨语	乌兹别克语	旁遮普语
Whisper	18.27	22.35	15.32	15.1
XLS-R(300M)	15.44	20.05	12.37	14.12
Whipser+Adapter1	17.39	20.88	14.39	14.15
Whipser+Adapter2	17.5	20.9	14.45	14.19

从整体结果来看，XLS-R模型在所有语言上的识别表现均优于Whisper模型，这一方面源自其大规模的无标注训练语料与参数量，另一方面也体现出其在跨语言建模方面更强的表征能力。例如在乌兹别克语任务中，XLS-R的CER相较Whisper从15.32%降至12.37%，呈现明显优势，柬埔寨语、缅甸语上和旁遮普语也分别降低了2.3%、2.83%和0.98%。该结果为后续基于XLS-R的表征迁移提供了理论支持。值得进一步指出的是，尽管旁遮普语在训练语料时长上最少，但在Whisper模型设置下均表现出相对最低的CER。我们分析认为，这一现象并不完全取决于语料规模，更与语料本身的质量密切相关。具体而言，旁遮普语语料整体发音清晰、字符集规模较小，且说话风格相对统一，可能有效降低了模型在学习过程中的模糊性与歧义性。这说明语料质量与语言复杂度同样对模型性能具有显著影响。

在此基础上，本文进一步引入了两种不同结构的Adapter模块，分别为采用双线性层构建的Bottleneck Adapter(Adapter1)和复制编码器最后一层结构的Transformer Block Adapter(Adapter2)。实验结果表明，这两种Adapter均能有效提升Whisper模型在低资源语言上的识别性能。例如在乌兹别克语上，两种Adapter分别将CER降至14.39%和14.45%，在柬埔寨语、缅甸语和旁遮普语上也取得了明显的改善，充分验证了引入适量可学习模块进行表征迁移的可行性。值得注意的是，Adapter1在四种语言上均取得了略优于Adapter2的效果，表现出更稳定的提升趋势。两种结构的具体差异及其对性能的影响将在后续实验中进一步分析。

综上所述，实验结果验证了本文方法在低资源语音建模任务中的有效性：(1)XLS-R模型的强表征能力可为Whisper提供明确的结构指导；(2)轻量级线性Adapter具备良好的迁移能力，在保证训练稳定性的同时提升了模型表达能力。最终模型在四种语言上均取得了优于原

始Whisper的CER结果，展现出良好的实际应用潜力。

4.5 实验分析

4.5.1 不同参数迁移策略对模型性能的影响

在本节实验中，我们探索了不同参数迁移策略对模型性能的影响，重点比较了三种代表性方法：使用双线性层构造的轻量型Adapter(记作Adapter1)、基于复制Whisper编码器最后一层结构的Adapter(记作Adapter2)，以及直接微调Whisper编码器的最后一层。在保持数据增强方式(随机音频拼接与语音扰动)一致的前提下，本文在缅甸语、柬埔寨语、乌兹别克语和旁遮普语四个低资源语言以及英语这种主流语言上评估了各策略的效果，结果如表2所示。

Table 2: 不同参数迁移策略在ASR任务中对比，四种低资源语言使用CER评估，英语使用WER评估，Whisper代表Whisper-base编码器。

模型配置	微调参数量	缅甸语	柬埔寨语	乌兹别克语	旁遮普语	英语
Whisper	-	18.27	22.35	15.32	15.1	9.7
Whisper+Adapter1	0.52M	17.39	20.88	14.39	14.15	9.7
Whisper+Adapter2	3.15M	17.50	20.90	14.45	14.19	9.7
Whisper+微调最后一层	3.15M	**17.35**	**20.85**	**14.33**	**14.1**	10.84

实验表明，在四种低资源语言上，三种迁移方式均优于原始Whisper模型。其中，直接微调编码器最后一层的方案在四个语言上分别取得了17.35%、20.85%、14.33%、14.1%的最低CER，显示出最强的表征适应能力。这说明，尽管该方法参数量较大(约3.15M)，但其对模型主干层的直接优化更有助于捕捉低资源语言中的细粒度语音特征，从而提升整体识别准确率。相比之下，两种Adapter方法虽然性能略逊一筹，但也取得了稳定的性能提升，且在参数量上展现出更强的效率优势。Adapter1仅引入约52.5万可训练参数，即可在四种语言上带来显著CER降低，体现了其优秀的参数效率与迁移能力。而Adapter2采用完整复制编码器最后一层结构的设计，虽然参数量与微调方法相当(约3.15M)，但因其更新范围局限于插入模块，整体性能略低，在三种方案种效果最差。然而，直接微调编码器虽然在目标低资源语言上效果最优，但也带来了模型泛化能力下降的问题。我们在英语上进一步评估了微调后的模型性能，发现其WER从原始模型的9.7%上升至10.84%，考虑到英语是Whisper训练数据中最主要的高资源语言之一，该结果表明微调策略可能破坏了原始模型对高资源语言的泛化建模能力。这种现象可以理解为模型在学习特定语言表示的同时“遗忘”了部分跨语言的通用表征能力。而对于两种Adapter方法，由于模块与原模型解耦，非目标语言的识别过程可以选择跳过Adapter，从而保留Whisper原有的参数与能力。

