arXiv:2305.05201v1 [cs.CL] 9 May 2023

日语自监督语音表征模型的语言依赖性探究

Takanori Ashihara、*Takafumi Moriya*、*Kohei Matsuura*、*Tomohiro Tanaka*

日本 NTT株式会社

摘要

自监督学习（SSL）在单语和跨语言场景中都取得了巨大成功。然而，由于这两种场景通常是分别研究的，因此鲜有研究关注跨语言模型相较于单语模型的效果如何。本文以日语自动语音识别任务为例，通过实证研究探讨这一基本问题。首先，在尽可能保持声学领域一致的情况下，我们比较跨语言模型和单语模型在两种不同语言任务上的自动语音识别性能。随后，我们考察需要收集多少日语无标注数据，才能达到与使用数万小时英语和／或多语言数据预训练的跨语言模型相当的性能。最后，我们全面研究自监督学习在日语中的有效性，并展示其在多项自动语音识别任务上达到最先进水平。鉴于目前尚无针对日语的全面自监督学习研究，我们希望本研究能为日语自监督学习研究提供指引。

索引词— 自监督学习，语音表征，语言依赖性，自动语音识别，低资源

1. 引言

近年来，自监督学习（SSL）在语音领域 [1]引起了广泛关注。由于 SSL 属于无监督表征学习，预训练模型能够捕捉无标注数据中所包含的语音信息。研究者提出了多种自监督学习方法，并取得了令人瞩目的性能，尤其是在英语自动语音识别（ASR）任务中 [2–7]。后续研究还表明，这些方法具有良好的任务泛化能力，在说话人验证和情感识别等多种下游任务上的表现优于人工设计的特征（例如，对数梅尔滤波器组输出） [8,9]。

除任务泛化能力外，自监督学习模型还展现出语言泛化能力，有助于缓解数据稀缺问题。例如，一些研究构建了多语言或英语预训练模型，以提升英语和低资源语言的性能 [10–13]。这种通用的—
语音表征使低资源语言中的语音任务更加普及，因此，目标语言的各种模型都可通过 HuggingFace 模型库等开放平台公开获取[1]。在上述研究和已发布的模型中，为评估其向其他语言的迁移能力，下游非英语模型通常是在迁移学习的背景下构建的，即使用少量目标语言数据对多语言模型或英语模型进行微调（图 1 中所示的模型 (a)）。

然而，尽管跨语言模型取得了成功，人们对语言迁移能力的有效性仍知之甚少-

[1]https://huggingface.co/models

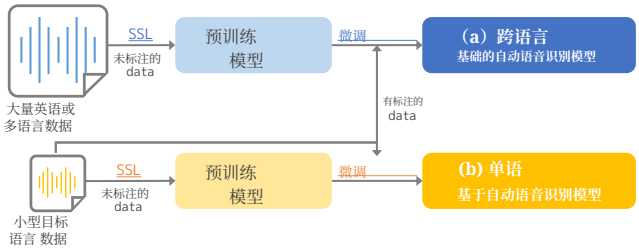


图 1。针对目标语言的基于自监督学习的模型。

多语言模型的性能如何？与在单语设置下进行预训练和微调的单语模型相比（图 1 中的模型 (b)），多语言模型表现如何？换句话说，自监督学习模型对语言的依赖程度有多强？更具体地说，我们可以问：如果能够为目标语言收集到足够的训练数据，单语模型能否比跨语言模型取得更高的准确率？如果可以，预训练需要多少目标语言的无标注数据，才能取得比使用数万小时数据进行预训练的多语言/英语模型更高的准确率？由于多次尝试自监督学习的计算与能源成本高昂，这个根本性问题对于判断预训练能否达到足够的性能至关重要。例如，一些低资源领域无法享受到泛化能力带来的益处 [14],，因此，这一启示有助于确定自监督学习的潜在应用。此外，在实际场景中，对于拥有一定数量日常使用者的语言（如亚洲语言或印度语言），通常可以收集数百小时的非英语无标注数据。

