通过多语言微调和持续预训练，使多语言语音表征模型适用于一种新的资源匮乏语言

Karol Nowakowski[a], **, Michal Ptaszynski[b], *, KyokoMurasaki[c] 和 Jagna Nieuważny[a]

[a]东北公益与科学大学，日本山形县酒田市[b]北见工业大学，日本北海道北见市[c]横滨国立大学（荣誉教授），日本神奈川县横滨市

arXiv:2301.07295v1 [cs.CL] 18 Jan 2023

文章信息

关键词：自动语音转写 自动语音识别 wav2vec 2.0 预训练变换器模型 语音表征模型 跨语言迁移 语言记录 濒危语言 资源匮乏语言 萨哈林阿伊努语

摘要

近年来，通过在大规模多语言文本或语音数据上进行自监督预训练而学习的神经模型，在资源匮乏语言上展现出良好的效果，尤其是在有相对大量相关语言数据可用时。尽管这项技术有望助力语言记录项目中的各项任务，例如语音转写，但为每一种新语言从头预训练多语言模型会非常不切实际。我们探讨了将现有的多语言 wav2vec 2.0 模型适配到新语言的可能性，重点关注一种极度濒危语言阿伊努语的实际田野调查数据。具体而言，我们（i）考察在微调过程中利用相似语言数据的可行性；（ii）验证通过目标语言数据进一步预训练能否提升模型性能。结果表明，持续预训练是将 wav2vec 2.0 模型适配到新语言的最有效方法，并能大幅降低错误率。此外，我们发现，如果有基于相关语音变体或音系特征相似的非相关语言预训练的模型可用，那么在目标语言有标注数据极少时，使用该语言的额外数据进行多语言微调，能够对语音识别性能产生积极影响。

1. 引言

语音转写的成本是语言记录项目面临的一大瓶颈。人们认为，利用语音识别技术可以使这项任务更容易完成，但此类项目所研究的语言通常可用数据（包括已标注和未标注数据）极为有限，这一事实阻碍了相关技术的应用。

近期研究（例如 Conneau、Baevski、Collobert、Mohamed 和 Auli，2021年）表明，在目标语言数据稀缺的情形下，多语言预训练语音表征有利于语音识别。然而，鉴于从头开始预训练此类表征的成本高得令人望而却步，预计大多数用户会仅对公开可用的模型进行微调。因此，本文探讨跨语言迁移的益处是否同样适用于微调。此外，我们还研究了利用目标语言少量可用数据进行额外预训练的效果。与以往大多数研究不同，我们关注的是有各种不足的真实田野调查数据，而非干净的自然语言处理数据集。具体而言，我们的目标是转写萨哈林阿伊努语的独特语音数据，这些数据录制在已有50年历史的录音带上。

我们的结果表明，在目标语言数据上进行持续预训练可显著降低错误率。此外，我们还证明，在目标语言的有标注数据极其匮乏的情况下，只要底层语音表征模型曾在相关语言上进行过预训练，那么在微调期间加入来自相关语音变体或具有相似音系特征的不相关语言的数据，也能提升语音识别性能。我们在阿伊努语上预训练的模型已公开发布[1]。

*通讯作者

**主要通讯作者karol@koeki-u.ac.jp (K.Nowakowski); michal@mail.kitami-it.ac.jp (M.Ptaszynski)开放研究与贡献者身份识别码(s)：0000-0001-7435-4061 (K. Nowakowski); 0000-0002-1910-9183 (M. Ptaszynski)

[1] https://huggingface.co/karolnowakowski/wav2vec2-large-xlsr-53-pretrain-ain

本文其余部分安排如下。在下一节中，我们将讨论语音转文本技术以及语音表征自监督学习的最新进展如何支持语言记录。我们还将介绍本研究所聚焦的数据。在第3节中，我们将介绍研究方法。在第4节中，我们将概述相关研究。第5 节介绍本研究使用的资源，包括语音表征模型及其训练数据。在第6节中，我们将介绍实验并分析实验结果。最后，第7 节总结全文并提出未来改进的思路。

2. 背景

2.1. 语言记录中的语音转写技术

濒危语言语言学研究的主要任务之一是收集和分析一手语言数据。典型的工作流程包括在田野调查期间录制语音，随后分析数据。在这一过程中，一个主要挑战是语音转写，这是一项非常耗时的任务[2]。因此，档案和个别研究者的收藏中有大量数据仍未转写。其中许多材料存储在过时的媒介上，例如录音带，而且保存条件不佳。在这些材料得到转写之前，其中一些可能会因意外而损毁，也可能在收集它们的研究者退休后遭到毁弃（Abney，2020年）。因此，这类材料通常被称为“濒危数据”。使用语音转文本技术或许可以解决（或至少缓解）转写瓶颈，从而加快语言记录进程（Hjortnaes、Partanen、Rießler 和 Tyers，2020年；Zahrer、Zgank 和 Schuppler，2020年）。然而，为了达到较高的准确率，传统方法需要大量标注训练数据（通常达数千小时（Baevski、Zhou、Mohamed 和 Auli，2020年）），而在语言记录场景中通常无法获得这类数据。

2.2. 面向低资源语音转写的跨语言自监督学习

过去几年，由于发展出了高效的语言表征模型自监督学习技术，自然语言处理的众多应用取得了显著进展，例如 BERT（Devlin、Chang、Lee 和 Toutanova，2019年）、ELECTRA（Clark、Luong、Le 和 Manning，2020年）（用于文本模型），wav2vec 2.0（Baevski 等人，2020年）和 HuBERT（Hsu、Bolte、Tsai、Lakhotia、Salakhutdinov 和 Mohamed，2021a）（用于语音表征）。研究表明，与传统的全监督方法相比，这些模型只需使用少得多的人工标注样本进行训练（此前已在大量未标注数据上预训练），便能取得有竞争力的结果。

在语言记录场景中应用自监督学习的一大障碍是，世界上大多数语言都没有大量的原始文本或语音数据。为缓解这一问题，近期研究探索了跨语言学习技术（Conneau、Khandelwal、Goyal、Chaudhary、Wenzek、Guzmán、Grave、Ott、Zettlemoyer 和 Stoyanov，2020年；Singh、McCann、Keskar、Xiong 和 Socher，2019年）。研究表明，利用多种语言的未标注数据训练单一模型，能够提升为每种语言计算出的表征质量。例如，Conneau 等人（2021年）在 53 种语言的 5.6 万小时未标注数据上预训练了语音表征模型，结果发现该模型的表现远胜于单语模型，尤其是在数据匮乏的语言上。然而，训练此类模型所需的计算资源成本以及由此消耗的能源都极其高昂[3]。

2.3. 萨哈林阿伊努语语音数据

本研究旨在开发一套用于自动转录阿伊努语的系统。阿伊努语是一种极度濒危的语言[4]，原生于日本北部、萨哈林和千岛群岛。具体而言，我们聚焦于转录阿伊努语在萨哈林使用的若干方言中尚未发表的材料。这些材料由 Ky¯oko Murasaki 教授于 20 世纪 60 年代和 70 年代¯录制，受访者是这些方言最后一批说话者中的几位：Haru Fujiyama（Rayciska 方言）、Ota Yuku（Maoka 方言）、Chikama Kimura（Shirahama 方言）等。相关材料最初录制在 30 多盘老式磁带上（即所谓的“卷对卷”磁带或“开盘式”磁带），并且……

