基于知识蒸馏的小样本外语听力理解模型构建研究

壮 蓉[1] 张文娴[1]
ZHUANG Rong ZHANG Wenxian

摘 要 在小样本条件下构建有效的外语听力理解模型是智能语言教学中的重要挑战。针对传统深度模型对数据量依赖较强的问题，文章提出一种基于知识蒸馏的 Teacher-Student 训练框架，引导轻量学生模型在少量数据上有效学习。实验基于 LibriSpeech 听力数据子集构建小样本实验集，结果显示，蒸馏模型在保持低复杂度的同时获得了显著性能提升。研究结果对职业教育中信息化、智能化教学辅助系统的构建具有一定参考价值。

关键词 小样本学习；知识蒸馏；外语听力；BiLSTM；深度学习

doi：10.3969/j.issn.1672-9528.2025.07.036

0 引言

在外语教学中，听力理解被视为语言运用能力的重要体现，其涉及语音识别、语义解析与情境推理等多个层次，是智能语言学习系统的重点攻关方向。当前，在高校和职业院校的信息化教学改革中，借助深度学习实现对听力材料的理解与答题辅助成为主流趋势。然而，传统深度神经网络模型普遍依赖大量高质量标注语料，导致其在数据受限的现实场景中推广困难，模型容易出现过拟合、泛化能力弱等问题。特别是在职业教育领域，不同院校资源配置不均、学生语言基础差异较大，小样本学习需求日益突出。

近年来，小样本学习（few-shot learning）技术迅速发展，其核心目标是在训练数据量有限的条件下实现模型的有效泛化[1]。与此同时，知识蒸馏（knowledge distillation）作为一种将复杂模型知识迁移至轻量模型的技术路径，广泛应用于模型压缩、移动端部署等场景[2]。通过引入教师模型（Teacher）输出的软标签作为辅助信号，引导学生模型（Student）学习更丰富的语义表达，已在图像识别、语音识别等任务中验证其有效性[3]。然而，现有研究多数集中于分类或识别类任务，在结合知识蒸馏实现小样本外语听力理解方面，尚缺乏针对性的探索与系统性评估。

为此，本文提出一种基于知识蒸馏的小样本外语听力理解模型构建方法。通过构建 Teacher-Student 模型结构，利用大型预训练语音模型提取的深层语义表示，引导轻量学生模型在小规模语料上获得高性能。本文选用 LibriSpeech 公开语音数据构建小样本听力理解实验任务，采用 BiLSTM 结构作为学生网络，在蒸馏温度与损失权重等参数上进行系统调优，最终实现了在资源有限条件下的准确理解与稳定预测。研究结果有望为教育领域的智能听力评估系统提供轻量、实用的技术路径。

1 相关技术基础

1.1 小样本学习与深度语音模型

小样本学习（few-shot learning, FSL）旨在通过最少的训练样本完成任务学习，尤其适用于标注数据稀缺的场景，如医学图像识别、低资源语种建模和教学场景中的自动化评估。其核心策略包括引入先验信息、迁移已有模型知识、构建元学习机制等，以降低模型对大规模标注语料的依赖。

在语音理解任务中，语音信号是一类高维、时序性强且非结构化的数据。为了提取其中的语义信息，通常需将音频信号转化为 Mel 频谱图或 MFCC（mel frequency cepstral coefficient）等特征表示，再输入深度网络进行建模。目前广泛应用的深度语音模型主要包括：循环神经网络（RNN）及其变种，如双向长短时记忆网络（BiLSTM），通过引入时间门控结构捕捉音频信号中的上下文依赖；卷积神经网络（CNN），适合提取局部短时语音模式，可作为前端特征提取器；Transformer 结构，基于多头注意力机制，能够捕捉语音信号中远距离的依赖关系，已成为主流的端到端语音建模架构。然而，上述深度模型普遍参数量大、结构复杂，对训练样本数量和计算资源高度依赖。尤其在外语听力理解等教育场景中，教学单位无法长期维护海量数据采集与标注，导致模型难以泛化到个体学习者。因此，构建轻量化、可迁移且具备稳定性能的小样本语音理解模型成为研究关键。

