arXiv:2303.00550v1 [eess.AS] 1 Mar 2023

通过精英采样和集成知识蒸馏实现自动语音识别的领域泛化

Rehan Ahmad、*Md Asif Jalal*、*Muhammad Umar Farooq*、*Anna Ollerenshaw*、*Thomas Hain*

谢菲尔德大学语音与听觉组，英国

摘要

知识蒸馏（KD）已广泛应用于语音应用中的模型压缩和领域自适应。当存在多个教师模型时，通过对模型输出取平均值，可以轻松地将知识迁移到学生模型。然而，先前研究表明，学生模型难以通过这种组合实现良好的适应。本文提出在集成教师模型的输出端采用精英采样策略，从完全域外的教师模型生成的话语中，选出解码效果最佳的话语，以泛化至未见领域。教师模型在 AMI、LibriSpeech 和 WSJ 上训练，而学生模型则针对 Switchboard 数据进行适应。结果表明，采用基于各个模型后验概率的选择策略后，学生模型的词错误率优于所有教师模型和基线系统，绝对提升至少约 8.4%。此外，本文还通过相关性分析研究了使用域外数据进行模型自适应的相关见解。

索引词— 知识蒸馏，自动语音识别，师生训练

1. 引言

知识蒸馏（KD） [1]是一种广泛使用的技术，可将单个或多个复杂教师模型中的知识蒸馏到一个学生模型中。教师模型的知识会在输出 [1] 或中间特征层 [2]进行传递。一种常见的技术是利用教师的后验概率分布 [3]传递知识。通常，学生模型使用教师的后验概率（软标签）以及原始训练标签（硬标签）进行训练。因此，总损失是监督损失与知识蒸馏损失的加权和。知识蒸馏可用于许多任务，例如领域适应 [4, 5, 6], 、领域泛化 [7, 8, 9], 和模型压缩 [10, 11, 12]。

在自动语音识别中，知识蒸馏是在帧级 [13] 或序列级 [14] span 上进行的。根据 ASR 任务的序列特性，研究表明，序列级迁移是更好的方法 [13]。序列级知识蒸馏最早在 [15], 中提出，其中教师模型在整个样本空间上提供序列级概率分布，以实现更好的知识迁移。对于多个教师模型，优化采样策略以获得尽可能好的输出至关重要。例如， [16] 提出了三种针对教师模型输出的选择策略，即加权、Top-1 和 Top-K。这些策略都使用错误度量作为选择标准，因此假定有标签数据可用。然而，这种方法并不适用于旨在适配未见域外（OOD）数据的场景。因此，需要一种无监督采样策略，以减少多个教师模型输出的不确定性，并选择最佳输出。

此外，底层教师模型对于声学样本的泛化表征也很重要。在自动语音识别知识蒸馏任务中，研究者使用了 BERT[17], 、双向长短期记忆网络 [18] 等不同模型，并采用不同的上下文窗口。自监督训练为通过无监督预训练学习通用数据表征铺平了道路。这类模型通过掩码片段进行训练，以获得语音的通用上下文潜在表征 [19]。研究发现，这些自监督模型能够快速适应新领域或跨域任务 [20]。

知识蒸馏中的泛化问题有两个方面。第一个问题是优化教师模型学到的表征，第二个问题是为学生模型选择适当的分布。这些问题在教师与学生之间存在分布外数据时会变得棘手。本文通过使用预训练 wav2vec 教师模型来解决第一个问题，其中每个模型都在域内语料库上进行微调。第二个问题则通过基于后验概率的精英采样策略来解决，该策略会从教师模型解码的话语中选出最佳话语。总之，教师模型集成是在与学生模型需要适应的数据完全分布外的数据上训练的。随后，在未标注的分布外数据上运行推理，以生成教师模型的软标签。最后，采用基于输出后验概率的精英采样策略，从教师模型解码的结果中选出最佳话语来训练学生模型。具体而言，使用朗读语音数据集 WSJ（LDC 目录号 LDC93S6A、LDC94S13A）、LibriSpeech（LS） [21]以及会议语音数据集 AMI [22] 来训练三个最先进的教师模型。随后使用这些模型对 Switchboard（SB） [23] 语料库进行解码，并训练一个学生模型。结果表明，通过对教师模型集成的输出进行此类筛选，学生模型的表现优于基线系统和所有单个教师模型。