分钟到一小时不等（Cieri、Miller 和 Walker，

，Conneau 等人(2021)的多语言模型使用64个图形处理器进行训练，而Conneau 等人(2020年)则使用500个图形处理器训练了他们的 XLM-R 模型。[4]从本世纪下半叶起，20th 阿伊努语便不再作为日常交流语言使用（Bugaeva，2012），因此许多专家认为它已经灭绝。不过，人们正努力复兴这种语言，学习阿伊努语的人也越来越多。

表 1 Hattori Hattori (1964年）中记录的萨哈林阿伊努语（Rayciska 方言）和北海道阿伊努语（Horobetsu 方言）对应

萨哈林阿伊努语	北海道阿伊努语	日语	英语
ku’ani ’enahkari nee ’阿伊努人 何时 a ’oman.	恩莫斯玛·嫩·卡·森内 ’oma。	除了我之外，谁 mo 不去。	没人去 but me.
’uneeno ’an ’itahpateh kisci。	sine ’itak ’ukoraci ’an ’itak patek ye kor an。	总是说同样的话。	他总是说同一件事。
这是阿伊努语篇章 ayyeepe ?	tanpe ’anak 阿伊努人’itak ’ari nekona ’aye ya .	kore wa ainugo de nan to iimasu ka?	这在阿伊努语中叫什么？

这是“濒危数据”的一个很好的例子。音质相当差，噪声很大（既有嘶嘶声和嗡嗡声等设 是由一位说话者录制的自发独白（主要是

录音总时长超过 20 小时，据我们所知，这使其规模超过了迄今为止已出版的任何萨哈林阿伊努语文本集[5]。部分录音已转写、译成日语并出版：Murasaki (1976) 出版了 Haru Fujiyama 讲述的十一篇民间故事，以及五段简短的 ̄两位母语说话者（Fujiyama 和 Ota）之间的对话。Murasaki 和 Fujiyama (2010) 汇编了同一篇民间故事《Wenenekaype》的三个不同版本，均由 Haru Fujiyama 讲述。最后，Murasaki 和 Fujiyama (2013年) 和Murasaki 和 Fujiyama (2016) 出版的两卷共收录 297 个句子。

萨哈林阿伊努语是阿伊努语中公认的三大方言群之一（另外两种是北海道阿伊努语和千岛阿伊努语）。萨哈林与北海道的方言差异显著，彼此无法互通（Refsing，1986；Murasaki，2009）（表1 列出了两种语音变体中语义相同的句子示例）。Vovin (2016) 将阿伊努语描述为一个包含两个主要分支的组合语系：北海道—千岛分支和萨哈林分支。北海道阿伊努语已有大量录制并出版的音频和书面材料，近年来，其中越来越多的材料被数字化并在线发布。然而，萨哈林阿伊努语的数据则稀缺得多[6]。

迄

3. 研究方法

鉴于多语言语音表征模型在资源匮乏语言上取得了令人鼓舞的成果，且从头预训练此类模型成本高昂，我们探索了两种方法，利用有限的可用语音数据，将现有模型适配到一种新语言。

[5]除了萨哈林阿伊努语之外，相关磁带中还包含 2 小时的北海道阿伊努语语音录音。[6]千岛阿伊努语几乎没有可用数据，只有几部词典和一些词表。

多语言微调 除了多语言自监督预训练之外，Conneau 等人 (2021年) 还开展了一项实验，同时使用来自 10 种不同语言的有标注数据进行微调，发现所得模型的表现可与分别针对每种语言单独微调的模型相媲美。另一方面，他们并未分析语言相似性与这类多语言微调效果之间的相关性[7]。本研究将考察跨语言迁移的益处是否也适用于微调，包括发生在密切相关的语言／方言之间，以及发生在共享某些音系特征但彼此无关的语言之间。为此，我们使用目标语言（萨哈林阿伊努语）的数据，并结合来自三种不同语音变体的大量数据开展微调实验：北海道阿伊努语、日语和英语。如第 2.3 节所述，萨哈林阿伊努语和北海道阿伊努语可以视为同一语言的差异较大的变体，也可以视为密切相关的语言。至于日语，大多数专家都不认同日语与阿伊努语之间存在谱系关系的理论（Refsing，1986；Shibatani，1990）。尽管如此，尽管阿伊努语和日语在允许出现在音节韵尾中的辅音、音调等方面存在显著差异，但可以说，阿伊努语的音系系统与日语的共同点更多[8,9]，而非例如英语。使用 `lang2vec` (Littell、Mortensen、Lin、Kairis、Turner 和 Levin，2017年)[10] 计算语言向量所得的分析也印证了这一判断：计算阿伊努语与数据库中所有其他语言的距离后（具体而言，我们比较了音系和音位库存特征，并取平均距离），发现日语在与阿伊努语的接近程度方面，在 8070 种语言中排名第 43rd 位。

持续预训练其次，我们探讨能否通过使用少量目标语言数据进行额外预训练，提升一个强大的多语言模型在初始预训练阶段未曾接触过的语言上的性能。

4. 相关工作

Schneider、Baevski、Collobert 和 Auli (2019年) 提出了一种名为 wav2vec 的技术，使用卷积神经网络从原始音频数据中自监督学习语音表征。该网络经过训练，能够区分真实音频样本与干扰样本。在仅使用少两个数量级的有标注数据的情况下，他们的方法在 WSJ 语音识别基准上的表现优于此前的最先进水平。Baevski、Schneider 和 Auli (2019年) 在此基础上加入了一个量化模块，用于计算音频片段的离散表示，并将离散化后的序列输入 Transformer（BERT）模型。随后，Baevski 等人(2020年) 提出了 wav2vec 2.0，进一步改进了这一方法。这是一种端到端框架，可联合学习离散化的语音单元和上下文化语音表示；他们还对所得模型进行了微调，以进行语音转写，而不是将预训练特征输入单独的下游模型。Conneau 等人 (2021年) 使用 53 种语言的 5.6 万小时语音数据预训练了单个 wav2vec 2.0 模型（称为 XLSR-53），其语音识别准确率高于单语模型和此前的方法。Hsu、Sriram、Baevski、Likhomanenko、Xu、Pratap、Kahn、Lee、Collobert、Synnaeve 和 Auli (2021b) 汇集了来自不同领域的语音数据，并研究了领域不匹配对自动语音识别自监督学习的影响。Xu、Baevski 和 Auli (2021年) 通过微调 XLSR-53 模型，并利用发音特征将训练语言的音位映射到目标语言，实现了对未见语言的零样本转录。Babu、Wang、Tjandra、Lakhotia、Xu、Goyal、Singh、von Platen、Saraf、Pino、Baevski、Conneau 和 Auli (2021年) 使用 wav2vec 2.0 和 128 种语言的 43.6 万小时未标注数据，训练了大规模（最多 20 亿参数）模型。这些模型经过微调后，在语音识别、语音翻译和语言识别方面达到了最先进的性能。Sriram、Auli 和 Baevski (2022) 通过对预训练数据应用数据增强技术（例如音高移调和向输入信号添加随机噪声），并对 wav2vec 2.0 架构进行若干修改，提升了自动语音识别性能。