1. 山东电子职业技术学院 山东济南 250200

[基金项目] 山东省职业技术教育学会 2025 年度职业教育科研课题“数智 + 教育新生态：职业院校通识课程教学智能评价模型构建研究”（KYKT2025G162）

为解决该问题，一种典型策略是通过“预训练+迁移”模式先在大规模语音数据集（如 LibriSpeech、CommonVoice）上训练强大的教师模型（Teacher），再在小样本任务中指导学生模型（Student）学习，从而实现“以大带小”的知识迁移。

1.2 知识蒸馏机制

知识蒸馏（knowledge distillation）是一种在模型压缩与迁移学习中广泛应用的技术，最早由 Hinton 等人提出。其基本思想是在训练小模型（Student）时，引入由大模型（Teacher）生成的“软标签”作为学习目标，帮助学生模型获得更丰富的类别关系信息和更强的泛化能力。在传统的深度学习中，模型通过最小化交叉熵损失函数学习真实标签，但该方式仅利用了“硬标签”的信息。而教师模型所生成的预测概率分布（softmax 输出）则包含了类别之间的相对相似性。这些“软标签”作为训练信号可以帮助学生模型更充分地理解类别结构和边界 [4]。

知识蒸馏根据蒸馏目标的不同，主要分为以下几类：输出层蒸馏（logits distillation）：使用教师模型的 softmax 概率指导学生模型，最常见也最有效；中间层特征蒸馏（feature distillation）：对齐教师与学生在中间层的激活输出；注意力蒸馏（attention distillation）：基于注意力权重进行知识传递；自蒸馏（self distillation）：同一模型在不同阶段或不同子结构间传递知识。在本研究中，蒸馏技术被用于将教师模型中训练良好的语义结构迁移到轻量的学生模型中，从而解决外语听力理解任务中的小样本问题。具体的蒸馏策略与损失函数构造将在第 2.2 节中详细展开。

2 模型结构与蒸馏训练方法

为在小样本条件下实现对外语听力材料的有效理解，本文构建了一个基于知识蒸馏的轻量模型训练框架，整体包括教师模型（Teacher）与学生模型（Student）两个模块。该架构在保持模型性能的同时，大幅降低了计算资源需求，提高在教学终端的适应性和可部署性。

2.1 教师 - 学生模型结构设计

教师模型采用 Facebook AI 提出的 wav2vec 2.0 base 框架 [5]，其通过自监督方式在大规模语音数据集上预训练获得强大的上下文建模能力。随后，使用 LibriSpeech 960h 数据集对其进行监督微调，任务设定为句子级听力理解分类。为避免过拟合，仅解冻其顶部 Transformer 编码层与分类模块，其余结构保持固定。微调后的模型具备较强的语义判别能力，可生成高质量的 logits 与中间层语义特征，作为蒸馏源信号。

学生模型采用双层双向长短时记忆网络（BiLSTM），作为低计算成本的轻量级候选。BiLSTM 能够有效捕捉语音序列中的前后依赖关系 [6]，最终通过全连接层与 Softmax 输出理解类别的概率分布。设输入为经过预处理的 Mel 频谱特征序列：

$$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}\ ,\ \boldsymbol{x}_t\in\mathbf{R}^d \tag{1}$$

模型通过双向 LSTM 映射为隐藏状态表示：

$$\boldsymbol{H}=\mathrm{BiLSTM}(\boldsymbol{X})\in\mathbf{R}^{T\times h} \tag{2}$$

并在全局平均池化后生成语义向量 $\boldsymbol{h}_{\mathrm{avg}}$，最终经分类器输出预测概率：

$$\hat{\boldsymbol{y}}=\mathrm{Softmax}(\boldsymbol{W}\cdot\boldsymbol{h}_{\mathrm{avg}}+\boldsymbol{b}) \tag{3}$$