2. 使用CTC损失进行蒸馏

在师生训练中，知识通常以后验分布的形式从教师模型迁移而来。教师模型已经在给定的带标签数据集上完成训练。训练学生模型时，力求使其输出分布更接近教师模型的后验分布。通常通过使用 Kullback–Leibler（KL）散度 [24] 损失来实现：

$$KL(\hat{y}_{tea}||\hat{y}_{stu}) = \sum_{i=1}^{N} \hat{y}^{i}_{tea} \log \frac{\hat{y}^{i}_{tea}}{\hat{y}^{i}_{stu}} \quad (1)$$

其中$\hat{y}_{tea}$ 和$\hat{y}_{stu}$ 分别表示教师模型和学生模型的输出后验概率， N 为样本总数。通常通过计算帧级KL散度来匹配这两个分布。学生模型的总损失是监督损失与KL损失的加权和：

© 2023 IEEE。允许将本材料用于个人用途。对于所有其他用途，无论使用当前还是未来的媒体，均须获得 IEEE 的许可，包括为广告或促销目的重印/再版本材料、创作新的合集、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分。

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{sup} + (1-\alpha)\mathcal{L}_{KL} \quad (2)$$

研究表明，在自动语音识别等序列任务中，与帧级信息相比，通过序列级信息能够更有效地传递知识。因此，本研究采用了CTC [25]损失。用于知识蒸馏的教师模型和学生模型均基于使用libriVox（LV-60K）进行预训练的wav2vec2 [19] 。在wav2vec2的输出端连接稠密神经网络层，并应用CTC损失。模型的原始输入表示为序列 $\mathbf{x} = [x_1, ...x_{T_n}] \in X$，每个序列长度为 T_n， X 表示所有训练序列。wav2vec2模型输出的上下文特征表示为 $\mathbf{c} = [c_1, ...c_{T_n}]$。将上下文向量 $\mathbf{c}$ 输入全连接稠密层后，得到表示为 $\mathbf{h} = [h_1, ...h_{T_y}]$的输出，其中 T_y 为输出长度，词汇表大小为 $G = [g_1, ...g_z]$，其中 z 表示字素。教师模型使用原始标签，通过CTC损失进行训练：

$$\mathcal{L}_{CTC} = -\sum_{B} \log p(\mathbf{y}|\mathbf{h}) \quad (3)$$

其中，$\mathbf{y}$ 是输出标签序列，$B = \{X, Y\}$ 是训练数据集。

在标准的蒸馏设置中，学生模型使用原始（硬）标签和教师模型提供的软标签进行训练。学生模型的总损失是硬标签损失与软标签损失的加权和，如公式 2 所示。然而，在我们的场景中，我们假设目标域的原始标签不可用。OOD 教师模型对目标数据进行推理，并提供软标签来训练学生模型。因此，公式 2 中的 α 值为零。

对于蒸馏损失，我们遵循了 [12, 14] 的研究，该研究提出在自动语音识别任务中使用序列级损失，而不是帧级损失。因此，公式 2 中的 $\mathcal{L}_{KL}$ 被软 CTC 知识蒸馏损失取代。其表示如下：

$$\mathcal{L}_{CTC-KD} = -\sum_{B} p_{tea}(\hat{\mathbf{y}}|\mathbf{h_{tea}}) \log p_{stu}(\hat{\mathbf{y}}|\mathbf{h_{stu}}) \quad (4)$$