序、语音特征等）的影响。他们发现，

[7]他们确实针对预训练进行了这类分析，并发现，低资源语言的性能从相似语言的额外数据中获益更多。[8]这些相似之处可能有一部分是接触诱发的语言变化所致（Bugaeva，2012）。[9]关于Nowakowski、Ptaszynski 和 Masui（2020年）使用日语语音模型识别并生成阿伊努语语音的案例，请参见相关研究。Matsuura、Ueno、Mimura、Sakai 和 Kawahara（2020年）使用额外的日语和英语数据，为北海道阿伊努语训练了一个端到端自动语音识别模型，并发现前者更有帮助。[10]`https://github.com/antonisa/lang2vec`

通过多语言微调和持续预训练，将语音表征模型适配到一种新语言

Han、Weinberger 和 Artzi （2022年）以及 Vyas、Hsu、Auli 和 Baevsk

先前研究发现，持续对文本语言表征模型进行自监督训练，是使其适应新领域的有效方法（Howard 和 Ruder，2018；Sun、Qiu、Xu 和 Huang，2019年；Gururangan、Marasović、Swayamdipta、Lo、Beltagy、Downey 和 Smith，2020 年），或将其覆盖范围扩展至初始预训练阶段未涉及的语言（ Pfeiffer 、 Vulić 、 Gurevych 和 Ruder，2020年；Tang、Tran、Li、Chen、Goyal、Chaudhary、Gu 和 Fan，2020年；Ebrahimi 和 Kann，2021年）。Kessler、Thomas 和 Karout 探讨了在新语言上进一步预训练语音表征模型的做法Kessler、Thomas 和 Karout （2021年）。然而，他们仅使用单语（英语）模型开展实验，并将重点放在高资源场景，即新增语言可用的语音数据充足（800 小时或以上）。Khurana、Laurent 和 Glass （2022年）使用自训练，在模拟的低资源场景中将单

人 （2020年），结果显示模型性能有所提升，尤其是在低资源语言上。

5. 材料

5.1. Wav2vec 2.0

本文所述的所有语音转写实验均使用了一个公开可用的预训练 wav2vec 2.0 模型。具体而言，我们使用了 XLSR-53——一个在 53 种语言的 5.6 万小时数据上训练的模型——该模型由[11] Conneau 等人汇编并发布 (2021年)。此外，我们还使用下一节所述的阿伊努语数据对 XLSR-53 模型进行了额外预训练。模型的预训练和微调均使用 `fairseq`库完成。

5.2. 预训练数据

在对 XLSR-53 进行持续预训练的实验中，我们使用了共 234 小时的萨哈林阿伊努语和北海道阿伊努语语音数据[12]。具体而言，我们使用第 2.3 节所述磁带中的数据，以及表 2 所列的公开数据集进行预训练。

通过将录音机连接到磁带录音机，将开盘磁带上的录音转录为数字格式（WAV）。磁带为双面，我们发现，同一盘磁带两面的录音中有许多内容都可以听见。经检查发现，在某些情况下，从反面提取的数据（或其中某些部分）在音频信号质量方面优于正面对应的数据。因此，我们将重复录音纳入预训练数据中，所以表中报告了两个数字2（括号中的数字对应唯一录音的时长）。

所有文件均转换为单声道 WAV 格式，采样率为 16 kHz。超过 15 秒的文件会按静音间隔自动拆分（使用 pydub），生成时长为 2 至 15 秒的独立片段。短于 1 秒的文件不纳入预训练。

5.3. 微调数据

微调所用数据列于表3中。单语言微调使用了两个来源的萨哈林阿伊努语数据：一则来自Murasaki 和 Fujiyama(2010)的故事（即 `Fu12-690401`；其余两段录音用于验证和测试），以及来自 Murasaki 和 Asai (2001)[13]的数据。多语言微调则加入了北海道阿伊努语数据（64.5 小时）、日语数据（取自 Common Voice 语料库 8.0 日语数据的经验证子集（Ardila 等人，2020年）[14], 版本 `ja_43h_2022-01-19`，以及 JSUT 语料库（Sonobe 等人，2017年）[15]；总计 50.9 小时）和英语数据（LibriSpeech 中 100 小时的“clean”子集）。

从Murasaki 和 Fujiyama (2010) 获取的语音数据根据静音区间自动切分（使用`pydub`），生成时长为 2 至 15 秒的独立片段。书中的转录文本经过数字化处理，并与音频片段进行了对齐。

所有音频文件均转换为单声道 WAV 格式，采样率为 16kHz。所有转录文本中的标点符号和元数据均已删除。萨哈林阿伊努语和北海道阿伊努语转录文本中的所有字母字符

[11] `https://github.com/pytorch/fairseq/tree/main/examples/wav2vec` [12] 虽然在仅使用萨哈林数据的情况下预训练模型也可能有所帮助，但由于预训练成本高昂，我们决定不这样做。[13] 也可在线访问：`http://www.aa.tufs.ac.jp/~mmine/kiki_gen/murasaki/asai01.html` [14] `https://commonvoice.mozilla.org/ja/datasets` [15] `https://sites.google.com/site/shinnosuketakamichi/publication/jsut`

表 2用于持续预训练的数据统计。对于从磁带获取的萨哈林阿伊努语数据，我们纳入了重复录音，因此报告了两个数字（括号中的数字对应于去重录音的时长）。

Data	（主要）语言/方言	总时长（小时）
萨哈林阿伊努语磁带	萨哈林阿伊努语	35.9（21.5）
Tuytah（Murasaki 和 Asai，2001）	萨哈林阿伊努语	8.9
“Wenenekaype”（Murasaki 和 Fujiyama，2010）	萨哈林阿伊努语	1.9
阿伊努语档案（An=ukokor Aynu ikor oma kenru（国家阿伊努博物馆），2017–2022）	北海道阿伊努语	103.1
鹈川阿伊努语词典 (千叶大学 研究生院 人文与社会科学, 2014)	北海道阿伊努语	26.5
阿伊努语与阿伊努口头文学 (二风谷阿伊努文化 Mu- 博物馆，无日期）	北海道阿伊努语	19.2
早稻田大学阿伊努语音频材料 知识库 (https://waseda.repo.nii.ac.jp/)	北海道阿伊努语	14.1
ILCAA 的 阿伊努语资料出版项目 阿伊努语资料出版项目 （信息资源中心， 语言与 语言 材料 亚洲与非洲语言文化研究所，东京外国语大学，无日期）	北海道阿伊努语	12.1
阿伊努民间故事带注释音频语料库（Nakagawa 等，2016）	北海道阿伊努语	6.2
萱野茂的阿伊努语词典（Kayano，1996）	北海道阿伊努语	3.0
会话阿伊努语主题词典（国家语言与语言学研究所，2015年）	北海道阿伊努语	2.3
Ny¯u ekusupuresu 阿伊努语（Nakagawa，2013年）	北海道阿伊努语	1.0

