arXiv:2302.01496v1 [cs.CL] 3 Feb 2023

高效 语音基础模型的领域自适应

Bo Li、Dongseong Hwang、Zhouyuan Huo、Junwen Bai、Guru Prakash Tara N. Sainath、Khe Chai Sim、Yu Zhang、Wei Han、Trevor Strohman、FrancoiseBeaufays

谷歌 LLC，美国{

boboli,tsainath}@google.com

摘要

基础模型（FM）通过大规模、多样化数据进行训练，并能够适配广泛的下游任务，因此引起了研究界的广泛关注。得益于来自不同模态、语言和应用领域等多样化数据源，基础模型展现出强大的泛化能力和知识迁移能力。本文介绍了一项开创性研究，旨在为基于基础模型的语音识别系统构建高效解决方案。我们采用近期开发的自监督 BEST-RQ 进行预训练，并提出使用 JUST Hydra，结合源域数据和无监督目标领域数据进行联合微调。随后，使用少量有监督域内数据，将基础模型编码器适配器和解码器微调至目标域。在大规模 YouTube 和语音搜索任务上，结果表明，我们的方法在数据和模型参数方面都十分高效。与使用额外 3 亿有监督域内数据从头训练的 7.311 亿参数模型相比，我们的方法仅需 2160 万有监督域内数据和 1.308 亿微调参数，即可达到相同的质量。

索引词——基础模型、领域自适应

1. 引言

人们通常希望大规模模型能够基于广泛的数据进行训练，因为它们可以适应各种下游任务。一项颇具影响力的研究方向是基础模型（FM）[1], , 它在研究界引起了极大的关注。基础模型展现出强大的泛化和知识迁移能力 [2,3], , 这得益于它们从多样化的数据源中学习，例如不同的模态、多种语言和各种应用领域。在语音领域，已有大量研究展示了此类模型的良好效果，并证明了它们的潜在优势[4–10]。根据是否使用有监督训练数据，可以将基础模型分为两类：自监督预训练模型 [11, 12] 和有监督多任务训练模型[13]。通过自监督训练，模型首先使用对比损失 [14] 或重构损失 [15], 在仅音频数据上进行训练，以学习良好的语音信号表征。随后，这些模型可直接用作下游任务的特征提取器 [16, 17]。由于不需要标签信息，这种方法能够轻松扩展到更多样的语音数据，而无需人工转录。与之类似，在有监督多任务学习中[13], , 不同任务被统一为一个异构判别任务，模型在这些任务上联合训练，例如多领域任务 [5, 18] 或多语言任务[19, 20]。这种方法的前提是，基础模型所训练的任务需要具备一些带标签的数据。近期研究还发现，自监督预训练可以提升标签数据效率。在本文中，我们主要关注通过这一流程训练的基础模型，以获得更好的识别效果。

现有工作主要集中于使用有监督域内数据，为目标任务联合训练或微调基础模型 [5, 7]。当出现新的领域时，使用有监督域内数据就需要重新训练基础模型。为此，我们提出利用公共领域 YouTube 数据构建基础模型。YouTube 汇集了来自全球 100 多个不同国家、涵盖 80 种语言的多样化语音数据，覆盖众多领域 [21]。数据量也十分庞大 [22]。这些特点使其成为构建语音基础模型的绝佳数据来源，而且这些模型还可在工业界和学术界之间共享，以促进合作。为了获得更高质量，基础模型往往规模庞大。微调此类模型会造成资源效率低下且耗时。能够高效地将基础模型适配到目标任务的技术至关重要。残差适配器 [23–25], 、提示学习 [26] 和神经重编程 [27]等现有工作已展现出这类技术的潜力。