其中，$p_{tea}(\hat{\mathbf{y}}|\mathbf{h_{tea}})$ 和 $p_{stu}(\hat{\mathbf{y}}|\mathbf{h_{stu}})$ 分别是教师模型和学生模型的后验概率。在这种情况下，教师模型的后验概率来自多个教师模型的筛选标准，详见第 3 节。

3. 话语选择策略

本研究利用多个教师模型，将域外数据泛化到学生模型。众所周知，与单个模型相比，集成模型通常能提供更好的预测。根据 [10, 26] 的研究，从教师模型中蒸馏知识最直接的方法之一是对教师的逐帧后验概率取平均，具体如下：

$$\mathbf{p_{tea}} = \frac{1}{K}\sum_{k} \mathbf{p_k} \quad (5)$$

其中，K 是教师的总数，$\mathbf{p_k}$ 是教师 k 对整个话语的后验概率。由此得到的 $\mathbf{p_{tea}}$ 是每个话语逐帧后验概率的平均值。不过，这种策略可能效果不佳，因为如果某些教师表现欠佳，整体平均结果也会变差。因此，我们设计了一种采样技术，从教师模型中选出最佳解码话语。

在 [16], 中，作者提出了三种蒸馏策略，即加权、Top-1 和 Top-K。作者提出在每种策略中使用错误率作为选择／赋权指标。例如，在加权方法中，每位教师的权重根据小批次上的平均错误率确定，而只有在有有标签数据时才能计算该指标。此外，在他们的方法中，学生模型使用教师已经训练过的同一批数据进行学习。然而，我们的场景并非如此：学生模型使用教师未见过且没有原始标签的数据进行学习。因此，这种技术无法直接应用。

对于无标签数据，在转录同一话语时，每个教师模型可能会犯不同的错误，其中一个教师的表现会优于其他教师。因此，我们提出从最擅长转录该话语的教师模型中选取该话语。为此，可以对教师的解码后验概率进行排序。在我们的选择策略中，我们提出对每个教师给出的整个话语后验概率取平均，具体如下：

$$q_k = \frac{1}{T_y}\sum_{i=1}^{T_y} p_{ik} \quad (6)$$

其中 T_y 是话语的长度，$\mathbf{h}_k$ 是教师模型。对于每个由教师模型解码的话语，都会计算该平均后验概率。选择平均后验概率最大的教师模型作为最佳教师模型
as:

$$b = \arg\max_{k} q_k \quad (7)$$

采用这种选择策略时，每个话语都从表现最佳的教师模型 b 中选出，其后验概率将代入公式（4）以训练学生模型。该选择策略背后的直觉是精英主义的，即假设如果一组字素作为整体具有最高置信度，那么它们作为个体也最为匹配。当我们最大化字素组的置信度分数时，字素-音素对应关系将达到最优。如果每个单独标签的后验概率都更高，则表明模型具有更高的置信度。对所有帧取平均值，可以得到模型对整个话语的置信度。因此，三个模型中最佳的话语应具有最高的平均后验概率。

4. 内部表示相关性

从核心角度理解模型学习，对于建立输入声学数据与目标标签之间的关系至关重要。在自动语音识别中，转录通常取决于转录人员所使用语言的常规拼写以及语言感知，包括其方言或所属国家。统计相关性分析用于理解神经网络层中各表示之间的关系 [27]。在模型训练/测试过程中，提取各层之间或模型输出之间的激活输出，从而能够观察不同表示之间的共享表示。SVCCA [28] 使用奇异值分解（SVD）和典型相关（CC）分析来获取神经表示之间的共享关系。该算法试图寻找一组基 v，s 使两个矩阵（在此情况下为神经网络层的集合）投影到这些基上后，它们之间的相关性最大化：

$$\frac{v^T \sum_{XY} s}{\sqrt{v^T \sum_{XX} v}\sqrt{s^T \sum_{YY} s}} \quad (8)$$