文本均转换为小写。为了避免输出词汇量大幅增加并减少数据稀疏性，日语转录文本使用 pykakasi 转写为片假名音节文字。这也适用于阿伊努语数据中的日语内容，其中包含许多日语单词和话语（语码转换、关于文本的评论、采访者提出的问题等）。LibriSpeech 转录文本则保留原始形式，即全部使用大写字母。

表4中展示了“Wenenekaype”转录文本的一个节选。

6. 实验

6.1. 持续预训练

随机抽取 1% 的数据用于验证。预训练使用四块 Nvidia GTX 1080Ti 图形处理器进行。我们继续预训练了 100k 次更新，总耗时 5 周。为防止灾难性遗忘，我们采用了较小的学习率（1e-4；Conneau 等人在初始预训练中使用的学习率为2021e-3）（Sun 等人，2019年）。我们使用的批量大小为每块图形处理器 150k 个样本，并采用梯度累积来模拟 512 块图形处理器的效果，达到 80 分钟的有效批量大小。其他超参数则依据LARGE 模型的配置设定，该配置由Baevski 等人（2020年）报告。

6.2. 微调

我们使用 CTC损失对预训练模型进行微调，以进行语音转写（Graves、Fernández、Gomez 和 Schmidhuber，2006）。每次实验运行的最佳检查点根据验证集上的词错误率选出（为此，我们使用了来自Murasaki 和 Fujiyama（2010）的最短录音，时长为10分钟，即Fu13-700326）。使用维特比解码器和四元语法语言模型对微调模型的输出进行解码；该语言模型基于相应模型微调时所用数据中的萨哈林阿伊努语部分进行训练（见表5）。

表3用于微调的语音数据和转录文本统计。日语书写时不使用空格，且我们未进行分词，因此未报告日语数据的词元数量和词汇表大小。

Data	（主要）语言/方言	总计时长（小时）	词数量	词汇size
Tuytah (Murasaki 和 Asai，2001)	萨哈林阿伊努语	8.9	52,172	4,415
“Wenenekaype” (Fu12-690401) (Murasaki and Fujiyama，2010)	萨哈林阿伊努语	0.8	5,817	1,165
阿伊努语档案（An=ukokor Aynu ikor omakenru（国立阿伊努博物馆），2017–2022）	北海道阿伊努语	62.2	396,755	10,618
会话阿伊努语主题词典（国家研究所日语语言学研究所，2015年）	北海道阿伊努语	2.3	13,007	2,260
Common Voice（日语）（Ardila 等人，2015年）	日语	40.6	N/A	N/A
JSUT（Sonobe 等人，2017年）	日语	10.3	N/A	N/A
LibriSpeech（Panayotov 等人，2015年）	英语	100.6	990,101	33,798

表4摘自用于微调我们的语音识别模型的“Wenenekaype”转录文本。

原始转录文本（Murasaki 和 Fujiyama，2010）	已预处理，供微调使用	英语译文
sine, sine.. ’oyanruru 村庄 ’ohta sine, ’oyanruru 村庄 ’an manu。’an 传说 很久以前 大 村庄 ’an manu。 但是 然后，什么 ka，什么 祖先 说， ’an 传说 neyan wen-poro 村庄 ’oma 阿伊努人 ka ’emuyke ruhpa ’ike tuy wa (tuy)pa wa ’isam。 ’isam mayne tani ’ampene ’oha 村庄 nee manu。	sine sine ’oyanruru 村庄 ’ohta sine ’oyanruru 村庄 ’an manu ’an 很久以前，大村庄…… manu，但是，然后，hemata ka hemata’oyasi hee ’an 传说 我 大 村庄，那里的阿伊努人也 ’emuyke ruhpa 但是 tuy wa tuypa wa ’isam ’isam mayne tani ’ampene ’oha 村庄 nee manu	那里曾有一个大村庄。一个非常 big 村庄。 但是后来，那里 有某种怪物把所有……都吃掉了 那座大村庄里的人，而那里 已经没有人了。他们全都离开了， 如今这里只剩下一座空荡荡的村庄。

表5用于解码的 KenLM 语言模型统计信息，包括其在评估数据上的困惑度和未登录词率。

训练 data	词汇 size	困惑度（包括词汇表外词）	困惑度（不包括词汇表外词）	词汇表外词元率
“Wenenekaype”	1,168	152.1	92.9	563/4911 (11.5%)
“Wenenekaype” +Tuytah	5,095	290.7	181.9	403/4911 (8.2%)

使用 KenLM 工具包计算语言模型[16]。语音转写性能在来自 Murasaki 和 Fujiyama (2010)（即 Fu11-690328）这一数据子集（时长 37 分钟）上进行了评估。在评估之前，将系统生成的转录文本中的所有字母字符转换为小写。我们报告字符错误率（CER）和词错误率（WER）。

所有微调实验均使用单台 Nvidia RTX 3090 图形处理器进行。基线模型仅使用目标领域（即“Wenenekaype”）中少量带标注的样本进行微调，共更新 15k 步，批量大小为 2.56M 个样本，学习率设为 3e-4。其他超参数按照LARGE 模型的配置设置，该模型由Baevski 等人（2020年）报告。在我们的系统上，采用这些设置进行微调所需时间不到

[16]https://kheafield.com/code/kenlm/

表 6在“Wenenekaype”测试集上的语音转写任务中，对比在单语或多语数据上微调的模型，以及是否使用目标语言数据进行额外预训练。我们报告字符错误率和词错误率。

微调数据	额外预训练步骤	维特比解码器		KenLM 解码器	
		CER	WER	CER	WER
“Wenenekaype”	0	13.9	42.3	18.0	36.5
萨哈林阿伊努语（“Wenenekaype” + Tuytah）	0	11.8	37.3	14.1	31.8
萨哈林阿伊努语 + 北海道阿伊努语	0	15.9	50.6	20.2	38.6
“Wenenekaype” +北海道阿伊努语	0	16.4	51.1	21.8	41.2
“Wenenekaype” +日语	0	15.7	41.8	19.0	38.0
“Wenenekaype” +英语	0	13.9	43.0	17.5	38.0
“Wenenekaype” + 北海道阿伊努语 +日语	0	16.0	52.0	22.1	41.1
“Wenenekaype” 北	0	15.2	47.4	20.5	40.4
“Wenenekaype”	100k	10.6	33.5	14.6	33.0
萨哈林阿伊努语（“Wenenekaype” + Tuytah）	100k	**9.7**	**29.8**	13.4	30.2
萨哈林阿伊努语 + 北海道阿伊努语	100k	10.1	31.1	12.6	30.8
“Wenenekaype” +北海道阿伊努语	100k	10.2	30.5	15.1	32.1
“Wenenekaype” +日语	100k	10.8	32.2	15.2	33.0
“Wenenekaype” +英语	100k	12.0	40.1	16.8	34.2
“Wenenekaype” + 北海道阿伊努语 +日语	100k	10.9	33.7	16.2	33.4
“Wenenekaype” 北	100k	10.9	33.9	16.8	34.3