我们的工作在多个方面对基础模型学习作出了贡献。首先，wav2vec 2.0 [12]和 wav2vec-BERT [28] 等传统的预训练+微调两阶段方案，会在微调期间更新预训练编码器；然而，对于大型基础模型而言，这通常代价高昂。因此，我们提出在微调期间冻结预训练基础模型，仅更新为目标域新增的轻量适配器。我们证明，该方法适用于 CTC、LAS 和 RNN-T 解码器。其次，大多数基础模型使用来自目标域的标注数据微调模型。我们证明，从源域（YouTube）恰当选择数据也能缩小源域与目标域之间的差距，并进一步提升模型在目标域（语音搜索）中的性能。第三，即便没有来自目标域的音频文本配对数据，我们也证明仅使用音频数据或纯文本数据都能助力基础模型微调。我们给出了完整方案，说明如何构建能够高效适配目标域的基础模型。尽管本文中的一些技术此前已有研究，但本文创新性地引入了一个统一框架，利用基础模型为下游任务构建高质量的语音解决方案。在这一框架中，我们探索了数据效率和模型效率技术，以构建与目标域建立粗略联系的基础模型，并使其能够利用有限的有监督域内数据快速适配。这为探索进一步提升基础模型在语音任务中能力上限的技术开辟了新的研究方向。

2. 方法

本研究聚焦于将基于 YouTube 数据构建的基础模型进行领域自适应，使其适用于目标域语音搜索。我们致力于以高成本效益的方式进行自适应，从而复用规模庞大的基础模型，并最大限度地减少微调工作。此外，日益严峻的隐私问题促使我们更好地保护

语音搜索等个人数据，避免其被广泛用于自动语音识别训练，同时更有效地利用 YouTube 等公开数据。在本节中，我们提出了多种高效的自适应技术，以提升自动语音识别质量。

2.1. 语音基础模型

我们采用 6 亿参数的 ConformerXL [7] 架构来构建语音基础模型。源域 YouTube 包含极其广泛的子域 [18, 21]。这种多样性对于确保基础模型具备良好的泛化能力至关重要，也有助于完成各种下游任务。在第一阶段，基础模型使用仅音频数据进行预训练。不同于 BigSSL [7],，我们采用基于 BERT 的语音预训练方法——随机投影量化器（BEST-RQ），该方法在许多任务上的表现优于 wav2vec 2.0 和 wav2vec-BERT [29]。该技术会随机初始化一个投影矩阵和一个码本。每个输入语音帧都会通过这个随机矩阵投影到隐空间中。随机码本中最近向量的索引将用作该帧的离散标签令牌。随后，我们可以对语音信号的这一离散表示进行 BERT 训练：首先对其进行掩码，然后优化模型，使其预测被掩码部分的标签令牌。在训练过程中，随机初始化的投影矩阵和码本始终保持冻结。输入语音数据经过归一化，使其均值为零、标准差为一。我们发现，这种归一化对于防止随机投影坍缩到码本中的少数几个向量至关重要。在第二阶段，我们使用源领域数据进行有监督微调。微调后的模型用作基础模型基线。

2.2. 领域自适应

领域自适应将语音基础模型与特定应用场景相连接。鉴于基础模型规模庞大且训练数据多样，高效的自适应是推动其在各种下游任务中实际应用的关键。在这项工作中，我们致力于实现两个目标：

- 模型效率：最好仅在目标域上训练少量参数，并最大限度地共享可跨领域和任务复用的模型参数。在域内数据上微调的参数越少越好。

- 数据效率：在多样化数据上进行自监督训练，加上基础模型的大容量，使其在下游任务中具有更高的标注数据效率 [7]。此外，如果还能利用语音专用和纯文本等丰富的域内数据，那就更好了。所需的有监督数据和域内数据越少越好。

就质量而言，自适应技术应达到与使用大量有监督域内数据对整个模型进行微调相近的质量。

2.2.1. 模型效率

考虑到有监督基础模型在质量上的优势，了解不同解码器是否会影响目标域中的质量将很有意义。在三种常用的解码器 CTC [30], LAS [31] 和 RNN-T [32], 中，CTC 是最简单的一种，它基本上只在编码器之上添加一个 Softmax 层。LAS 使用注意力机制，在每个解码步骤汇总整个话语，这可能带来潜在的质量优势，而其他解码器通常不会显式利用未来帧。此外，RNN-T 已显示出比注意力模型更好的长语音鲁棒性 [33]。因此，本文对这三种解码器进行了比较。