这里，$\sum_{XX}$, $\sum_{XY}$, $\sum_{YY}$分别是矩阵的交叉协方差和协方差。采用逐层方法评估相关性的系数；对于 $\mathcal{N}$个数据点，汇总了 l_1和 l_2的投影视图，其中 $l_1 = \{z_1^{l_1}, ..., z_{N_1}^{l_1}\}$和 $l_2 = \{z_2^{l_2}, ..., z_{N_2}^{l_2}\}$。使用 SVD 剪除并保留代表性维度中排名前 99% 的维度，由此形成子空间 $l_1' \subset l_1$ 和 $l_2' \subset l_2$。使用典型相关分析找出向量 v 和 s，使 l_1' 和 l_2'的投影之间的相关性 ρ 最大。在这个

情形下， ρ 的值与神经表示中编码的信息相关：

$$\rho = \frac{\langle v^T l_1', s^T l_2' \rangle}{||v^T l_1'|| \, ||s^T l_2'||} \tag{9}$$

本文训练了两个 Transformer 模型（12 个编码器、12 个解码器层），分别使用原始 Switchboard 转录文本和由教师模型集成生成的伪转录文本。第 6.1 节使用 SVCCA 分析并讨论了各层的嵌入表示。

5. 实验

5.1. 数据

实验使用了四种不同的语料库，涵盖朗读语音和会话语音。AMI、LS 和 WSJ 用于训练教师模型，而 Switchboard（SB）用于训练学生模型。LS、WSJ 和 SB 的录音对象均为美国英语母语者；AMI 则是在包括爱丁堡、Idiap 和 TNO 在内的三家机构录制的，也涵盖非英语母语者。在这些语料库中，AMI 和 SB 属于会话语音语料库，而 WSJ 和 LS 属于朗读语音语料库。属于同一类别并不意味着它们相似。所有这些语料库在录音条件、说话风格等方面的分布各不相同。

LS、WSJ、AMI 和 SB 的总时长分别为 960 小时、272 小时、100 小时和 300 小时。由于 LS 的时长最长，我们仅考虑其被称为 LS360 的 360 小时子集。此外，除 SB 外，所有数据集的采样率均为 16KHz。因此，SB 从 8KHz 上采样至 16KHz。为分析模型性能，我们考虑了各语料库的标准测试集。LS-test-clean 和 LS-test-other 分别是 LS 的纯净和噪声测试集。我们考虑了 WSJ 的一个测试集，即 WSJ-Eval92-20K，其词汇表包含 2 万个词。从 AMI 中，我们选取了完整语料库自动语音识别测试集 AMI-FCASC。对于 SB，我们考虑了两个测试集：以 CH 表示的 eval00-CallHome（LDC97S42）和 eval00-Switchboard（SB）。eval00 结合了这两个 CH 和 SB 测试集。

5.2. 实验细节

教师模型和学生模型的基础模型是使用 LV-60K 预训练的 wav2vec2-large。wav2vec2 的输出连接到两个全连接层，并采用 CTC损失。训练期间，教师模型和学生模型的 wav2vec2 都会进行微调。此外，wav2vec2 要求音频以 16KHz 的采样率进行采样，因此只有 SB 音频会进行上采样。所有模型均使用基于字素的分词器。

实验中，首先分别使用 AMI、WSJ 和 LS360 数据集训练独立的教师模型。这些模型在多个测试集上进行评估，包括表1所列的域内和域外测试集。该评估有助于了解模型在完全域外分布上的表现，从而分析跨域适应。解码时使用的语言模型（LM）是一个三元语法模型，使用所有教师模型的训练转录文本进行训练。

使用教师模型对 SB 数据进行解码，并根据第 3 节讨论的选择策略，选取最佳的解码后的话语来训练学生模型。训练学生模型时，由于原始训练标签不可用，学生模型并不知道这些标签。因此，学生模型只能根据教师模型提供的后验概率进行泛化。此外，与声学模型类似，语言模型也是 SB 数据的域外模型。为保持本研究的一致性，选用该语言模型仅评估未见领域自适应。本文将所提出的方法与单个教师模型以及两个基线进行比较。对于第一个基线，计算所有教师模型的后验概率逐帧平均值，