本文中，我们重点研究使用开放获取语料的日语自监督学习。首先，我们比较图 1 中所示模型 (a) 和 (b) 在 JNAS 数据集 [15] 上的自动语音识别性能，以确定自监督学习的语言依赖性。由于 JNAS 是一个低资源语料库，包含一组日语报纸朗读语音，其声学领域与 CommonVoice [16], LibriSpeech [17], 多语言 LibriSpeech [18]，和 Libri-Light [19] 相似，唯一区别在于语言。其次，我们分析日语无标注数据量与性能之间的关系，以выяс定在 JNAS 任务上达到与多语言自监督学习模型相当的水平需要多少数据。上述探索结果表明，语言依赖性比声学依赖性更为关键；此外，使用 200 小时数据预训练的单语模型，其性能可与多语言模型相媲美。最后，据我们所知，与其他非英语语言 [20–22]，不同，目前尚无关于日语自监督学习的论文，因此我们全面研究了一种基于自监督学习的自动语音识别模型，并在多项日语自动语音识别任务上取得了最先进的错误率。[2]

[2]一些日语预训练模型将通过以下网址发布，以加速日语自监督学习研究：https://huggingface. co/ashi-ta/japanese-pretrained-ckpts

2. 相关工作

先前的研究报告了数据集领域对自监督学习（SSL）的影响。例如，[23] 研究了预训练所使用的领域内和领域外数据集如何影响下游自动语音识别（ASR）的性能，并证明领域内数据集在任何情况下都具有良好的效果。[24] 还探讨了数据集偏差的影响，例如说话者的性别或语速。这些研究主要关注无标注数据的数据集领域对英语下游任务性能的影响程度，而对于语言可迁移性和语言依赖性，尚未进行充分阐明。由于已有研究指出，自监督学习表征可能与音素结构部分相关，且这种相关性与语言存在一定联系，[2]，本文研究了语言依赖性对下游自动语音识别系统性能的影响程度。

3. 方法

3.1. 自监督学习

3.1.1. wav2vec 2.0

本文聚焦于各种自监督学习方法中的 wav2vec 2.0 [3]。一般而言，预训练模型的架构由两个主要组件构成：

$$\mathbf{H}^0 = \mathrm{Convolution}(\mathbf{X}), \tag{1}$$

$$\mathbf{H}^l = \mathrm{Encoder}(\mathbf{H}^{l-1}), \tag{2}$$

其中，卷积(·)是一个多层卷积网络（卷积神经网络），用于将输入音频 $\mathbf{X}$ 映射为潜在特征 $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_T\}$，并进行下采样。经过第一个组件后，表征被输入到编码器(·)模型（如 Transformer）中，以输出隐藏语音表征 $\mathbf{H}^l = \{\mathbf{h}_1^l, \cdots, \mathbf{h}_T^l\}$。其中，$l$ 和 T 分别表示编码器(·)的块索引和输出序列长度。我们将编码器的最终输出 $\mathbf{H}^l$ 定义为上下文表征 $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_T\}$。

对于训练目标，wav2vec 2.0 使用从$\mathbf{Q} = \{\mathbf{q}_1, \cdots, \mathbf{q}_T\}$ 计算得到的量化表示作为训练目标。在预训练时，编码器(·)的输入在一定比例的时间步上被掩码，并根据$\mathbf{C}$ 和$\mathbf{Q}$计算对比损失。

3.1.2. 输入音频特征

虽然卷积(·)的输入通常采用原始波形（WAV），但深层卷积神经网络可能无法从少量 WAV 中获取足够的特征表示。因此，我们不仅研究将 WAV 输入自监督学习模型 [25]，还研究将对数梅尔滤波器组输出（滤波器组特征）作为输入。具体而言，我们从原始音频中提取 80 维滤波器组特征，窗口宽度为 25 ms，帧移为 10 ms。为对特征进行归一化，我们对数据集应用预先估计的全局倒谱均值方差归一化（倒谱均值方差归一化）。采用滤波器组特征后，卷积(·)的结构也从原始 wav2vec 2.0 模型中使用的 7 层一维卷积神经网络改为 2 层二维卷积层。每个二维卷积层有 32 个通道，卷积核大小为 3，步幅为 2，因此输出的时间分辨率相较滤波器组特征序列降低为四分之一，从而减少 GPU 内存消耗。

3.1.3. 编码器架构

除音频输入的类型外，我们还对比了用于编码器（·）组件的 Conformer [26] 编码器与 Transformer 编码器。近期研究表明，Conformer 编码器对英语 [25, 27]，自监督学习的有效性已得到验证，但目前尚无针对日语的比较研究。本文采用一种 Conformer 模块，该模块由带相对位置编码的多头自注意力模块、卷积模块和前馈模块组成。