式中：$\hat{\boldsymbol{y}}\in\mathbf{R}^C$，$C$ 为分类类别数。

输入特征处理方面，原始音频采用 16 kHz 采样率，通过 40 维 Mel 滤波器组提取特征，构建二维时频图，有助于保留语音重要信息的同时控制维度，适合小样本建模场景。

2.2 蒸馏策略与训练流程

学生模型的训练不仅依赖人工标注标签，还需尽可能复现教师模型的预测分布与中间语义结构，从而提升模型稳定性与泛化能力。本文蒸馏训练包括三项关键机制：

（1）软标签蒸馏（Soft Target Distillation）

使用带温度调节的 Softmax 平滑教师与学生模型输出 logits，形成 soft probability 分布：

$$p_i^{(T)}=\frac{\exp(z_i^{(T)}/T)}{\sum_j\exp(z_j^{(T)}/T)},\quad p_i^{(S)}=\frac{\exp(z_i^{(S)}/T)}{\sum_j\exp(z_j^{(S)}/T)} \tag{4}$$

式中：T=2，用于扩大类别间边界信息。

使用 KL 散度度量两分布间差异：

$$L_{\mathrm{KD}}=T^2\sum_{i=1}^{C}p_i^{(T)}\log\left(\frac{p_i^{(T)}}{p_i^{(S)}}\right) \tag{5}$$

（2）硬标签监督（Hard Label Supervision）

结合人工标签，采用传统交叉熵损失：

$$L_{\mathrm{CE}}=-\sum_{i=1}^{C}y_i\log(\hat{y}_i) \tag{6}$$

两者加权组成主蒸馏目标：

$$L_{\mathrm{total}}=\alpha\cdot L_{\mathrm{CE}}+(1-\alpha)\cdot L_{\mathrm{KD}} \tag{7}$$

（3）中间层特征对齐（Feature Alignment）

为提升结构信息的迁移效果，进一步引入中间层蒸馏损失，将教师模型特征激活值 $\boldsymbol{h}^{(T)}\in\mathbf{R}^d$ ，与学生模型中间表示 $\boldsymbol{h}^{(S)}\in\mathbf{R}^d$ 对齐，采用均方误差形式定义：

$$L_{\mathrm{feat}}=\frac{1}{N}\sum_{j=1}^{N}\left\|F_T^{(j)}-F_S^{(j)}\right\|_2^2 \tag{8}$$

最终总损失函数整合三项子目标：

$$L_{\mathrm{final}}=\lambda_1\cdot L_{\mathrm{CE}}+\lambda_2\cdot L_{\mathrm{KD}}+\lambda_3\cdot L_{\mathrm{feat}} \tag{9}$$

式中：λ_1=0.6、λ_2=0.3、λ_3=0.1，控制各项损失的贡献比例。

通过上述结构设计与蒸馏流程，学生模型在小样本条件

下能够充分吸收教师模型的语义表示与结构表达能力，在显著压缩模型规模的同时保持优异的听力理解性能。

3 实验设置与结果分析

为验证所提出的基于知识蒸馏的小样本外语听力理解模型的有效性，本文设计了一组对比实验，围绕模型在小规模语料条件下的表现，分别从分类性能、模型复杂度和训练开销等多个维度开展评估分析。实验目标不仅是验证蒸馏机制的有效性，同时探讨模型在实际教学环境中部署的可行性与稳定性。

3.1 数据集与评估指标

本实验基于 LibriSpeech-clean-100 语音语料构建小样本测试集。具体地从语料中抽取总时长约 30 min 的英文语音段落，围绕其内容人工设计了 10 道听力理解题目，涵盖主旨提取、细节识别、推理判断等多个认知维度。每道题目与一段音频材料配对，并设置 3 个选项，构成“音频 – 问题 – 选项”的三元组结构，确保测试任务具有语义多样性与语言真实性。所有音频数据均统一采用 16 kHz 采样率，经标准化预处理后转化为 40 维 Mel 频谱图 [7]，作为模型输入特征。