表1：教师模型在有无语言模型条件下，针对域内和域外测试集的词错误率（%）评估。三个独立教师模型分别使用 AMI、LS360 和 WSJ 数据集训练。

	AMI		LS360		WSJ	
LM / 测试集	✗	✓	✗	✓	✗	✓
AMI-FCASC	15.8		51.69	47.42	60.60	57.32
LS-test-clean	14.61	12.41	3	3.04	10.14	8.39
LS-test-other	29.69	26.32	10.76		27.19	24.43
WSJ-Eval92-20K	15.87	13.11	9.60	7.14	2.39	1.47
eval00	52.04	49.21	45.86	42.59	66.43	64.20

表2：无语言模型时学生模型和教师模型在 Switchboard 测试集上的词错误率（%）评估。

	教师模型（不使用语言模型）			学生模型（不使用语言模型）		
				教师平均	教师最大值（我们的方法）	
sets	AMI	LS360	WSJ	SWBD		
eval00	52.04	45.86	66.43	58.11	51.71	**37.38**
CH	56.74	50.55	73.48	62.68	54.63	**41.15**
SB	47.09	40.96	59.07	53.33	48.67	**33.45**

并使用得到的后验概率训练学生模型。对于第二个基线，找出教师模型中最大的逐帧后验概率，并将其选作最终分布。随后使用得到的后验概率训练学生模型。后验概率的逐帧最大值表示模型对特定标签的置信度。因此，选取逐帧置信度较高的软标签。所有结果均在表2和表3中讨论，其中第一个基线以 Tea avg（教师平均法）表示，第二个则以 Fw max（逐帧最大值法）表示。

6. 结果与讨论

在各个语料库上分别训练的教师模型，最初在域内和域外测试集上进行了评估。表1展示了用于评估的各个测试集上的词错误率。每一列（AMI、LS360 和 WSJ）代表在相应语料库上训练的模型，每个测试集均分别在使用和不使用语言模型的情况下进行评估。可以看出，每个模型在其自身的测试集上都取得了最先进的结果，阴影单元格对此进行了标示。对于两个 LS测试集，跨语料库测试中词错误率最低的是 WSJ，其次是 AMI。这是因为 WSJ 同样是朗读语音数据，且语音来自英语母语者。然而，WSJ 和 LS 的录音及环境条件各不相同。因此，无论是否使用语言模型，WSJ 模型在 LS-test-clean 和 LS-test-other 上的词错误率都高出 2.5 倍以上。另一方面，AMI 是会话语音语料库，其中也有非英语母语者。此外，与 AMI 相比，WSJ 和 LS 的语音相对更干净，因为 AMI 中存在许多环境噪声和人为噪声。WSJ-Eval92-20K 测试集也呈现出类似的情况。对于 AMI测试集，LS360 和 WSJ 的结果都不如 AMI 模型本身。这同样是因为会话语音中存在不同的说话风格和简短的自发话语。最后，属于 SB 数据的 eval00测试集由 LS360 转写效果最佳。这是由于其训练数据量较大（360 小时），且说话人数多于 AMI 和 WSJ。有人可能会认为，AMI 应该表现更好，因为它包含会话语音。然而，AMI 模型仅使用 86 小时数据训练，且数据中既有英语母语者，也有非英语母语者；而 SB 中只有英语母语者。

表3：使用基于 AMI、LS360 和 WSJ 转录文本训练的三元语言模型，在 Switchboard 测试集上评估学生模型和教师模型的词错误率（%）。

	教师模型（使用域外语言模型）			学生模型（使用域外语言模型）		
				教师平均	教师最大值	我们的方法
sets	**AMI**	**LS360**	**WSJ**	**SWBD**		
eval00	49.21	42.59	64.20	46.53	38.09	**32.00**
CH	54.03	47.18	71.70	51.84	42.16	**35.72**
SB	44.17	37.80	56.37	40.99	33.84	**28.13**