表1。使用 ESPnet 配方进行预处理后，日语数据集的领域及数据量（小时）。每列分别显示训练集、开发集和多个评估集的数据量。根据训练数据量，四项任务可分为上方两行（高资源）和下方两行（低资源）。

数据集	领域	训练	Dev	Test
CSJ	自发的语音	515.6	6.5	开发集 / 评估集1 / 评估集2 / 评估集3 6.5 / 1.8 / 1.9 / 1.3
LTVS	广播语音	2031.2	5.0	开发集 / 开发集-4k / TEDx 日语-10k 5.0 / 13.7 / 8.9
JNAS	报纸朗读语音	47.8	2.5	开发集 / 测试集-100 / 测试集-500 2.5 / 1.6 / 1.9
JSUT	Text 朗读语音	9.7	0.3	开发集 / 评估集1 0.3 / 0.3

3.2. 自动语音识别（ASR）系统

在自监督学习后，使用连接时序分类（CTC）[28] 准则优化预训练模型。在微调时，我们研究类似于 SpecAugment [29] 的特征掩蔽应在何处应用。在原始 wav2vec 2.0 论文中，此操作应用于卷积（·）之后（后置卷积神经网络），进行时间／通道掩蔽。然而，输入滤波器组特征时，同样的变换可以应用于卷积（·）之前（前置卷积神经网络），进行时间／频谱掩蔽，这与原始 SpecAugment 更为相似。由于除了编码器（·）之外，卷积模块也可以进行正则化，因此基于滤波器组特征的模型的自动语音识别性能可以进一步提升。

4. 实验设置

4.1. 数据集

对于日语语音数据，我们使用了四个公开可用的语料库：自发日语语料库（CSJ）[30], Laboro TVSpeech（LTVS）[31], 日语报纸文章句子语料库（JNAS）[15], 以及东京大学猿渡研究室的日语语音语料库（JSUT）[32]。表 1 总结了使用 ESPnet [33]中的各个配方完成数据准备后，各数据集所属的领域及其数据量。由于 JNAS 由朗读报纸文章的说话者录制，因此其声学领域与 LibriSpeech（LS）[17], CommonVoice [16], Multilingual LibriSpeech [18] 和 Libri-Light [19], 较为接近，这些语料库均来源于有声读物或朗读句子。另一方面，CSJ 包含学术报告演讲和模拟公众演讲的集合，与 TED-LIUM-v2（TED2）语料库 [34] 类似，后者采集自 TED 演讲。

对于自监督学习，我们仅使用高资源语料作为无标注数据，具体包括 CSJ、LTVS 和 LS。对于微调，我们定义了六项使用全部有标注数据的自动语音识别任务，并按资源情况分为两类：CSJ、LTVS 和 LS 属于高资源情形，JNAS、JSUT 和 TED50h 属于低资源情形。这里的 TED50h 是从 TED2 语料中随机抽取的英语数据集，时长仅为 50 小时，使其数据量大致与 JNAS 相当。除非另有说明，我们使用各语料库中的全部数据。

4.2. 训练与评估

4.2.1. 预训练

本文探索的模型架构基于 [3]，Transformer 编码器采用与 wav2vec 2.0 BASE 模型相同的设置。对于 Conformer 编码器，我们使用了 12 个模块，每个模块包含一个嵌入维度为 768、具有 12 个注意力头的多头自注意力模块、一个卷积核大小为 31 的卷积模块，以及一个维度为 3072 的前馈模块。我们基本上采用 16 位浮点数（FP16）进行运算，仅在遇到梯度溢出问题时使用 FP32。请注意，预训练和微调时若以滤波器组特征作为输入，与输入长度相关的超参数会从采样点单位转换为帧单位。例如，若 WAV 输入的每个 GPU 最大批次大小为 87.5 秒，则滤波器组特征输入对应 8750 帧。

在我们的实验中，我们还与 fairseq 提供的已发布预训练模型进行了比较。[3]