在编码器端，我们采用残差适配器 [23]，以便在适配目标域的同时冻结 FM编码器。文献中还有许多其他变体，例如前缀调优、LoRA 和并行适配器 [26]。然而，在我们的任务中，这些变体并未带来质量提升，因此我们仅采用标准残差适配器。

2.2.2. 数据效率

数据效率的目标是减少适配过程中使用的域内数据量。使用的有监督域内语音-文本对越少越好。此外，与配对数据相比，使用未配对的音频或文本数据更为理想。

源领域数据过滤：YouTube 源域包含各种各样的子域，例如不同的话题、风格、说话人等。其中很可能存在与目标域相似的数据。为了提升 FM 在目标域上的质量，我们可以筛选出这类数据，并用它们定制 FM。这有可能减少所需的域内数据量。对于语音搜索，语音数据大多由单个说话人录制，且搜索用途通常更偏好书面形式的转录文本。基于这些假设，我们通过说话人聚类筛选 YouTube 数据，找出仅包含单个说话人且转录文本的书面形式与口语形式一致的话语。我们发现，无需任何域内数据，这种方法也有助于减少 FM 在目标域上的识别错误。还有许多其他过滤技术，例如 [34], , 将在未来进行探究。

仅音频域内数据：源域与目标域之间一种潜在的不匹配是声学差异。为了提升基础模型在目标域上的质量，我们可以引入仅音频域内数据来解决这种不匹配。联合监督与无监督训练（JUST）[35] 及其变体 JUST Hydra [9] 已证明联合训练的优势。在这种设置下，我们通过联合训练，将源域的语音-文本配对数据（监督）与目标域的仅音频数据（无监督）结合起来。另一种常见方法是噪声学生训练（NST） [36, 37]，它利用教师模型为训练数据提供伪标签，从而使用无监督语音。我们发现，这种方法在提升质量方面非常有效。

仅文本域内数据：另一种可能的不匹配来自不同领域间的语言模型差异。各领域使用的词语及其用法可能有所不同。为了解决这一问题，我们可以将目标域文本数据注入基础模型。最简单的方法是使用在目标域训练的语言模型对识别假设重新评分。最大熵语言模型在语音搜索任务中取得了非常有前景的结果[38]。除了重新评分，在束搜索解码中融合目标域语言模型也展现出更好的效果 [39], , 本文将对此进行介绍。近期，通过联合训练注入文本数据的方法，例如 SLAM [40], MAESTRO [41]和 JOIST [42]，也取得了颇有前景的提升，这部分内容留待未来工作探讨。

3. 实验设置

3.1. 数据

3.1.1. YouTube

对于源域，我们收集了两个基于 YouTube 的数据集。无监督 YouTube 数据，即 YT-U，是一个

使用语音活动检测模型切分的多语言 YouTube 数据集[43]。该数据集为 FM预训练提供了丰富多样的语音变化。监督式 YouTube 数据，即**YT-T**，是一个仅包含英语的数据集，由用户上传了转录文本的视频组成。这些视频首先使用一个采用双向LSTM编码器的 1 亿参数 RNN-T 模型进行切分[44]。随后移除 YT教师模型识别出的非语音片段，得到约 50 万小时的无标注音频数据。不过，用户提供的转录文本会被弃用，我们使用同一个 YT教师模型生成伪标签。源域的测试集由人工转录 YouTube 上的热门视频构建，包含 11 小时音频，时长为 2 至 10 分钟。

3.1.2. 语音搜索

同样，对于目标域，我们使用从英语语音搜索流量中收集的两个数据集。无监督语音搜索数据（即 VS-U）包含 3.837 亿条话语，相当于 53.29 万小时的语音数据。这些数据已去标识化，且未转录。使用时，可以不提供转录文本，也可以使用教师模型生成的机器转录文本。有监督语音搜索数据（即 VS-L）共包含 2160 万条来自不同英语国家的英语话语，相当于 3 万小时的语音搜索流量数据，这些数据已去标识化并经过人工转录。未配对文本数据（即 VS-Text）包含超过 1000 亿条话语，规模远大于我们的音频数据集。此外，我们还纳入了 VS-L 集中的所有文本数据。测试集包含从英语语音搜索流量中抽取的 1 万条话语，与训练集没有重叠，且均经过人工转录。所有数据均已去标识化，其收集和处理遵循谷歌人工智能原则[45]。