4.5.2 不同表征蒸馏方法对模型性能的影响

在本节实验中，我们探索了不同表征蒸馏方法对模型性能的影响，进一步对比了两种典型的序列间对齐方法：Wasserstein距离与Soft-DTW(Soft Dynamic Time Warping)。两种方法均在相同的模型结构(Whisper+Adapter1)及相同的数据增强配置(随机音频拼接与语音扰动)下进行，仅表征蒸馏方法不同。

Table 3: 不同表征蒸馏方法对模型在四种语言ASR任务的影响(评价指标：CER%)。

表征蒸馏策略	缅甸语	柬埔寨语	乌兹别克语	旁遮普语
Wasserstein距离	**17.39**	**20.88**	**14.39**	**14.15**
Soft-DTW	17.44	20.93	14.46	14.19

实验结果如表3所示，使用Wasserstein距离的模型在缅甸语(17.39%)、柬埔寨语(20.88%)、乌兹别克语(14.39%)和旁遮普语(14.15%)上均略优于使用Soft-DTW的模型。从机制上看，Soft-DTW更侧重于寻找两段特征序列之间的“最优对齐路径”，对于特征之间存在轻微位置偏移的情况具有较好的鲁棒性。然而，在Whisper和XLS-R这类结构差异较大的模型之间，由于二者输

出特征的分布存在明显差异，Soft-DTW在对齐时容易偏向于局部的匹配路径，导致对整体表征结构的捕捉能力有限。相比之下，Wasserstein距离从整体分布的角度出发，通过学习一种全局的质量分配策略来进行特征对齐，不依赖于具体的时间路径选择，因此在面对模型表征差异较大或特征时序不一致的情况下，具有更强的对齐能力和泛化能力。

4.5.3 不同数据增强方法对模型性能的影响

在本节实验中，我们探索了不同数据增强方法对跨模型表征蒸馏效果的影响，重点分析了随机音频拼接策略与语音扰动策略在低资源语言ASR任务中的作用。实验采用Whisper+Adapter1结构，评估指标为四个低资源语言(缅甸语、柬埔寨语、乌兹别克语和旁遮普语)的字符错误率(CER)，结果如表4所示。

Table 4: 不同数据增强方法对模型在四种语言ASR任务的影响(评价指标：CER%)，Whisper代表Whisper-base编码器。

数据增强方法	缅甸语	柬埔寨语	乌兹别克语	旁遮普语
Whisper	18.27	22.35	15.32	15.1
Whisper + Adapter1	17.53	21.02	14.52	14.28
+ 随机音频拼接	17.43	20.91	14.45	14.2
+ 语音扰动	**17.39**	**20.88**	**14.39**	**14.15**

首先，相较于基线的Whisper编码器，引入XLS-R特征蒸馏与Adapter模块后，模型性能在四个语言上均有显著提升，表明从XLS-R迁移知识能够有效增强Whisper对低资源语言的建模能力。在此基础上，加入随机音频拼接策略进一步带来了小幅度的性能提升，其原因可以归结为两个方面：一方面，拼接策略扩展了训练样本的长度分布和发音多样性，有助于模型学习更鲁棒的对齐关系；另一方面，这一策略间接缓解了Whisper编码器输入机制带来的对齐偏差。由于Whisper编码器的输入窗口固定为30秒，对于较短语音往往通过末尾填充零向量(zero-padding)以凑齐窗口长度，而XLS-R模型则不会进行填充。随机拼接多个语音片段构成更接近完整窗口长度的输入，有效减少了Whisper端的padding占比，从而提升了跨模型特征对齐的准确性。在拼接策略基础上引入语音扰动策略，进一步提升了模型的泛化能力，四种语言上CER均有细微下降。这说明在蒸馏过程中加入轻微扰动可以使Adapter学习到更加稳定且对输入变化具有鲁棒性的表征分布，从而提升最终识别性能。

为进一步分析语音扰动策略中不同扰动类型的具体贡献，我们补充了在缅甸语任务上的消融实验。具体设置包括：仅使用语速扰动、仅使用音高偏移，以及联合使用两者，均在“随机音频拼接”策略的基础上进行。实验结果如表5所示，语速扰动与音高偏移均对模型性能有正面影响，分别将CER由17.43%降低至17.41%和17.40%，当两者联合使用时，CER进一步降至17.39%。该结果表明，多样化的语音扰动策略能够提供互补的表征增强效果，有助于提升模型在低资源条件下的泛化能力。