1.5 小时。加入更多萨哈林阿伊努语数据后，我们以 1e-4 的学习率进行了 20k 步微调。虽然批量大小降至 800k 个样本，但我们采用梯度累积来模拟 16 个图形处理器，从而实现 12.8M 的有效批量大小，训练时间约为 10 小时。在此设置下，“Wenenekaype”数据的过采样倍数为 10（我们发现，不进行过采样时，验证集上的性能会差得多）。对于使用两种语音变体数据组合进行微调的模型，我们将学习率设为 3e-5，并执行了 80k 次更新。将萨哈林阿伊努语和北海道阿伊努语数据合并进行微调时，“Wenenekaype”数据的过采样倍数为 200，Tuytah 数据的过采样倍数为 10。在未使用后者数据的实验中，“Wenenekaype”数据的过采样倍数为 100。在这种情况下，单次微调约需 2 天。加入第三种语言（日语）的数据后，我们进行了 120k 步微调，耗时近 3 天。最后，使用全部四种语言的数据进行微调时，我们执行了 160k 次更新，耗时 4 天。在这两种情况下，“Wenenekaype”数据均经过过采样，约占训练数据的一半。

6.3. 结果与分析

在表 6 中，我们比较了在测试集上进行语音识别时，使用单语或多语数据微调的模型，以及是否在阿伊努语数据上进行额外预训练所得的错误率。图 1 分析了使用目标语言数据进行额外预训练的影响。

我们的结果表明，持续预训练显然是使语音表征模型适应新语言的最有效方法。仅经过 1 万次更新（在我们的系统上耗时半周），我们便将字符错误率降低了近 13%（使用维特比解码器对模型输出进行解码时）。经过两周和 4 万次更新后，性能提升达到 24.5%。在字符错误率方面，测试数据上的最佳相对性能是在 6 万次更新后取得的。就词错误率而言，使用维特比解码器进行解码时，6 万次和 9 万次更新后测得的数值最低（分别为 32.7 和 32.5）；使用语言模型时，则是在 4 万步后达到最低值（32.3）。

使用所有可用的萨哈林阿伊努语数据进行微调，并在不使用语言模型的情况下进行解码，取得了整体最佳结果。在进一步预训练之前的模型上，无论采用哪种语言组合，也无论目标语言数据有多少，多语言微调都没有帮助（加入日语数据确实使词错误率略有降低，但字符错误率有所上升）。相当令人意外的是，加入北海道阿伊努语数据不仅对模型性能产生了负面影响，错误率也显著高于使用日语和英语数据进行的实验。这种现象可能有两个原因：（i）由于未能学到任何一种阿伊努语的表示，模型无法利用两者之间的相似性；（ii）

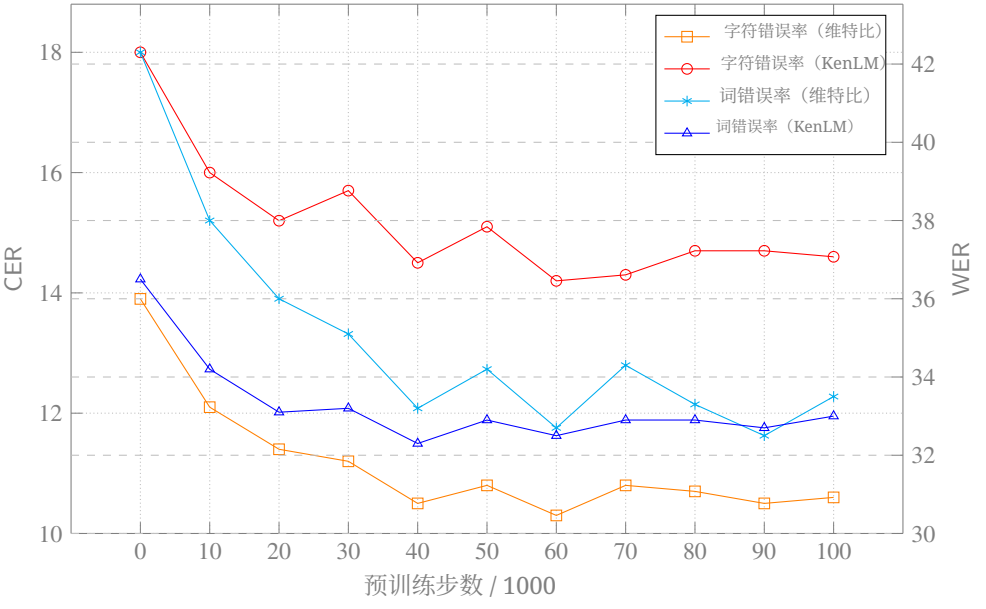

图1： 使用目标语言数据继续预训练的效果。模型仅使用“Wenenekaype”数据进行微调。

日语和英语与目标语言的差异足够大，使模型能够轻易区分它们，因此因语言混淆而产生的错误相对较少。

额外预训练后，多语言微调的结果截然不同：加入另一种阿伊努语变体的数据，得到的模型性能最佳；其次是在具有某些音系特征的无关语言上进行微调（即阿伊努语和日语）；而将阿伊努语与英语结合则明显有害。与使用 10 小时萨哈林数据训练的模型相比，加入北海道阿伊努语的额外数据并未带来提升；但在可用目标语言数据少于 1 小时的情况下，结果确有改善。在这种情况下，加入日语数据后，词错误率也有所下降。

在使用两种或三种其他语言的标注数据进行实验时，我们观察到结果更差，即使这些语言单独使用时曾带来帮助（即北海道阿伊努语和日语）。我们认为，这种现象可能与微调后的模型分配给每种语言的容量较低有关，但仍需进一步分析。

对使用维特比解码器生成的转录文本进行分析后发现，部分错误是由书写系统选择不当造成的（即使用片假名字符表示阿伊努语的语音，或者反过来，使用拉丁字母转录日语部分；表7展示了一个例子。事实上，语言学家经常使用罗马字转写系统，而非日文文字，来转录阿伊努语文本中的日语语码转换词。因此，我们决定考察：如果放宽问题设定，将系统输出和真实标注数据中的所有日文字符都转换为拉丁字母，结果会有何变化。具体而言，我们使用`pykakasi`按照赫本罗马字转写系统将日语转写为罗马字。结果见表8。虽然这一修改使几乎所有配置的字符错误率都略有降低，但使用日语数据微调的模型取得了最大的提升。根据这一评估方案，可以得出结论：联合使用“Wenenekaype”数据和日语语音数据进行微调显然有益。

上述结果与我们的假设一致：在微调中利用相似语言的有标注数据，可以提升模型在目标语言上的表现。另一方面，这些结果表明，只有满足以下两个条件，才能取得任何改进：(i) 模型最初是在微调所涉及的语言上进行预训练的，或者至少是在其中一种语言上进行预训练的（即，除了目标语言数据之外，还打算使用其数据的迁移语言）；(ii) 目标语言中的有标注数据量极少。

系统

表7微调模型生成的转录文本摘录，展示了因书写系统选择错误而导致的错误。

模型输出：	tanna ’an ’opompaki nah ramupe オota’asi nee manu ’タ ’asi ソ’oヤw nah ’anramu ’ampe アノ ジンkoy nee manu
转写：	[tanna ’an ’opompaki nah ramupe oota’asi nee manu ’ta ’asi so’oyaw nah ’anramu ’ampe ano jinkoy nee manu]
真实标注：	tani neya アノ ’opompaki nah ramupe ’ota’asi nee manu ’ota’asi suy ’oyaw nah ’anramu’ampe アノ cinkoy nee manu
转写：	[tani neya ano ’opompaki nah ramupe ’ota’asi nee manu ’ota’asi suy ’oyaw nah ’anramu’ampe ano cinkoy nee manu]

