TICL+：面向儿童语音识别的语音上下文学习案例研究

Haolong Zheng、*Yekaterina Yegorova*、*Mark Hasegawa-Johnson*

伊利诺伊大学厄巴纳-香槟分校{haolong2、
yay2、jhasegaw}@illinois.edu

arXiv:2512.18263v1 [eess.AS] 20 Dec 2025

摘要

由于声学和语言变异性显著、标注数据有限，且儿童语音与成人语音存在较大差异，儿童语音识别仍面临挑战。语音基础模型可通过语音上下文学习（SICL）应对这些挑战，无需微调即可适应新领域。然而，SICL 的效果取决于如何选择上下文示例。我们扩展了一种现有的基于检索的方法——用于 SICL 的文本嵌入 K 近邻（TICL），并加入声学重排序步骤，构建 TICL+。这一扩展优先选择与测试输入在语义和声学上均相匹配的示例。在四个儿童语音语料库上的实验表明，TICL+ 相较于零样本性能，词错误率最高相对降低 53.3%；相较于基线 TICL，最高降低 37.6%，凸显了结合语义和声学信息对于实现稳健、可扩展的儿童语音自动语音识别的重要价值。

索引词— 上下文学习、自动语音识别、大型多模态模型

1. 引言

根据《残障人士教育法》（IDEA），需要言语与语言服务的儿童超过一半，约有340万名。存在言语或语言相关问题的儿童，其学业和社会情感发展可能会落后 [1]。通常，越早发现这些问题并通过基于能力的干预加以解决，这些儿童在学业和社会经济方面取得良好发展的可能性就越大。然而，言语与语言病理学家（SLPs）的数量与需要其服务的儿童数量之间存在严重失衡，因此人们越来越关注将这些任务自动化，以提高语言障碍筛查的效率。 [2]。

此类自动化的成功在很大程度上取决于集成到这些流程中的自动语音识别（ASR）系统的准确性和稳健性。ASR 儿童语音仍是一项低资源任务。由于儿童语音具有显著的声学和语言变异性，直接使用现成的自动语音识别系统时，性能会明显落后。这种变异性包括因发育速度不同而产生的说话人间变异，以及因发音能力尚未充分发展而产生的说话人内变异 [3–8]。由此造成的性能下降十分显著，因为用于训练大规模自动语音识别模型的数据中基本不存在这些变异。为应对这一挑战，研究人员采用迁移学习技术，将成人自动语音识别系统中的知识应用于儿童语音识别，具体做法是使用儿童语音数据对 Whisper [9, 10] 和 Wav2Vec2 [11] 等模型进行微调。为减轻使用不同于预训练数据的领域数据微调自监督学习模型时产生的数据偏差， [12]提出了一种领域负责任的适配与微调策略，并报告称，使用儿童语音数据集进行微调后，多个语音模型的词错误率（WER）均有所改善 [12, 13]。

除微调方法外，上下文学习（ICL）[14, 15] 已成为大语言模型（LLMs）的一种灵活适应范式，既能缓解灾难性遗忘，也无需更新参数。SICL 研究依赖随机采样来选择上下文示例 [16–18], 然而，先前研究表明，上下文示例的选择会显著影响 ICL 的性能 [19–21]。为使上下文示例的选择更具针对性，我们此前提出了用于 SICL 的文本嵌入 K 近邻（TICL）流程：先为测试样本生成伪标签，再检索语义相似的示例，从而提升 SICL 的性能 [22]。该方法依赖伪标签，而在儿童语音等低资源场景中，伪标签的质量可能大幅下降。在这些场景下，将声学相似度作为额外的相似度度量可能会有所助益。为此，我们扩展了 TICL 的检索阶段，引入基于声学的重排序步骤，优先选择声学特征更接近测试话语的示例，由此提出了 **TICL+** 流程。如图 1 所示，这种双重标准选择策略能够改善 SICL 在儿童语音等低资源语音领域中的上下文构建。