4.2.2. 微调

我们根据第 4.1 节所述的资源情况，采用两种配置对自监督学习模型进行微调。对于高资源情况，我们采用与使用 LS 960h [3] 微调 wav2vec 2.0 BASE 模型时相同的设置，但下文所述的掩蔽参数除外。另一种配置适用于低资源情况，训练步数上限降至 80 k，学习率峰值降至 3×10^{-5}。如第 3.2 节所述，在微调期间，我们采用两种掩蔽位置及超参数：对于后置卷积神经网络，时间（通道）掩蔽的长度为 10（64），概率为 0.5（0.1）；对于前置卷积神经网络，时间（频谱）掩蔽的长度为 20（30），概率为 0.65（0.1）。此外，仅在采用后置CNN掩码时，我们才冻结特征编码器卷积(·)。所有任务的输出词汇表均为字符；CSJ、LTVS、JNAS、JSUT、LS 和 TED50h 任务的词汇表数量分别为 3259、3949、2272、2739、28 和 31。

4.2.3. 解码

我们使用一个对多个检查点的权重取平均的CTC模型进行了维特比解码。在微调期间，我们针对高资源任务每6400步评估一次模型，针对低资源任务每1600步评估一次，并保存排名前5的检查点用于平均。这里，评估指标为日语任务的字符错误率（CER）和英语任务的词错误率（WER）。

上述所有训练和解码过程均已实施-
并使用 fairseq[4] [35] 实现。

5. 结果

这里，我们首先介绍第5.1节至第5.3节的分析结果，然后给出基于自监督学习的自动语音识别模型在各种日语任务上的性能。为简洁起见，除第5.4节外，我们均使用经过后置CNN掩码微调的基于Transformer的自监督学习模型，

5.1. WAV 与滤波器组特征的比较

首先，为了验证实验中所用输入音频类型的影响，我们比较了输入 WAV 和滤波器组特征的模型性能。表2展示了 LS 和 CSJ 模型的结果，这些模型使用相同的数据进行预训练和微调

[3]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec[4] https://github.com/facebookresearch/fairseq

表2。在相同数据上进行预训练和微调时，输入类型对 LS 和 CSJ 的影响。请注意，LS 的 WAV输入模型是公开提供的 wav2vec 2.0 BASE。

（未）标注 data	输入 type	WER			
		开发集-干净	开发集-其他	测试集-干净	测试集-其他
LS	WAV	3.2	8.9	3.4	8.4
	滤波器组特征	3.2	9.0	3.4	8.7
（无）标签 data	输入 type	CER			
		dev	评估集1	评估集2	评估集3
CSJ	WAV	4.0	4.7	3.5	3.8
	滤波器组特征	3.4	4.6	3.2	3.7

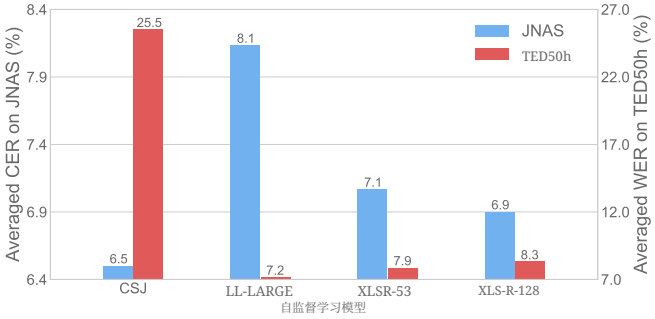


图2。不同自监督学习模型在 JNAS（左侧柱）和 TED50h（右侧柱）任务上的评估结果。LL-LARGE 的结果取自发布的、在 Libri-Light 上预训练的 wav2vec 2.0 LARGE 模型 [19]。

集。LS 模型的性能没有明显下降，而 CSJ 模型的性能反而有所提升。这可能是因为 CSJ 的数据量约为 LS 的一半，导致难以优化将 WAV 编码为潜在特征的深层卷积神经网络。基于这些结果，后续分析基本上使用输入滤波器组特征的日语自监督学习模型。

5.2. 语言依赖性

本节中，我们使用四种自监督学习模型分析语言依赖性：在 CSJ 上预训练的日语模型、公开提供的在 Libri-Light 上预训练的 wav2vec 2.0 LARGE 模型 [19], XLSR-53 [12], 以及 XLS-R-128[5] [13]。图2展示了两项日语（JNAS；左侧柱）和英语（TED50h；右侧柱）低资源自动语音识别任务的性能。请注意，每项错误率均为该任务中所有测试集的平均值。结果表明，多语言模型展现出明显的语言泛化能力（XLSR-53 为 7.1%/7.9%，XLS-R-128 为 6.9%/8.3%），但每种单语模型在相应语言的任务上错误率更低；CSJ 模型在 JNAS/TED50h 任务上的错误率为 6.5%/25.5%，LL-LARGE 模型则为 8.1%/7.2%。由于 CSJ 和 LL-LARGE 自监督学习模型的数据集领域分别与 TED50h 和 JNAS 相似，这表明，在同一种语言上进行预训练和微调，比在相同声学领域上进行训练更为关键。