表2展示了各个教师模型和学生模型在 SB 测试集上不使用语言模型时的词错误率。在学生模型中，教师平均并不是一种理想的集成模型策略，其词错误率远高于两个教师（AMI 和 LS360），也高于其他学生模型。主要原因是其中一位教师，即 WSJ，表现较差，从而拉低了教师后验概率的平均值，最终影响了学生模型的训练。而教师最大值优于大多数教师模型和教师平均，这是因为该策略会从教师中选取最大的后验概率，对应于置信度较高的预测。这种技术的缺点是，它会从不同教师模型中选取后验概率，从而影响输出的序列特性。

此外，除了所提出的方法外，在教师模型和学生模型中，所有测试集上词错误率最低的最佳模型是 LS360。然而，不能据此断定，LS360 模型在解码 SB 的所有话语方面都优于 AMI 和 WSJ 模型，因此适合用于训练学生模型。所提出的学生模型的评估结果清楚地表明，与所有教师模型和基线学生模型相比，该模型的词错误率最低。对于 CH 测试集，所提出的学生模型的词错误率为 41.15%，绝对差值分别比 LS360 低近 9.4%、比教师平均低 21.5%，并比教师最大值低 13.4%。同样，SB 相较于 LS360 模型至少提升了约 7.5%。总体而言，对于 eval00，所提出的学生模型比表现最佳的 LS360 模型好近 8.4%。

此外，表3显示，当教师模型和学生模型都使用域外语言模型时，结果更好。与表2中的结果相比，使用域外语言模型后，所提出的学生模型绝对提升了至多 5.3%，使 eval00 的词错误率降至 32.00%。此外还观察到，教师平均和教师最大值模型的结果均显著改善，最终教师最大值的表现超过了所有教师模型和教师平均。这表明，如果解码时使用语言模型，逐帧后验采样似乎是更好的选择。总体而言，结果显示了教师集成及选择策略对于训练学生模型的重要性。由于教师模型是在完全域外的数据上训练的，仍然可以有效地将知识从集成模型蒸馏出来，从而更好地适配学生模型。因此，学生模型的表现远优于基线系统和各个教师模型。

6.1 表征分析 图 1 展示了相关性

通过使用 Switchboard 数据训练的模型和使用 Switchboard 伪标签训练的模型来计算相关系数。收敛模式表明，在训练过程中，模型各层首先最大化声学输入与该层潜在表示之间的互信息，然后最小化该层表示与目标字素之间的互信息 [29, 30]。图 2 展示了使用 Switchboard 原始转录训练的模型与另一个模型各层相关系数的差异

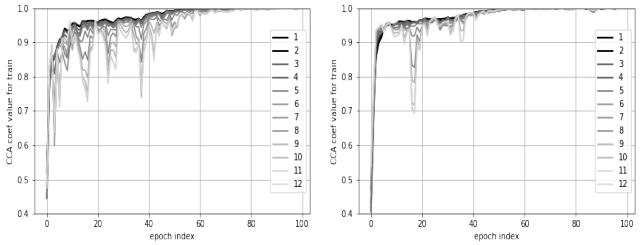

图 1：在使用原始 Switchboard 转录文本训练的 Transformer 第 1 至 12 层模型（左）和使用伪标签转录文本训练的 Transformer 第 1 至 12 层模型（右）中，随着性能逐渐收敛，SVCCA 相关系数随时间的变化

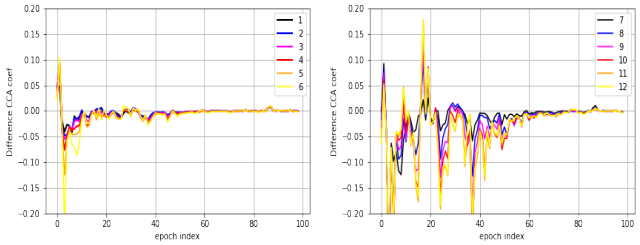

图 2：使用原始标签训练的 Switchboard 模型与使用伪标签训练的 Switchboard 模型之间，Transformer 第 1 至 6 层（左）和第 7 至 12 层（右）中逐层 SVCCA 相关性随时间的差值。