3.2. 模型架构

输入网络的对数梅尔滤波器组特征首先经过卷积下采样模块，将帧率从 10 毫秒调整为 40 毫秒。随后使用 Conformer [46]，（卷积增强型 Transformer）。每个 Conformer 层由 8 头注意力、前馈模块和卷积模块组成。模型共有 24 个 Conformer 块，模型维度为 1024。在 Conformer 块之后使用输出投影层，以生成最终的编码器输出。这些编码器输出要么在经过额外的投影层后用作 CTC [30] 模型的输入，要么与 6 个 LSTM 层一起传递给 RNN-T [32] 解码器或 LAS [47] 解码器。每个 LSTM 层的单元维度为 768，隐藏维度为 3072，遵循 [7]。将基础模型适配到目标语音搜索域时，使用 128 维残差适配器。

3.3. 模型训练

本研究使用的语音数据均统一重采样至 16 KHZ。原生采样率不同的数据会进行升采样或降采样。我们使用 128 维对数梅尔特征，采用 32 毫秒窗口和 10 毫秒步长计算。使用 SpecAugment [48] 提高模型的抗噪鲁棒性。具体而言，使用两个最大长度为 27 的频率掩码，以及两个最大长度为 50 的时间掩码。使用 4K 子词模型（WPM）对训练转录文本进行分词。

所有模型均使用 TensorFlow 和 Lingvo [49]工具包，在谷歌的张量处理单元（TPU）V3 [50] 上进行训练，全局批次大小为 4,096 个话语。模型使用 512 个 TPU 核心进行训练，并采用同步随机梯度下降进行优化。使用 Adafactor [51]，其参数为 β_1=0.9 和 β_2=0.99。采用 Transformer 学习率计划 [52]，峰值学习率为 3e-4，预热步数为 10K。使用指数移动平均来稳定模型权重更新。

4. 结果

本节介绍我们关于将基于 YouTube 训练的基础模型适配至目标语音搜索域的实验研究。为验证其有效性，我们选用一个现有的 Conformer RNN-T 模型，该模型架构相同，在 VS-L 上训练，并额外使用了另外 3 亿个英语多领域音频文本对[4]。该模型在 VS 测试集上的词错误率为 4.3%。我们将其称为目标领域模型“T0”。本研究旨在以更少的有监督域内数据（即仅使用 2160 万条 VS-L 数据）和更少的可适配参数，达到这一目标。

4.1. 基础模型

我们在 YT-U 上使用基于 BEST-RQ 的自监督训练对 FM 编码器进行预训练（表 1中的 S0）。随后，我们添加 LAS 解码器，遵循 [7]，，并仅在 YT-L 上对解码器进行微调（S1）。较高的词错误率表明，经过自监督训练的编码器针对自动语音识别任务的优化不够充分，因此不适合直接用作自动语音识别编码器。接下来，我们对整个模型进行微调（S2），在 VS 上取得了 12.4% 的词错误率。由于基于注意力的模型存在长语音问题 [53]，，我们用 RNN-T 解码器替换 LAS 解码器（S3）。该模型在短语音 VS 测试集上的词错误率略低，在长语音 YT 测试集上的表现则明显更好。我们尝试减少解码器层数（S4），也尝试使用 CTC 解码器，在编码器与输出 softmax 层之间不设置任何层（S5）。这两种方案在 VS 上的表现都不如 S3。因此，除非另有说明，后续研究均以 S3 作为基础模型基线。

表 1：在 YT-U 上进行预训练（PT），并在 YT-T 上进行微调（FT）的不同基础模型的 WER（%）性能。

ID	解码器			训练	YT	VS
	Type	层数	参数（M）			
S0	-	-	-	PT	-	-
S1	LAS	6	164.3	解码器微调	92.2	65.7
S2	LAS	6	164.3	全部微调	62.8	12.4
S3	RNN-T	6	124.4	全部微调	14.3	12.3
S4	RNN-T	2	39.4	全部微调	14.0	12.5
S5	CTC	0	2.6	全部微调	13.6	13.9