在性能评估方面，本文采用准确率（Accuracy）与 F_1 值（F_1-score）作为核心指标。其中，准确率反映模型整体分类能力，而 F_1 值综合考虑精确率与召回率，更能客观评价在样本不均衡情况下的识别稳健性。此外，为全面刻画模型效率，实验还报告了各模型的参数规模（以百万参数 10^6 为单位）与平均训练时长，用于分析不同模型在资源占用与实际部署方面的表现。

3.2 对比方案设计

为系统评估知识蒸馏机制对小样本任务性能的提升效果，本文设置了 3 种对比模型作为实验对象，具体如下：

（1）Student（无蒸馏）

该模型为轻量 BiLSTM 结构，仅依靠真实标签（hard label）进行训练，不引入任何教师模型信息。该模型反映了在缺乏迁移机制下，纯监督小样本学习的性能上限。

（2）Student + KD（蒸馏）

在 Student 基础上引入知识蒸馏机制，训练过程中同时融合教师模型输出的 soft label 与真实标签进行联合优化，目标在于增强学生模型的表达能力与泛化能力。

（3）Teacher（教师模型）

即为微调后的 wav2vec 2.0 base 模型，具备完整参数与完整语料支持，用作性能上限对照组。虽然该模型不可直接部署于教学终端，但能作为学生模型性能参考基线。

为保证对比公平，所有模型使用一致的输入特征表示方式与训练参数配置。训练优化器统一采用 Adam，初始学习率设为 1×10^{-4}，训练轮数固定为 50 epoch，batch size 设为 8。所有实验在同一平台（Intel i9 CPU + RTX 3090 GPU）下完成，确保硬件环境一致。

3.3 实验结果与分析

实验结果如表 1 所示。在基线条件下，Student 模型在测试集上取得 68.4% 的准确率与 65.3% 的 F_1 值，表现出轻量模型在小样本条件下的性能瓶颈。而引入知识蒸馏机制后，Student + KD 模型的准确率显著提升至 78.9%，F_1 值达到 77.1%，相较无蒸馏方案分别提升了 10.5 与 11.8 个百分点，充分验证了知识蒸馏在引导小模型学习语义信息、缓解过拟合方面的有效性。

表 1 各模型在测试集上的性能对比结果

模型名称	Accuracy/%	F_1-score/%	参数量 /10^6
Teacher（教师模型）	85.6	84.1	94
Student（无蒸馏）	68.4	65.3	9.2
Student + KD（蒸馏）	78.9	77.1	9.2

同时，Student 模型的参数量维持在 9.2×10^6，相较于 Teacher 模型的 94×10^6 缩小超过 90%，大幅降低了存储与计算成本。在训练效率方面，蒸馏后的学生模型平均训练时间为 31 min，约为教师模型的 55%，展示出良好的收敛速度与部署可行性。

总体而言，实验结果充分验证了本文所构建的基于知识蒸馏的小样本外语听力理解模型在精度、效率和可部署性方面的优势。该方法不仅可在资源受限场景下实现性能突破，也为后续在实际教学系统中的落地应用提供了技术支撑。

4 结论与展望

针对外语教学场景中普遍存在的听力标注数据稀缺问题，本文提出了一种基于知识蒸馏的 Teacher-Student 框架，用于构建小样本条件下的外语听力理解模型。该方法充分利用预训练大模型所蕴含的语义表示能力，通过软标签与特征对齐机制引导轻量化学生模型在小样本语料上实现更高的分类准确率和泛化能力。实验结果表明，引入知识蒸馏策略后，学生模型的准确率和 F_1 值均获得了显著提升，且在模型参数量和训练资源开销方面远低于大型教师模型，兼顾了性能与效率，为小样本智能教学任务提供了可行的建模路径。