5.3. 无标注数据量

第 5图2 展示的结果突显了自监督学习的语言依赖性，而非声学依赖性，也体现了多语言模型明显的语言泛化能力。不过，只有在收集到足量目标语言的无标注数据时（例如 CSJ），上述结果才有可能实现。

[5]我们使用了通过以下链接获取的 XLS-R (0.3B)，其基于与 XLSR-53 相同的架构，并在 128种语言上进行了预训练：
https://github.com/facebookresearch/fairseq/tree/ main/examples/wav2vec/xlsr

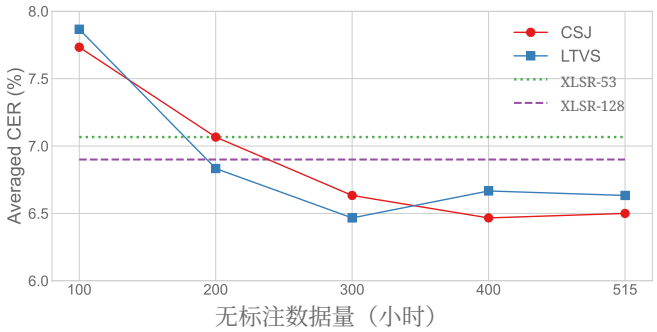


图3。在 JNAS 上使用不同数量的无标注数据进行评估的结果；对于所有模型而言，JNAS 均属于域外设置。

因此，我们假设日语无标注数据量各不相同的若干种情形，并探究需要收集多少无标注数据才能使多语言模型的性能优势发生逆转。具体而言，在这项分析中，我们使用五种不同数据规模构建日语自监督学习模型，并与多语言模型进行比较；同时，在 LTVS 和 CSJ 上开展相同实验以作验证。图 3 展示了日语无标注数据量与 JNAS 任务三个评估集平均字符错误率之间的关系。结果表明，使用约 200 小时数据预训练的模型取得了与多语言模型相当的性能；对于 CSJ 和 LTVS，这一趋势似乎也几乎相同。

5.4. 日语自动语音识别任务的比较

在第5.3节中，使用近 200 小时的无标注数据后，多语言模型的性能追上了单语模型；这一数据量远少于仅使用日语开放获取语料库所需的数据量。因此，我们结合高资源数据（即 CSJ 和 LTVS）构建了自监督学习模型，并比较了使用第3节所介绍的训练技术进行微调的各种模型。表3展示了四项日语任务中两个基线模型（B1–B2）、日语单语自监督学习模型（M1–M6）、跨语言模型（C1、C2）以及现有模型（E）的字符错误率。除掩码位置配置外，各任务的性能趋势相似，因此受篇幅所限，我们仅报告 CSJ 任务的完整结果，以及其他任务的最佳性能。请注意，对于低资源任务，我们仅展示使用后置卷积神经网络微调的模型性能，因为与高资源任务不同，该模型的表现优于使用前置卷积神经网络微调的模型。从 CSJ 的结果可以看出，自监督学习模型（C1、C2、M1–M6）优于基线模型（B1–B2），其中日语自监督学习模型（M1–M6）的效果尤为显著。为验证其他语言的数据是否能提高目标语言任务的准确率，我们还测试了使用 CSJ、LTVS 和 LS 的组合进行预训练的模型（M6），但该模型在任何任务上都没有带来进一步提升。据我们所知，我们模型的字符错误率在大多数测试集上优于现有报告（E）中的最先进性能。

[6]https://github.com/espnet/espnet/blob/master/egs2/laborotv/asr1/README.md[7] https://github.com/espnet/espnet/blob/master/ egs/jnas/asr1/RESULTS[8] https://github.com/espnet/espnet/blob/master/egs2/jsut/asr1/README.md

表3。CSJ、LTVS、JNAS 和 JSUT 在不同无标注数据、编码器及掩码策略下的评估结果。请注意，每项任务的微调仅使用该任务中的有标注数据（例如，CSJ 任务中的 515.6 小时有标注数据）。第四列中的 T 和 C 分别表示 Transformer 和 Conformer 编码器，第三列中的 pre 和 post 分别表示前置卷积神经网络掩蔽和后置卷积神经网络掩蔽。