4.2. 数据高效自适应

尽管 YouTube 中的数据具有很大的多样性，S3 的 12.3% 与域内模型 T0 相差甚远。为提升目标域上的质量，我们可以从源域中挖掘与目标域相匹配的数据。为实现快速迭代，我们使用占用内存更少且训练速度快得多的 CTC 解码器（即 S5）来验证数据选择。我们改变说话人相似度阈值，以生成 YT-L数据的不同子集，然后在表2中报告对每个相应子集上的 S6 进行微调所得的词错误率。这种简单的启发式过滤降低了目标域上的词错误率，也减少了微调阶段所需的数据量。

现有实验均只使用了源领域数据。尽管在目标域上的效果尚可，但与域内模型 T0 相比仍相差甚远。为解决这一问题，我们从无监督语音数据入手，并采用了 JUST Hydra [9] 训练。它将 YouTube 数据上的有监督微调损失与语音搜索仅音频数据上的自监督损失结合起来。加入联合

表2：使用筛选后的 YT-T 数据时的词错误率（%）比较。

ID	YT-T 小时数（K）	YT	VS
S5	500.0	13.6	13.9
D1	456.7	13.8	13.0
D2	271.8	13.9	13.2
D3	154.9	13.9	13.2
D4	13.4	13.9	14.2

训练到 S3 (G1) 后，我们可以将 VS 词错误率从 12.3% 降至 6.9%，这清楚地表明了域内数据的益处。

另一种广受欢迎的仅使用音频数据的方法是进行噪声学生训练（NST）[37]。我们首先将仅使用 YouTube 域数据训练的模型 S3 作为教师模型（G2）。使用教师转录文本确实有助于将词错误率从 12.3% 降至 11.6%，但教师模型较高的词错误率限制了改进幅度。相比之下，当我们使用目标域模型 T0 作为教师模型（G3）时，即使只使用 10% 的 VS-U 数据，也能达到 4.5% 的词错误率。这证实了构建高质量域内教师模型的必要性，也表明 VS-U 数据集存在大量冗余。开发有效的数据选择技术将很有用，我们将在未来的工作中对此进行探索。

表3：使用非配对目标域数据（VS-U、VS-Text）时的词错误率（%）比较。

ID	数据类型	数据比例	VS
S3	-	-	12.3
G1 JUST Hydra	VS-U	100%	6.9
G2 使用 S3 的 NST	VS-U	100%	11.6
G3 使用 T0 的 NST	VS-U	100%	4.5
	VS-U	10%	4.5
	VS-U	3%	4.6
	VS-U	1%	5.1
H1 重评分	VS-Text	100%	11.1
H2 浅层融合	VS-Text	100%	12.0

与配对语音-文本数据相比，文本是另一种更容易获取的目标域数据来源。我们研究了两种基本技术，即最大熵语言模型重打分（表 3 中的 H1）[38] 和 Conformer语言模型浅层融合（H2），[54], 以将 VS-Text 注入仅使用 YouTube 数据训练的基础模型。这两种方法都降低了 VS 上的词错误率，但收益相对较小。许多错误是由于 YouTube 与语音搜索之间的领域不匹配，导致模型转写了背景语音。这表明，后续注入文本可能更为可取。

通过这些消融研究，我们可以在没有任何有监督目标领域数据的情况下，通过筛选源领域数据、与目标领域纯音频数据进行联合训练，以及整合目标领域语言模型，来提升基础模型在目标域上的质量。然而，与目标领域模型 T0 相比，质量仍存在差距。

4.3. 参数高效适配

鉴于使用域内数据进行训练是最有效的方法，我们研究了使用域内数据高效适配模型的方法。我们使用少量有监督域内数据 VS-L，将 YouTube基础模型适配到 VS（表 4）。我们采用 S3，并在 VS-L 上对整个模型进行微调（E1）。其在 VS 上的词错误率为 4.4%，与 T0 的 4.3% 相近，这表明使用预训练基础模型能够提高标注数据效率 [7]。