在未来研究中，本文将从以下几个方向进一步拓展和深化：首先，在当前音频建模的基础上，考虑引入语音与文本的多模态蒸馏策略，实现对语义、语调与语言结构的协同理解；其次，探索结合学习者特征的个性化学生建模机制，根据不同语言水平、认知风格构建适应性理解模型，提升模型在教育场景下的实用性和适配性；最后，计划将本模型集成进智能外语教学平台中，结合教学任务实现实时交互式验证与人机共学，推动小样本人工智能技术在外语教育中的应用落地。

基于 CosyVoice2 的语音克隆技术的研究

冯 豹[1] 王 芬[1] 余海军[2]
FENG Bao WANG Fen YU Haijun

摘 要 DeepSeek 的横空出世几乎占领了全球 AI 用户的榜首，这也让我们领悟了 AI 的魅力，而语音克隆是 AI 工具功能的重要一环。阿里巴巴的 CosyVoice2 作为语音克隆领域的代表之一，其中部分功能仍有可拓展性，如多帧局部信息之间的细粒度并不饱和以及方言的精确感知。因此，文章基于 CosyVoice2 的语音克隆技术的研究，提出语义特征识别模型 SFRM 以提高语音克隆技术的精确性。包括设置音频项矩阵初筛、利用语义特征训练学习、利用麦克斯相似度来语气改进等。其使输入新文本生成的音频的情感、自然度有了进一步提高。

关键词 语音克隆；CosyVoice2；SFRM；音频项矩阵；语义特征

doi：10.3969/j.issn.1672-9528.2025.07.037

0 引言

随着 AI 工具的兴起，语音克隆技术也在生活中被广泛应用，文献 [1] 中提出语音识别模型——SpeechTransformer，该模型虽提高了一定效率，但忽略了汉字多帧信号之间的局部特征信息，如方言无法识别。文献 [2] 中则提出了语音克隆模型解码方法，但在局部和全局特征上仍然存在问题，尤其情感语气上。本文基于 CosyVoice2 的语音克隆技术进行研究，在模拟训练、克隆模型、Conformer 编码器卷积模块上进行优化。通过对 CosyVoice2 方法的改进，克隆的新文本音频的语气情感，特别是方言精确度上有了进一步提高。

1 语音克隆模型

基于 CosyVoice2 的语音克隆技术是一种监督离散语音标记技术。通过对语义特征数据进行深度学习处理，实现高质量的识别。其在音频的训练、识别认证等方面智能识别能力较强。模型通过大规模语料库进行预训练，能够生成连贯、合理的文本序列。而本文研究的是根据其提出语义特征识别模型 SFRM（semantic feature recognition model），SFRM 核

1. 江苏开放大学 江苏南京 210036
2. 中车大同电力机车有限公司 山西大同 037038

参考文献：
[1] 赵凯琳 , 靳小龙 , 王元卓 . 小样本学习研究综述 [J]. 软件学报 , 2021,32(2):349-369.
[2] 黄震华 , 杨顺志 , 林威 , 等 . 知识蒸馏研究综述 [J]. 计算机学报 ,2022,45(3):624-653.
[3] 刘兵 , 史伟峰 , 刘明明 , 等 . 融合知识蒸馏与记忆机制的无监督工业缺陷检测 [J]. 中国图象图形学报 ,2025,30(3):660-671.
[4] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL].(2024-12-25)[2025-03-09]. https://doi.org/10.48550/arXiv.1503.02531.
[5] BAEVSKI A, ZHOU H, MOHAMED A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations[EB/OL].(2020-10-22)[2025-06-06].https://doi.org/10.48550/arXiv.2006.11477.
[6] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural networks, 2005,18(5/6): 602-610.
[7] MÜLLER M. Fundamentals of music processing: audio, analysis, algorithms, applications[M/OL]. Springer, 2015[2024-02-01]. https://link.springer.com/book/10.1007/978-3-319-21945-5.

【作者简介】
壮蓉（1972—），女，江苏常州人，硕士，副教授，研究方向：外语教学。
张文娴（1983—），女，山东德州人，硕士，讲师，研究方向：外语教学。

（收稿日期：2025-06-11 修回日期：2025-07-08）