表 8将维特比解码器的输出和真实标注数据中的日文字符转写为罗马字后，所得的评估结果。括号中的数字表示与表 6 中结果相比，字符错误率/词错误率的变化。

微调数据	额外预训练步骤	CER	WER
“Wenenekaype”	0	13.7 (-0.2)	42.3
萨哈林阿伊努语（“Wenenekaype” + Tuytah）	0	11.6 (-0.2)	37.3
萨哈林阿伊努语 + 北海道阿伊努语	0	15.6 (-0.3)	50.5 (-0.1)
“Wenenekaype” +北海道阿伊努语	0	16.0 (-0.4)	51.0 (-0.1)
“Wenenekaype” +日语	0	13.4 **(-2.3)**	40.9 **(-0.9)**
“Wenenekaype” +英语	0	13.6 (-0.3)	43.0
“Wenenekaype” + 北海道阿伊努语 +日语	0	15.8 (-0.2)	52.0
“Wenenekaype” 北	0	14.8 (-0.4)	47.4
“Wenenekaype”	100k	10.5 (-0.1)	33.5
萨哈林阿伊努语（“Wenenekaype” + Tuytah）	100k	**9.6** (-0.1)	**29.8**
萨哈林阿伊努语 + 北海道阿伊努语	100k	9.8 (-0.3)	31.1
“Wenenekaype” +北海道阿伊努语	100k	9.9 (-0.3)	30.5
“Wenenekaype” +日语	100k	10.1 (-0.7)	31.9 (-0.3)
“Wenenekaype” +英语	100k	11.9 (-0.1)	40.1
“Wenenekaype” + 北海道阿伊努语 +日语	100k	10.3 (-0.6)	33.5 (-0.2)
“Wenenekaype” 北	100k	10.9	33.9

这

），而在训练数据中，它有两种不同的转写方式：大 （11 例）或w 未来，我们将研究自动检测数据中此类差异的

6.3.1. 解码方法的影响

Baevski 等人(2020年) 报告称，使用文本语言模型对其模型输出进行解码时，词错误率大幅降低。如表6所示，对于额外预训练之前的模型，我们也观察到词错误率显著降低，但字符错误率有所上升。例如，与维特比解码器相比，使用 4元语法模型对在所有萨哈林阿伊努语数据上微调的模型（“Wenenekaype” + Tuytah）进行解码时，词错误率降低了 15%，而字符错误率则上升了近 20%。表9展示了测试数据中的一个样本分别采用这两种解码方法生成的转录文本示例。

在阿伊努语数据上进行持续预训练，缩小了两种解码方法在词错误率方面的差距（见图 1）；总体而言，不使用语言模型进行解码时效果最佳。这一结果表明，在学会目标语言的表征后，语音表征模型能够从微调数据中学习到隐式语言模型，其能力强于根据相同数据计算得出的基于计数的n元语法模型。Baevski 等人Baevski 等人(2020年）的设置与我们的设置之间一个主要差异是，他们的语言模型是在一个包含超过8亿个词元的图书语料库上训练的（Panayotov 等人，2015年），

表9针对单个测试样本，经微调模型使用两种不同方法解码后的输出。KenLM 模型会强制使用词汇表内词语（例如，使用 *reekoh* 而不是 *reepoh* ，以及使用 *'ekasihi* 而不是 *'ekasihii*），这通常会降低词错误率。另一方面，维特比解码器更擅长处理词汇表外项目（例如，*'ankopuri*被转写为*'ankopuuri*，而语言模型则将其替换成一个完全不同的词*'ankopisi*），从而降低字符错误率。

维特比解码器：	ne'ohah nay kohnean tani macirih 'oho nay kohne 'anu wa 很久以前 ne'an cispuurikara 'ankii 传说 'awwen 'ekasihii 'ireske 'ankamuy henke 'ohta 'ankopuuri 'ahsin 传说 reepoh ne'an henke 'ihunke 这则传说很久以前
KenLM 解码器：	'oha nay 这种 tani 方言 'ohonkesehe ne'an cispuurikara 'ankii 传说 'awwen 'ekasihi 'ireske 'ankamuy henke 'ohta 'ankopisi 'asin 传说 很久以前 ne'an henke 'ihunke kii 传说 很久以前
真实标注：	'ohah naykoh 不是 tani macirihi 'ohoo naykoh 不是 nu wa 很久以前 不是 cispuurikara 'ankii 传说 'awwen 'ekasihi 'ireske 'ankamuy henkehe 'ohta 'ankopuri'ahte 传说 很久以前 ne'an henke 'ihunke kii 传说 很久以前

我们能够获取的萨哈林阿伊努语文本数据仅限于语音数据的转录文本，而这些数据也用于微调我们的模型。由此可见，在通常无法获得大量文本数据的语言记录场景中，使用外部语言模型进行解码可能作用有限。不过，未来的研究应探究，使用其他类型的语言模型（例如字符级语言模型）时，是否也会出现类似趋势。另一个可能影响结果的因素是书写系统的特征：Baevski 等人(2020年）在英语上进行了实验，而英语的字位与音位之间存在多对多的对应关系。相比之下，阿伊努语采用音位正字法转录，因此无需明确掌握各个词项的正确拼写。

6.3.2. 重叠字符词汇表的影响

Conneau 等人 (2021年)发现，在多语言微调中使用共享音位词汇表，比为每种语言分别维护词汇表能取得更好的结果。虽然我们的系统并非基于音位运行，但我们想验证，表示目标语言的字符词汇表与微调所用其他语言的字符词汇表之间的重叠，是否会影响语音转写的准确率。具体而言，我们使用以下数据微调了额外的模型：(i) 转换为大写字母的北海道阿伊努语转录文本；(ii) 转换为小写字母的 LibriSpeech（英语）转录文本；以及 (iii) 转换为小写拉丁字母的日语转录文本。在所有情况下，我们都使用了基于阿伊努语数据预训练的模型。模型输出由维特比解码器解码。为了尽量减少阿伊努语数据中使用日文文字书写的语码转换片段所造成的影响，生成的转录文本和真实标注数据中的片假名字符在评估前均被转写为拉丁字母。

结果见表10。目标语言数据少于 1 小时（即仅有“Wenenekaype”）且额外加入相近语言（北海道阿伊努语或日语）的数据时，使用共享输出词汇表可降低错误率。将英语文本转换为小写的结果不一：词错误率有所改善，但字符错误率更高。增加带标注的萨哈林阿伊努语数据后，结果则相反：使用大写的北海道阿伊努语文本训练的模型表现更好。与共享字符词汇表的实验不同，在这种设置下，同时使用萨哈林阿伊努语和北海道阿伊努语数据进行微调，相较于仅使用萨哈林阿伊努语数据，性能略有提升。