编码器有 606.6M 个参数，约占整个模型的 83%，因此我们希望冻结这部分，以限制

表 4：使用 VS-L 数据微调的词错误率（%）对比。

ID	对比项	# 参数(M)	训练速度（话语/秒）	VS
S3	-	-	-	12.3
E1	完整模型	731.1	3100	4.4
E2	dec.	124.4	3600	8.1
E3	编码器适配器	6.4	4989	6.7
E4	编码器适配器 +解码器	130.8	3500	4.5

微调参数的数量。仅适配解码器（E2）可达到 8.1% 的词错误率。若在每个编码器层之间插入残差适配器（E3），只需适配 6.4M 个参数即可达到 6.7% 的词错误率。结合 E2 和 E3 后，只需使用 130.8M 个参数（E4）即可将词错误率降至 4.5%。与 E1 相比，E4 的参数效率更高（可适配参数为 130.8M，而 E1 为 731.1M）；与 T0 相比，E4 的数据效率也更高（有监督数据约为 20M，而 T0 为 320M），且两者达到的质量相近。

4.4. 最终方案

最后，我们结合前面各节探讨的各种技术，为我们的基础模型和目标域适配制定最终方案。具体包括：

1. F1：在 YT-U 上进行基于 BEST-RQ 的自监督训练（S0）；
2. F2：添加一个包含 6 个 LSTM 层的 RNN-T 解码器（S3），并使用筛选后的 YT-T（D1）和 VS-U 数据（G1）进行 JUST Hydra 训练；3. F3：在每个 Conformer 层之间插入残差适配器（E3），并使用目标域监督数据 VS-L 微调编码器适配器和解码器（E4）。

在 F3 上，采用 H1 或 H2 均未带来质量提升。采用这两种方案时，我们在 VS 上取得了与域内模型 T0 相近的 4.4% 词错误率，但所用的域内有监督数据少得多（21.6M 对比 321.6M），微调的参数也少得多（130.8M 对比 731.1M）。

表 5：最终方案中不同阶段的词错误率（%）。

ID	Data	VS
F1 S0，预训练	YT-U	-
F2 6L RNN-T 解码器，JUST	YT-T(D1)，VS-U	6.9
F3 编码器适配器 + 解码器	VS-L	4.4

5. 结论

本文研究语音基础模型（基础模型）的高效领域自适应问题。我们以 YouTube 为基础模型训练所用的源域，以语音搜索为目标域，针对如何提升基础模型在目标域上的质量，实证研究了不同技术。为提高数据效率，我们展示了仅使用语音数据和仅使用文本数据来提升基础模型在目标域上质量的潜力。然而，与使用域内配对语音文本数据训练的模型相比，仍存在较大差距。为提高参数效率，我们通过编码器适配器和解码器对基础模型进行适配，使其达到与全模型微调相同的质量。本文探讨的技术只是可用于语音基础模型高效领域自适应的众多方法中的一小部分。仍有许多问题有待解答，也有许多潜在方向值得探索。我们希望本研究能让更多人关注基于基础模型的语音解决方案这一范式。