CSJ

ID	未标注 data	Enc.	Mask	CER dev	评估集1	评估集2	评估集3
B1	-	T	pre	4.2	5.3	3.8	4.1
B2	-	C	pre	4.1	5.0	3.8	3.9
M1	CSJ	T	post	3.4	4.6	3.2	3.7
M2	CSJ+LTVS	T	post	3.5	4.5	3.2	3.7
M3	CSJ+LTVS	T	pre	3.5	4.2	3.3	3.7
M4	CSJ+LTVS	C	post	3.3	4.2	3.0	3.5
M5	CSJ+LTVS	C	pre	**3.1**	**3.9**	**2.9**	**3.2**
M6	CSJ+LTVS+LS	C	pre	3.4	4.3	3.1	3.4
C1	XLSR-53 [12]		post	3.8	4.7	3.6	3.8
C2	XLS-R-128 [13]		post	3.7	4.7	3.5	3.8
E	Karita 等人 [36]			**3.1**	4.1	3.2	3.5

LTVS

ID	未标注 data	Enc.	Mask	CER dev	开发集-4k	TEDx 日语-10k
B1	-	T	pre	10.0	8.0	13.4
B2	-	C	pre	9.4	7.7	13.1
M5	CSJ+LTVS	C	pre	**8.7**	7.4	**12.6**
E	ESPnet[6]			9.7	**7.0**	13.0

JNAS

ID	未标注 data	Enc.	Mask	CER dev	测试集-100	测试集-500
B1	-	T	pre	12.0	8.9	12.4
B2	-	C	pre	14.7	11.1	14.4
M4	CSJ+LTVS	C	post	**5.9**	**3.7**	**5.5**
E	ESPnet[7]			11.0	8.9	11.3

JSUT

ID	未标注 data	Enc.	Mask	CER dev	评估集1
B1	-	T	pre	30.0	31.5
B2	-	C	pre	33.4	34.9
M4	CSJ+LTVS	C	post	**11.7**	**12.6**
E	ESPnet[8]			12.0	13.9

6. 结论

本文通过比较日语自动语音识别任务中的单语模型和跨语言模型，对语言依赖性进行了实证研究与分析。研究结果表明，在预训练和微调过程中，基于自监督学习的模型受语言因素的影响大于受声学因素的影响；此外，使用 200 小时数据预训练的单语模型，其准确率可与跨语言模型相媲美。基于这一观察，我们对原始 wav2vec 2.0 模型进行了改进，构建了多种模型，并在多个日语自动语音识别任务上取得了最先进的性能。

7. 参考文献

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, “自监督语音表征学习：综述”, *IEEE* 信号处理专题期刊，2022。[2] A. v. d. Oord, Y. Li, and O. Vinyals, “基于对比预测编码的表征学习”，*arXiv preprint arXiv:1807.03748*，2018。[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0：自监督语音表征学习框架”，载于神经信息处理系统大会，2020。[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhut- dinov, and A. Mohamed, “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，2021。[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “WavLM：面向全栈语音处理的大规模自监督预训练”，*IEEE* 信号处理专题期刊，2022。[6] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “data2vec：面向语音、视觉和语言自监督学习的通用框架”，*ICML*，2022。[7] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, “w2v-BERT：结合对比学习与掩码语言建模的自监督语音预训练”，载于自动语音识别与理解研讨会，2021。[8] S. w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. y. Lee, “SUPERB：语音处理通用性能基准”，载于*Interspeech*，2021。[9] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, “深度与宽度之争：面向自监督语音模型任务无关知识蒸馏的学生架构分析”，载于*Interspeech*，2022。[10] M. Rivi`ere, A. Joulin, P.-E. Mazar´e, and E. Dupoux, “无监督预训练可在不同语言间有效迁移”，载于*ICASSP*，2020。[11] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord, “学习稳健的多语言语音表征”，载于自然语言处理实证方法会议，2020。[12] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “用于语音识别的无监督跨语言表征学习”，载于*Interspeech*，2021。[13] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “XLS-R：大规模自监督跨语言语音表征学习”，载于*Interspeech*，2022。[14] L. P. Violeta, W. C. Huang, and T. Toda, “病理语音识别的自监督预训练框架研究”，载于*Interspeech*，2022。[15] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS：用于大词汇量连续语音识别研究的日语语音语料库”，日本音响学会志，1999。[16] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Hen- retty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common Voice：大规模多语言语音语料库”，载于*LREC*，2020。[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech：基于公有领域有声读物的自动语音识别语料库