Table 5: 不同语音扰动策略在缅甸语ASR任务上的消融实验（CER%）。所有设置均基于Whisper+Adapter1+随机音频拼接。

数据增强方法	缅甸语
无扰动（仅拼接）	17.43
拼接+ 语速扰动	17.41
拼接+ 音高偏移	17.40
拼接+ 语速扰动与音高偏移	**17.39**

4.5.4 不同模型容量对模型性能的影响

在本节实验中，我们系统评估了不同参数规模的教师模型(XLS-R 300M与1B)与学生模型(Whisper-base与Whisper-small)组合对模型性能的影响。实验采用Whisper编码器+Adapter1结构,在保持数据增强策略(随机音频拼接与语音扰动)一致的前提下，我们在缅

甸语、柬埔寨语、乌兹别克语和旁遮普语四个低资源语言上评估了不同模型容量的影响，结果如表6所示。

Table 6: 不同参数量的模型对模型在四种语言ASR任务的影响(评价指标：CER%)，Whisper-base代表Whisper-base编码器，Whisper-small代表Whisper-small编码器。

模型设置	缅甸语	柬埔寨语	乌兹别克语	旁遮普语
Whisper-base	18.27	22.35	15.32	15.1
Whisper-small	16.08	19.52	13.12	11.76
XLS-R(300M)	15.44	20.05	12.37	14.12
XLS-R(1B)	14.74	18.76	11.2	12.41
Whisper-base+XLS-R(300M)	17.39	20.88	14.39	14.15
Whisper-base+XLS-R(1B)	17.38	20.76	14.45	14.07
Whisper-small+XLS-R(1B)	15.64	18.89	12.4	11.79

结果显示，学生模型由Whisper-base升级为Whisper-small在所有语言上显著降低了CER，验证了更大容量模型的建模优势。在同一学生模型下(如Whisper-base)，使用更大规模的教师模型(如XLS-R 1B)通常带来更低的CER，表明增加教师模型的参数规模对于提升蒸馏效果具有一定的积极作用。这一现象归因于大模型在自监督预训练阶段所学到的更丰富的跨语言语音表征，从而为学生模型(Whisper)提供了更具泛化能力的监督信号。然而，在乌兹别克语上，1B模型表现反而略逊于300M，可能因语言特异性或蒸馏过程中的过拟合所致。进一步来看，Whisper-small搭配XLS-R(1B)取得了整体最佳性能，表明双重容量提升有助于增强低资源语言建模能力。但蒸馏并非总是有效：例如在旁遮普语上，教师模型(12.41%)性能弱于学生模型(11.76%)，导致蒸馏后表现略降(11.79%)，说明教师模型质量不足时反而可能抑制学生原有能力。此外，XLS-R(1B)的计算与显存开销显著增加，尤其在长音频输入时更易出现溢出，需在实际应用中权衡性能与资源消耗。

5 结论

本文提出了一种基于自监督表征蒸馏的Whisper低资源语音识别优化方法，旨在通过引入多语言自监督模型XLS-R的语音表征，引导Whisper模型学习更具泛化能力的低资源语言特征。实验结果表明，所提出的方法在缅甸语、柬埔寨语、乌兹别克语和旁遮普语四种低资源语言上均取得了显著的性能提升。与原始Whisper模型相比，表征蒸馏方法在字符错误率(CER)方面实现了稳定的下降，证明了引入自监督模型表征作为知识引导对低资源语音识别具有积极效果。同时，轻量化的Adapter模块在提升低资源语言性能的同时，亦能有效保留模型在高资源语言上的识别能力，体现出良好的迁移效率与通用性。

未来工作中，我们计划进一步探索跨模型表征蒸馏的层级策略、语言特定调控机制，以及其他大规模多语言语音模型在跨语种迁移中的潜力，以持续提升多语言、尤其是极低资源语言下的语音识别性能。

参考文献

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. *arXiv preprint arXiv:2406.06371*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Marco Cuturi and Mathieu Blondel. 2017. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR.

Xabier de Zuazo, Eva Navas, Ibon Saratxaga, and Inma Hernáez Rioja. 2025. Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint arXiv:2503.23542*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Jinpeng Li, Yu Pu, Qi Sun, and Wei-Qiang Zhang. 2024. Improving whisper's recognition performance for under-represented language kazakh leveraging unpaired speech and text. *arXiv preprint arXiv:2408.05554*.

Yist Y Lin, Tao Han, Haihua Xu, Van Tung Pham, Yerbolat Khassanov, Tze Yuang Chong, Yi He, Lu Lu, and Zejun Ma. 2022. Random utterance concatenation based data augmentation for improving short-video speech recognition. *arXiv preprint arXiv:2210.15876*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Haruki Nagasawa, Shinta Otake, and Shinji Iwata. 2025. Task vector arithmetic for low-resource asr. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Victor M Panaretos and Yoav Zemel. 2019. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Leena G Pillai, Kavya Manohar, Basil K Raju, and Elizabeth Sherly. 2024. Multistage fine-tuning strategies for automatic speech recognition in low-resource languages. *arXiv preprint arXiv:2411.04573*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

第二十四届中国计算语言学大会论文集，第563页-第573页，济南，中国，2025年8月11日至14日。