6. 参考文献

[1] R. Bommasani, D. A. Hudson 等，“基础模型的机遇与风险”，arXiv:2108.07258，2021年。[2] T. Brown, B. Mann 等，“语言模型是少样本学习器”，NeurIPS论文集，第33卷，第1877–1901页，2020年。[3] D. Adiwardana, M.-T. Luong 等，“迈向类人开放域聊天机器人”，arXiv:2001.09977，2020年。[4] A. Narayanan, A. Misra, K.C. Sim 等，“通过大规模训练实现领域不变语音识别”，载于SLT论文集，2018年。[5] W. Chan 等，“Speechstew：将所有可用语音识别数据简单混合，训练一个大型神经网络”，arXiv:2104.02133，2021年。[6] G. Chen 等，“Gigaspeech：不断扩展的多领域自动语音识别语料库，包含10,000小时转录音频”，arXiv:2106.06909，2021年。[7] Y. Zhang, D. Park 等，“Bigssl：探索用于自动语音识别的大规模半监督学习前沿”，IEEE信号处理精选主题期刊，2022年。[8] Alec Radford, Jong Wook Kim 等，“通过大规模弱监督实现鲁棒语音识别”，技术报告，OpenAI，2022年。网址：https://cdn. openai. com/papers/whisper. pdf，2022年。[9] Dongseong Hwang, Khe Chai Sim 等，“伪标签优于人工标签”，arXiv:2203.12668，2022年。[10] Sanchit Gandhi, Patrick von Platen 等，“ESB：多领域端到端语音识别基准测试”，arXiv:2210 13352，2022年。[11] J. Devlin, M.-W. Chang 等，“BERT：用于语言理解的深度双向变换器预训练”，arXiv:1810.04805，2018年。[12] Alexei Baevski, Yuhao Zhou 等，“wav2vec 2.0：语音表征自监督学习框架”，NeurIPS论文集，第33卷，第12449–12460页，2020年。[13] C. Raffel, N. Shazeer 等，“探索使用统一文本到文本变换器进行迁移学习的极限”，《机器学习研究杂志》，第21卷，第140期，第1–67页，2020年。[14] Aaron van den Oord, Yazhe Li 等，“基于对比预测编码的表征学习”，arXiv:1807.03748，2018年。[15] Jan Chorowski, Ron J Weiss 等，“使用WaveNet自编码器进行无监督语音表征学习”，IEEE/ACM音频、语音与语言处理汇刊，第27卷，第12期，第2041–2053页，2019年。[16] Wen-Chin Huang, Shu-Wen Yang 等，“S3PRL-VC：基于自监督语音表征的开源语音转换框架”，载于ICASSP论文集，IEEE，2022年，第6552–6556页。[17] G.-T. Lin, C.-J. Hsu 等，“分析无监督语音识别的鲁棒性”，载于ICASSP论文集，IEEE，2022年。[18] A. Narayanan, A. Misra 等，“通过大规模训练实现领域不变语音识别”，载于SLT论文集，IEEE，2018年。[19] B. Li, R. Pang 等，“扩展端到端模型，实现大规模多语言自动语音识别”，载于ASRU论文集，IEEE，2021年，第1011–1018页。[20] B. Li, R. Pang 等，“大规模多语言自动语音识别：终身学习解决方案”，载于ICASSP论文集，IEEE，2022年，第6397–6401页。[21] Hank Liao, Erik McDermott 等，“利用半监督训练数据进行大规模深度神经网络声学建模，以实现YouTube视频转录”，载于ASRU论文集，IEEE，2013年，第368–373页。[22]谷歌，“YouTube数据统计”，https://blog.youtube/press/。[23] Neil Houlsby, Andrei Giurgiu 等，“自然语言处理的参数高效迁移学习”，载于ICML论文集，PMLR，2019年，第2790–2799页。[24] Dongseong Hwang, Ananya Misra 等，“使用自监督与半监督学习进行大规模自动语音识别领域自适应”，载于ICASSP论文集，IEEE，2022年，第6627–6631页。[25] Fadi Biadsy, Youzheng Chen 等，“一种使用子模型进行训练和推理的可扩展模型专门化框架及其在语音模型个性化中的应用”，arXiv:2203.12559，2022年。[26] Junxian He, Chunting Zhou 等，“迈向参数高效迁移学习的统一视角”，arXiv:2110.04366，2021年。