后者使用 Switchboard 伪标签转录进行训练。图 2 显示，与声学输入距离较近的各层之间，使用原始转录训练的 Switchboard 模型与使用伪标签转录训练的 Switchboard 模型几乎没有差异，甚至毫无差异。然而，编码器最后三层（10-12）呈现出显著差异，这也解释了标签之间的差异。可以认为，直到最后几层之前，两个模型学习到的表示都相似。因此，生成的伪标签与原始标签并非完全不同。图 1 和图 2 清楚地表明，尽管中间层学到的表示相似，但靠近目标类别分布（字素）的层较为敏感，稳定性也低于使用原始 Switchboard 转录训练的模型。这一现象归因于伪标签生成过程中的混淆，尤其是在学习声学边界时，词语前的犹豫和短促插话会造成影响。由于话语中存在犹豫、笑声及其他会话语音特有的细微差异，相同的话语会生成不同的伪标签，导致模型在训练时产生困惑。域内朗读语音集成自动语音识别模型并未涵盖所有这些会话语音细微特征。使用与未见数据域相似的语言模型生成伪标签，可以缓解这些问题。

7. 结论

为了实现对未见域外数据的泛化，本文提出结合集成知识蒸馏与精英采样技术。采样依据输出后验概率进行，并从教师模型中选出最佳话语来训练学生模型。结果表明，与教师模型和基线系统相比，采用该技术的学生模型能够更好地泛化到域外数据。研究仅使用域外语言模型来考察跨域声学模型的差异性。此外，还分析了不同模型和层中的中间神经表示，以了解声学数据与转录文本之间的关系。未来的工作中，

可以利用迭代伪标签和域内语言模型改进采样策略与伪标签校正。

8. 致谢

本研究在谢菲尔德大学的Voicebase/Liveperson语音与语言技术中心开展，该中心由Liveperson公司提供支持。

9. 参考文献

[1] G. Hinton、O. Vinyals和J. Dean，“蒸馏神经网络中的知识”，载于*NIPS*深度学习与表征学习研讨会，2015年。[2] A. Romero、N. Ballas、S. E. Kahou、A. Chassang、C. Gatta和Y. Bengio，“FitNets：薄型深度网络的提示”，*arXiv* 预印本 *arXiv:1412.6550*，2014年。[3] J. Gou、B. Yu、S. J Maybank和D. Tao，“知识蒸馏：综述”，国际计算机视觉期刊，第129卷，第6期，第1789–1819页，2021年。[4] T. Asami、R. Masumura、Y. Yamaguchi、H. Masataki和Y. Aono，“利用知识蒸馏实现深度神经网络声学模型的领域自适应”，载于*2017*年 *IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2017年，第5185–5189页。[5] Z. Meng、J. Li、Y. Gaur和Y. Gong，“通过师生学习实现端到端语音识别的领域自适应”，载于*2019*年 *IEEE*自动语音识别与理解研讨会（*ASRU*）。IEEE，2019年，第268–275页。[6] H. Zhu、J. Zhao、Y. Ren、L. Wang和P. Zhang，“利用类别相似性实现鲁棒语音识别的领域自适应”，*arXiv*预印本 *arXiv:2011.02782*，2020年。[7] Y. Wang、H. Li、L.-p. Chau和A. C Kot，“拥抱暗知识：利用正则化知识蒸馏实现领域泛化”，载于第*29*届*ACM*国际多媒体会议论文集，2021年，第2595–2604页。[8] B. Kim、S. Yang、J. Kim和S. Chang，“利用残差归一化实现高效声学场景分类的领域泛化”，*arXiv*预印本 *arXiv:2111.06531*，2021年。[9] G. Fang、Y. Bao、J. Song、X. Wang、D. Xie、C. Shen和M. Song，“拼接蒸馏：从域外数据中进行知识蒸馏”，神经信息处理系统进展，第34卷，第11920–11932页，2021年。[10] Y. Chebotar和A. Waters，“从神经网络集成中蒸馏知识以用于语音识别”，载于*Interspeech*，2016年，第3439–3443页。[11] H.-G. Kim、H. Na、H. Lee、J. Lee、T. G. Kang、M.-J. Lee和Y. S. Choi，“利用输出误差对自注意力端到端模型进行知识蒸馏”，载于*ICASSP 2019—2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第6181–6185页。