” 中*ICASSP* ，2015。

[18] V. Pratap、Q. Xu、A. Sriram、G. Synnaeve 和 R. Collobert，《MLS：用于语音研究的大规模多语言数据集》，发表于 Interspeech，2020。[19] J. Kahn、M. Rivi`ere、W. Zheng、E. Kharitonov、Q. Xu、P. E. Mazar´e、J. Karadayi、V. Liptchinsky、R. Collobert、C. Fuegen、T. Likhomanenko、G. Synnaeve、A. Joulin、A. Mohamed 和 E. Dupoux，《Libri-Light：用于有限监督或无监督自动语音识别的基准》，发表于 ICASSP，2020。[20] K. Deng、Z. Yang、S. Watanabe、Y. Higuchi、G. Cheng 和 P. Zhang，《利用预训练声学模型和语言模型改进非自回归端到端语音识别》，发表于 ICASSP，2022。[21] R. Al-Ghezi、Y. Getman、A. Rouhe、R. Hild´en 和 M. Ku-rimo，《面向低资源二语瑞典语的自监督端到端自动语音识别》，发表于 Interspeech，2021。[22] K. D. N、P. Wang 和 B. Bozza，《使用大型自监督模型进行低资源语音识别》，发表于 Interspeech，2021。[23] W.-N. Hsu、A. Sriram、A. Baevski、T. Likhomanenko、Q. Xu、V. Pratap、J. Kahn、A. Lee、R. Collobert、G. Synnaeve 和 M. Auli，《稳健的 wav2vec 2.0：分析自监督预训练中的领域偏移》，发表于 Interspeech，2021。[24] Y. Meng、Y.-H. Chou、A. T. Liu 和 H.-y. Lee，《别说得太快：数据偏差对自监督语音模型的影响》，发表于 ICASSP，2022。[25] Y. Zhang、J. Qin、D. S. Park、W. Han、C.-C. Chiu、R. Pang、Q. V. Le 和 Y. Wu，《突破自动语音识别半监督学习的极限》，arXiv 预印本 arXiv.2010.10504，2020。[26] 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》，发表于 Interspeech，2020。[27] A. Sriram、M. Auli 和 A. Baevski，《Wav2Vec-Aug：利用有限数据改进自监督训练》，发表于 Interspeech，2022。[28] A. Graves、S. Fernandez、F. Gomez 和 J. Schmidhu-ber，《连接时序分类：利用循环神经网络对未分段序列数据进行标注》，发表于 ICML，2006。[29] D. S. Park、W. Chan、Y. Zhang、C.-C. Chiu、B. Zoph、E. D. Cubuk 和 Q. V. Le，《SpecAugment：一种用于自动语音识别的简单数据增强方法》，发表于 Interspeech，2019。[30] K. Maekawa，《自发日语语料库：设计与评估》，发表于 SSPR，2003。[31] S. Ando 和 H. Fujihara，《基于电视录音构建大规模日语自动语音识别语料库》，发表于 ICASSP，2021。[32] R. Sonobe、S. Takamichi 和 H. Saruwatari，《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 预印本 arXiv:1711.00354，2017。[33] S. Watanabe、T. Hori、S. Karita、T. Hayashi、J. Nishitoba、Y. Unno、N. Enrique Yalta Soplin、J. Heymann、M. Wiesner、N. Chen、A. Renduchintala 和 T. Ochiai，《ESPnet：端到端语音处理工具包》，发表于 Interspeech，2018。[34] A. Rousseau、P. Del´eglise 和 Y. Est`eve，《通过精选数据扩充 TED-LIUM 语料库，以用于语言建模和更多 TED 演讲》，发表于 LREC，2014。[35] M. Ott、S. Edunov、A. Baevski、A. Fan、S. Gross、N. Ng、D. Grangier 和 M. Auli，《fairseq：一种快速、可扩展的序列建模工具包》，发表于 NAACL-HLT，2019。[36] S. Karita、Y. Kubo、M. A. U. Bacchiani 和 L. Jones，《日语语音识别神经网络架构与训练方法的比较研究》，发表于 Interspeech，2021。