根据这些结果，我们得出结论：当标注的目标语言数据非常少，并且使用了来自相似语音变体的额外数据时，采用共享字符词汇表更好，因为这似乎有助于跨语言迁移。不过，这项建议应谨慎对待，因为语言之间的相似性未必与其书写系统（或转写方法）的相似性相对应，因此，其他语言组合的结果可能会有所不同。另一方面，只要音位—字位映射的差异具有可预测性，就可以通过对转录文本应用基于规则的预处理来处理不匹配。随着可用的目标语言数据增多，模型更常受到跨语言信号的干扰，而非从中受益，因此，最好可能只进行单语言微调（这也能节省计算资源），但也值得尝试使用独立字符词汇表。未来研究应对其他语言开展类似分析。

表10在微调中，为目标语言和其他语言使用独立或共享字符词汇表的效果。

微调数据	字符词汇表大小	CER	WER
“Wenenekaype”	76	10.5	33.5
“Wenenekaype”（小写） +北海道阿伊努语（大写）	102	10.2	31.2
“Wenenekaype”（小写） +北海道阿伊努语（小写）	86	9.9	30.5
“Wenenekaype”（字母） + 日语（片假名）	137	10.1	31.9
“Wenenekaype”（字母） + 日语（字母）	119	9.7	31.0
“Wenenekaype”（小写） + 英语（大写）	102	11.9	40.1
“Wenenekaype”（小写） + 英语（小写）	86	12.3	39.6
萨哈林阿伊努语	94	9.6	29.8
Sakh. 阿伊努语（小写） +北海道阿伊努语（大写）	111	**9.4**	**29.6**
Sakh. 阿伊努语（小写） + 北海道阿伊努语（小写）	95	9.8	31.1

7. 结论与未来工作

我们证明，强大的多语言语音表征模型（例如 XLSR-53）可以通过多语言微调和额外预训练适配一种新的低资源语言，从而提升下游任务的表现。通过对萨哈林阿伊努语自动转录开展实验，我们发现，在目标语言数据上持续预训练能够大幅降低错误率。此外，我们的结果表明，在标注的目标语言数据极其匮乏的情况下，如果模型在预训练期间见过相关语音变体的数据（或者在较小程度上，见过音系特征相似的非相关语言的数据），那么在微调期间加入这些数据，模型就能从中受益。

我们的研究结果证实了这一假设：应当考虑语言相似性，并且可以在多语言微调过程中加以利用。研究结果还表明，对语言表征模型进行自监督预训练不仅有助于使其适应特定语言，而且在这一过程中学到的表征还可以通过跨语言监督，作为迁移到相似语言的桥梁。我们预计，这一观察结果同样适用于同一语言内部的跨领域监督——未来的工作应当对这一假设进行研究。

们

[illegible]对齐，来增加带标注的领域内数据量。此外，我们还将[illegible]

致谢

本研究得到日本学术振兴会科学研究费助成事业（JSPS KAKENHI）资助，项目编号为 JP22K17952。

参考文献

Abney, S., 2011. 语言数字化。网址：http://www.vinartus.net/spa/p102-v2.pdf。一=ukokor Aynu ikor oma kenru（国立阿伊努博物馆），2017–2022。《阿伊努语 ¯档案 [阿伊努语档案]。网址：https://ainugo. nam.go.jp/。Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common Voice：一个大规模多语言语音语料库，载于：第12届语言资源与评估会议论文集（LREC 2020），第4211–4215页。Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M., 2021. XLS-R：大规模自监督跨语言语音表征学习。arXiv abs/2111.09296。Baevski, A., Schneider, S., Auli, M., 2019. vq-wav2vec：离散语音表征的自监督学习。CoRR abs/1910.05453。网址：http://arxiv.org/abs/1910.05453，arXiv:1910.05453。Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. wav2vec 2.0：语音表征自监督学习框架。ArXiv abs/2006.11477。

Bugaeva, A., 2012. 南北海道阿伊努语，载于：Tranter, N.（编），《日本与朝鲜的语言》。Routledge，伦敦，第461–509页。千叶大学人文社会科学研究科，2014。《阿伊努语鹿川方言日语—阿伊努语词典》¯《鹿川阿伊努方言日语—阿伊努语词典》 [《阿伊努方言词典》]。网址：https://www.gshpa.chiba-u.jp/cas/Ainu-archives/index.html。Cieri, C., Miller, D., Walker, K., 2004. Fisher语料库：面向新一代语音转文本技术的资源，载于：第四届语言资源与评估国际会议论文集（LREC'04），欧洲语言资源协会（ELRA），里斯本，葡萄牙。网址：http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf。Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA：将文本编码器预训练为判别器而非生成器，载于：ICLR。网址：https://openreview.net/pdf?id=r1xMH1BtvB。Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M., 2021. 用于语音识别的无监督跨语言表征学习，载于：Interspeech。Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. 大规模无监督跨语言表征学习，载于：计算语言学协会第58届年会论文集，计算语言学协会，线上。第8440–8451页。网址：https://aclanthology.org/2020.acl-main.747，doi:10.18653/v1/2020.acl-main.747。Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT：面向语言理解的深度双向Transformer预训练，载于：计算语言学协会北美分会2019年会议：人类语言技术论文集，第1卷（长篇与短篇论文），计算语言学协会，明尼阿波利斯，明尼苏达州。第4171–4186页。网址：https://aclanthology.org/N19-1423，doi:10.18653/v1/N19-1423。Ebrahimi, A., Kann, K., 2021. 如何将预训练多语言模型适配至1600种语言，载于：计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集（第1卷：长篇论文），计算语言学协会，线上。第4555–4567页。网址：https://aclanthology.org/2021.acl-long.351，doi:10.18653/v1/2021.acl-long.351。Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. 连接主义时序分类：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，计算机协会，纽约，美国纽约州。第369–376页。网址：https://doi.org/10.1145/1143844.1143891，doi:10.1145/1143844.1143891。Gries, S.T., Berez, A.L., 2017. 语料库语言学中的语言学标注，载于：Ide, N., Pustejovsky, J.（编），《语言学标注手册》。Springer，第379–409页。Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. 不要停止预训练：将语言模型适配到领域和任务，载于：计算语言学协会第58届年会论文集，计算语言学协会，线上。第8342–8360页。网址：https://aclanthology.org/2020.acl-main.740，doi:10.18653/v1/2020. acl-main.740。Hattori, S., 1964. 阿伊努语方言词典¯阿伊努语方言词典 [阿伊努语方言词典]。岩波书店，东¯京¯。Hjortnaes, N., Partanen, N., Rießler, M., Tyers, F.M., 2020. 面向科米语语音识别器：一种濒危且资源匮乏的乌拉尔语系语言，载于：第六届乌拉尔语系计算语言学国际研讨会论文集，计算语言学协会，维也纳，奥地利。第31–37页。网址：https://aclanthology.org/2020.iwclul-1.5，doi:10.18653/v1/2020.iwclul-1.5。北海道¯Utari Ky¯okai，1994。《我们的语言》 [《我们的语言》]。北海道¯Utari Ky¯okai，札幌。Howard, J., Ruder, S., 2018. 用于文本分类的通用语言模型微调，载于：计算语言学协会第56届年会论文集（第1卷：长篇论文），计算语言学协会，墨尔本，澳大利亚。第328–339页。网址：https://aclanthology.org/P18-1031，doi:10.18653/v1/P18-1031。Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021a. HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习。《IEEE/ACM音频、语音与语言处理汇刊》29，3451–3460。Hsu, W.N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., Auli, M., 2021b. 鲁棒的wav2vec 2.0：分析自监督预训练中的领域偏移，载于：Interspeech。亚洲・非洲语言文化研究所信息资源中心、东京外国语大学，未注明日期。AA-ken阿伊努语资料公开项目¯《阿伊努语资料出版项目》 [ILCAA的阿伊努语资料出版项目]。网址：http://ainugo.aa-ken.jp/。Kayano, S., 1996. 《萱野茂阿伊努语词典》 [《萱野茂的阿伊努语词典》]。三省堂¯，东¯京¯。Kessler, S., Thomas, B., Karout, S., 2021. Continual-wav2vec2：持续学习在自监督自动语音识别中的应用。ArXiv abs/2107.13530。Khurana, S., Laurent, A., Glass, J., 2022. 用于单语wav2vec-2.0跨语言适配的魔法粉尘，载于：ICASSP 2022—2022 IEEE声学、语音与信号处理国际会议（ICASSP），第6647–6651页。doi:10.1109/ICASSP43922.2022.9746276。Littell, P., Mortensen, D.R., Lin, K., Kairis, K., Turner, C., Levin, L., 2017. Uriel和lang2vec：将语言表示为类型学、地理和系统发育向量，载于：计算语言学协会欧洲分会第15届会议论文集：第2卷，短篇论文，第8–14页。Matsuura, K., Ueno, S., Mimura, M., Sakai, S., Kawahara, T., 2020. 阿伊努民间故事语音语料库与阿伊努语端到端语音识别，载于：LREC。Murasaki, K., 1976. 桦太阿伊努语／萨哈林雷奇斯卡阿伊努方言。Kokusho Kank¯okai，东京。Murasaki, K., 2009. 桦太阿伊努语入门会话／萨哈林阿伊努语入门。绿莺社，钏路。Murasaki, K., Asai, T., 2001. 《桦太阿伊努民间故事：Tuytah》 [《萨哈林阿伊努民间故事：Tuytah》]。S¯of¯ukan，东京。Murasaki, K., Fujiyama, H., 2010. 萨哈林阿伊努民间故事（ucaskuma）：Wenenekaype。《亚洲・非洲语言文化研究所东北亚研究》第2卷。亚洲・非洲语言文化研究所、东京外国语大学，东京。Murasaki, K., Fujiyama, H., 2013. Fujiyama Haru讲述的萨哈林阿伊努语短句（1）。《萨哈林阿伊努语丛书》第3卷。Aynu Teetawanoankur Kanpinuye Cise（阿伊努与原住民研究中心），北海道大学，札幌。