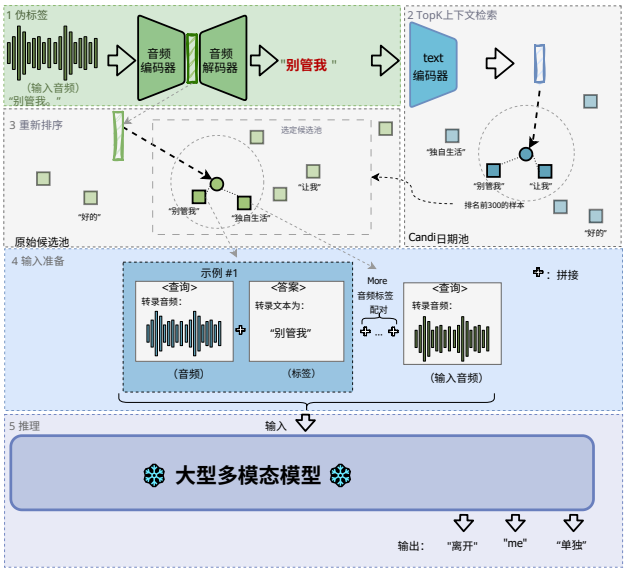

图 1。TICL+ 流程概览

2. 方法

2.1. 语音上下文学习

ICL 并非通过更新模型参数来适应目标领域，而是以从目标领域中选取的示例作为条件，使模型适应该领域。SICL 将基于文本的 ICL 扩展为同时以配对的音频和文本词元为条件。给定一个测试语音样本 $\boldsymbol{s}^*$，模型 Λ 在上下文 C的条件下生成转录文本$\hat{\boldsymbol{y}}$：

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \Pr(\boldsymbol{y} \mid C, \boldsymbol{x}_{\boldsymbol{s}}^*, \Lambda),$$

其中，$\boldsymbol{x}_{\boldsymbol{s}}^*$ 表示 $\boldsymbol{s}^*$的音频编码。

上下文 C 由查询—答案对 $c^{(i)}$ =($q^{(i)}$, $a^{(i)}$)组成，其中每个查询 $q^{(i)}$ 是编码后的音频片段，每个答案 $a^{(i)}$ 则对应于该音频片段的转录文本。

2.2. 基于文本嵌入的 K 近邻候选选择

TICL 流程 [22] 引入了一种基于文本嵌入的 K 近邻候选选择方法，旨在为给定的测试样本识别有效的 SICL 上下文 C。为构建 C，TICL 从候选数据集 $\mathcal{C} = \{(\boldsymbol{s}^{(i)}, \boldsymbol{y}^{(i)})$ 中检索转写文本与测试话语在词汇上相似的语音—转写文本配对$\}_{i=1}^{N}$，其中 $\boldsymbol{s}^{(i)}$ 表示语音音频，$\boldsymbol{y}^{(i)}$ 表示其对应的转写文本。

冻结的文本编码器 ϕ： $\mathcal{Y}_{\text{text}} \to \mathbb{R}^d$ 将每个转录文本 $\boldsymbol{y}$ 映射为 d维句子嵌入。ℓ_2归一化后的嵌入定义为：

$$\bar{\phi}(\boldsymbol{y}) = \frac{\phi(\boldsymbol{y})}{\|\phi(\boldsymbol{y})\|_2}$$

对于每个候选项 $c_i \in \mathcal{C}$，其归一化嵌入预先计算为 $\bar{\boldsymbol{z}}^{(i)} = \bar{\phi}(\boldsymbol{y}^{(i)})$。在推理过程中，测试话语 $\boldsymbol{s}^*$ 的真实转录文本不可用。相反，使用冻结的自动语音识别模型 f_θ 生成伪转录文本 $\tilde{\boldsymbol{y}} = f_\theta(\boldsymbol{s}^*)$： $\mathcal{X}_{\text{音频}} \to \mathcal{Y}_{\text{text}}$，其中 $\mathcal{X}_{\text{audio}}$，$\mathcal{Y}_{\text{text}}$分别表示音频空间和文本空间。