[12] R. Takashima, S. Li, and H. Kawai, 《CTC-T模型知识蒸馏方法研究》，载于*2018*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第5809–5813页。

[13] J. H. Wong and M. Gales, 《深度神经网络的序列学生-教师训练》，2016年。

[14] M. Huang, Y. You, Z. Chen, Y. Qian, and K. Yu, 《序列模型的知识蒸馏》，载于 *Interspeech*，2018年，第 3703–3707 页。[15] Y. Kim and A. M Rush, 《序列级知识蒸馏》，*arXiv* 预印本 *arXiv:1606.07947*，2016 年。[16] Y. Gao, T. Parcollet, and N. D Lane, 《为联合CTC-注意力端到端语音识别从声学模型集成中蒸馏知识》，载于 *2021*年*IEEE*自动语音识别与理解研讨会 *(ASRU)*。IEEE，2021年，第 138–145 页。[17] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, 《蒸馏BERT的知识以用于序列到序列自动语音识别》，*arXiv* 预印本 *arXiv:2008.03822*，2020年。[18] G. Kurata and K. Audhkhasi, 《改进从双向到单向LSTM CTC的知识蒸馏，以用于端到端语音识别》，载于 *2018*年*IEEE*口语技术研讨会 *(SLT)*，2018年，第 411–417 页。[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, 《wav2vec 2.0：用于语音表征自监督学习的框架》，神经信息处理系统进展，第 33 卷，第 12449–12460 页，2020年。[20] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, 《鲁棒wav2vec 2.0：分析自监督预训练中的领域偏移》，*arXiv* 预印本 *arXiv:2104 01027*，2021年。[21] 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》，载于 *IEEE*声学、语音与信号处理国际会议，2015年，第 5206–5210 页。[22] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, 《AMI会议语料库：预发布》，载于多模态交互的机器学习。2006年，第 28–39 页，施普林格·柏林·海德堡。[23] J.J. Godfrey, E.C. Holliman, and J. McDaniel, 《Switchboard：用于研究与开发的电话语音语料库》，载于 *IEEE*声学、语音与信号处理国际会议，1992年，第 1 卷，第 517–520 页，第 1 卷。[24] J. M. Joyce，库尔贝克-莱布勒散度，第 720–722 页，施普林格·柏林·海德堡，柏林，海德堡，2011年。[25] A. Graves, 《连接主义时序分类》，载于使用循环神经网络进行监督式序列标注，第 61–93 页。施普林格，2012年。[26] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, 《从教师集成中高效蒸馏知识》，载于*Interspeech*，2017年，第 3697–3701 页。[27] A. Ollerenshaw, M. A. Jalal, and T. Hain, 《端到端语音识别神经表征的见解》，*2021*年*Interspeech*会议论文集，第4079–4083 页，2021年。[28] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, 《SVCCA：用于深度学习动态与可解释性的奇异向量典型相关分析》，神经信息处理系统进展，第 30 卷，2017年。[29] R. Shwartz-Ziv and N. Tishby, 《通过信息揭开深度神经网络的黑箱》，*arXiv* 预印本 *arXiv:1703.00810*，2017年。[30]N. Saphra and A. Lopez, 《使用SVCCA理解语言模型的学习动态》，*arXiv* 预印本 *arXiv:1811.00225*，2018年。