通过多语言微调，将语言模型迁移到Fujiyama种新语言

Fujiyama Haru 讲述的萨哈林阿伊努语短句（2）。第4卷，萨哈林阿伊努语丛书。Aynu Teetawanoankur Kanpinuye Cise（阿伊努与原住民研究中心），北海道大学，札幌。Nakagawa, H., 2013. Ny¯u ekusupuresu Ainugo. 白水社，T¯oky¯o. Nakagawa, H., Bugaeva, A., Kobayashi, M., Kimura, K., 2016. 阿伊努民间故事带注释音频语料库。网址：https://ainu.ninjal.ac.jp/ folklore/corpus/en/。日本国立国语研究所，2015年。《会话阿伊努语主题词典》。网址：https://ainu.ninjal. ac.jp/topic/。二风谷阿伊努文化博物馆，未注明日期。《阿伊努语与阿伊努口头文学》。网址：http://www.town.biratori.hokkaido.jp/biratori/ nibutani/culture/language/。Nowakowski, K., Ptaszynski, M., Masui, F., 2019. MiNgMatch——一种用于阿伊努语分词的快速N元语法模型。《Information》10，317。doi: 10.3390/info10100317。Nowakowski, K., Ptaszynski, M., Masui, F., 2020. 为低资源语言学习增添趣味：阿伊努语Pepper机器人初步实验，载于：第6届对话代理的语言学与认知方法研讨会（LaCATODA 2020）。网址: http://arakilab.media.eng.hokudai.ac.jp/IJCAI2020/LACATODA2020/Program_files/Spicing%20up%20the%20Game%20for%20Underresourced%20Language%20Learning%3A%20Preliminary%20Experiments%20with%20Ainu% 20Language-speaking%20Pepper%20Robot.pdf。Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. LibriSpeech：基于公有领域有声读物的自动语音识别语料库，载于：2015年IEEE声学、语音与信号处理国际会议（ICASSP），第5206–5210页。doi: 10.1109/ICASSP.2015.7178964。Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S., 2020. MAD-X：用于多任务跨语言迁移的基于适配器的框架，载于：2020年自然语言处理实证方法会议（EMNLP）论文集，计算语言学协会，线上。第7654–7673页。网址: https://aclanthology.org/2020.emnlp-main.617，doi：10.18653/v1/2020.emnlp-main.617。Refsing, K., 1986. 阿伊努语：静内方言的形态与句法。奥胡斯大学出版社，奥胡斯。Sanabria, R., Hsu, W.N., Baevski, A., Auli, M., 2022. 衡量自监督预训练中各个领域因素的影响。arXiv abs/2203.00648。Schneider, S., Baevski, A., Collobert, R., Auli, M., 2019. wav2vec：用于语音识别的无监督预训练，载于：INTERSPEECH。Shibatani, M., 1990.《日本的语言》。剑桥大学出版社，伦敦。Singh, J., McCann, B., Keskar, N.S., Xiong, C., Socher, R., 2019. XLDA：用于自然语言推理和问答的跨语言数据增强。arXiv abs/1905.11471。Sonobe, R., Takamichi, S., Saruwatari, H., 2017. JSUT语料库：用于端到端语音合成的免费大规模日语语音语料库。arXiv abs/1711.00354。Sriram, A., Auli, M., Baevski, A., 2022. Wav2vec-aug：利用有限数据改进自监督训练。arXiv abs/2206.13654。Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. 如何微调BERT以进行文本分类？，载于：Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y.（编），《中文计算语言学》，施普林格国际出版公司，Cham。第194–206页。Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., 2020. 借助可扩展的多语言预训练与微调实现多语言翻译。arXiv abs/2008.00401。Vovin, A., 2016. 论北海道的语言史前史。《东方学研究》117，29–38。Vyas, A., Hsu, W.N., Auli, M., Baevski, A., 2022. 按需减少计算资源：随机wav2vec 2.0。arXiv abs/2204.11934。Wu, F., Kim, K., Pan, J., Han, K.J., Weinberger, K.Q., Artzi, Y., 2022. 语音识别无监督预训练中的性能—效率权衡，载于：ICASSP 2022——2022年IEEE声学、语音与信号处理国际会议（ICASSP），第7667–7671页。doi: 10.1109/ICASSP43922.2022.9747432。Xu, Q., Baevski, A., Auli, M., 2021. 简单有效的零样本跨语言音素识别。arXiv abs/2109.11680。Zahrer, A., Zgank, A., Schuppler, B., 2020. 探索构建语言记录自动转录系统：来自穆尤语的经验，载于：第12届语言资源与评测会议论文集，欧洲语言资源协会，法国马赛。第2893–2900页。网址: https://aclanthology.org/2020.lrec-1.353。