[27] C.-H. Yang、Y.-Y. Tsai 等人，《Voice2series：为时间序列分类重新编程声学模型》，载于ICML 会议论文集。PMLR，2021年。[28] Yu-An Chung、Yu Zhang 等人，《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 ASRU 会议论文集。IEEE，2021年，第244–250页。[29] C.-C. Chiu、J. Qin 等人，《采用随机投影量化器进行语音识别的自监督学习》，arXiv:2202.01855，2022年。[30] A. Graves、S. Fernandez 等人，《连接时序分类：使用循环神经网络标注未分段序列数据》，载于 ICML 会议论文集，2006年。[31] W. Chan、N. Jaitly、Q. V. Le 和 O. Vinyals，《Listen, attend and spell》，CoRR，第abs/1508.01211卷，2015年。[32] A. Graves，《使用循环神经网络进行序列转导》，CoRR，第abs/1211.3711卷，2012年。[33] A. Narayanan、R. Prabhavalkar 等人，《使用流式端到端模型识别长篇语音》，载于 ASRU 会议论文集。IEEE，2019年。[34] C. Park、R. Ahmad 等人，《利用对比损失比进行语音识别的无监督数据选择》，载于 ICASSP 会议论文集。IEEE，2022年。[35] Junwen Bai、Bo Li 等人，《多语言自动语音识别的无监督与有监督联合训练》，载于 ICASSP 会议论文集。IEEE，2022年，第6402–6406页。[36] Daniel S Park、Yu Zhang 等人，《改进的自动语音识别噪声学生训练》，arXiv:2005.09629，2020年。[37] Q. Xie、M.-T. Luong 等人，《采用噪声学生进行自训练可提升 ImageNet分类性能》，载于 CVPR 会议论文集，2020年，第10687–10698页。[38] Fadi Biadsy、Mohammadreza Ghodsi 等人，《有效构建融入非语言信号的太规模最大熵语言模型》，载于 Interspeech 会议论文集，第2710–2714页，2017年。[39] Shubham Toshniwal、Anjuli Kannan 等人，《编码器—解码器语音识别中语言模型集成技术的比较》，载于 SLT 会议论文集。IEEE，2018年，第369–375页。[40] Ankur Bapna、Yu-an Chung 等人，《SLAM：通过语音文本联合预训练实现语音与语言建模的统一编码器》，arXiv:2110.10329，2021年。[41] Z. Chen、Y. Zhang 等人，《MAESTRO：通过模态匹配获得匹配的语音文本表征》，arXiv:2204.03409，2022年。[42] Tara N Sainath、Rohit Prabhavalkar 等人，《JOIST：语音与文本联合流式模型》，arXiv:2210.07353，2022年。[43] Hendrik Purwins、Bo Li 等人，《音频信号处理中的深度学习》，IEEE信号处理精选主题期刊，第13卷第2期，第206–219页，2019年。[44] C.-C. Chiu、A. Narayanan 等人，《RNN-T模型无法泛化至域外音频：原因与解决方案》，载于 SLT 会议论文集。IEEE，2021年。[45]谷歌，《谷歌人工智能：我们的原则》，https://ai.google/principles/。[46] A. Gulati、J. Qin 等人，《Conformer：卷积增强型Transformer语音识别模型》，载于 Interspeech 会议论文集，2020年。[47] I. Williams、A. Kannan 等人，《在端到端神经网络系统中使用束搜索进行上下文语音识别》，载于 Interspeech 会议论文集，2018年。[48] D. S. Park、W. Chan 等人，《SpecAugment：一种用于自动语音识别的简单数据增强方法》，载于 Interspeech 会议论文集，2019年9月。[49] J. Shen、P. Nguyen 等人，《Lingvo：用于序列到序列建模的模块化、可扩展框架》，arXiv:2005.08100，2019年。[50]《云张量处理单元（TPU）》，https://cloud.google.com/tpu/docs，访问日期：2021年9月30日。[51] N. Shazeer 和 M. Stern，《Adafactor：具有次线性内存开销的自适应学习率》，载于 ICML 会议论文集。PMLR，2018年，第4596–4604页。[52] A. Vaswani、N. Shazeer 等人，《Attention Is All You Need》，CoRR，第abs/1706.03762卷，2017年。[53] Chung-Cheng Chiu、Wei Han 等人，《长篇语音识别端到端模型比较》，载于 ASRU 会议论文集。IEEE，2019年，第889–896页。[54] D. Le、M. Jain 等人，《结合基于Trie的深度偏置与浅层融合的上下文化流式端到端语音识别》，载于 Interspeech，布尔诺，2021年。