随后，将伪标签编码为归一化的词汇嵌入 $\bar{\boldsymbol{z}}^* = \phi(\tilde{\boldsymbol{y}})$。为选择相关的上下文示例，我们计算 $\bar{\boldsymbol{z}}^*$ 与每个候选嵌入 $\bar{\boldsymbol{z}}^{(i)}$之间的欧几里得距离：

$$r(i) = \left\|\bar{\boldsymbol{z}}^* - \bar{\boldsymbol{z}}^{(i)}\right\|_2.$$

最 K 相似的候选项被检索为

$$\mathcal{N}_K(\boldsymbol{s}^*) = \text{TopK}_{i\in[N]}\big(-r(i)\big),$$

并用于构建最终上下文 C。

我们在 Phi-4-MultiModal-Instruct (Phi-4-MM) [23]上评估了 TICL。

2.3. 用声学重新排序优化上下文选择

为了使上下文 C 更好地匹配测试音频的声学特征，我们通过引入声学重新排序步骤，将 TICL 流程扩展为 TICL+。先前的结果表明，在语义相似度之后，Whisper嵌入是效果第二好的检索方法 [22]。Whisper嵌入能够捕捉输入语音的许多特征，例如韵律、说话人身份和发音，而这些特征无法通过纯粹的词汇表示体现出来 [24]。受此启发，我们采用基于声学的距离度量来优化上下文示例的选择。

使用语义上最相似的前 M 个候选项 $\mathcal{N}_M(\boldsymbol{s}^*)$，其中 $M = 300$，我们利用冻结的语音编码器 g 预先计算的嵌入来计算声学相似度： $\mathcal{X}_{\text{audio}} \to \mathbb{R}^p$。我们的实验使用了 Whisper-large-v3-turbo。经 ℓ_2归一化的声学嵌入预先计算如下：

$$\bar{g}(\boldsymbol{s}) = \frac{g(\boldsymbol{s})}{\|g(\boldsymbol{s})\|_2}.$$

随后计算测试音频与每个候选项之间的声学距离，计算方式如下：

$$r_{\text{acoustic}}(i) = \|\bar{\boldsymbol{a}}^* - \bar{\boldsymbol{a}}^{(i)}\|_2.$$

其中，$\bar{\boldsymbol{a}}^* = \bar{g}\ (\boldsymbol{s}^*)$，$\bar{\boldsymbol{a}}^{(i)} = \bar{g}\ (\boldsymbol{s}^{(i)})$。根据 r_{acoustic} (i) 对候选样本重新排序，并选择在声学上最接近的前 K 个样本作为：

$$\mathcal{N}_K^{\text{acoustic}}(\boldsymbol{s}^*) = \text{TopK}_{i\in\mathcal{N}_M(\boldsymbol{s}^*)}\big(-r_{\text{acoustic}}(i)\big).$$

最终的 SICL 上下文 C 由 $\mathcal{N}_K^{\text{acoustic}}(\boldsymbol{s}^*)$ 构建而成。这一两阶段检索过程确保所选的上下文示例既具有词汇相关性，在声学上也与测试话语相似。

3. 实验结果与分析

为评估儿童语音识别性能，我们使用了四个语料库：My Science Tutor（MyST）[25], , 包含针对 8–11 岁学生的科学辅导对话；OGI 儿童语音语料库 [26] , 包含约 100 小时、由 5–16 岁儿童朗读或根据提示进行的语音；埃德蒙顿叙事规范工具（ENNI）[27, 28],, 包含 4–9 岁儿童对故事的复述和续写；以及雷德蒙句子复述（RSR）[29] , 由面向 5–9 岁儿童的句子复述任务组成，其中包括患有发育性语言障碍的儿童。所有数据集均经过预处理，并按照 [22] 中概述的流程生成相应的候选集。我们使用 Phi-4-MM 评估 TICL+。

表 1 显示，在 TICL 流程中加入声学重排序，可显著提升所有数据集上的识别准确率。TICL+ 的性能较零样本性能最高提升 53.3%，较 TICL 最高提升 37.62%。这些提升可能源于伪标签生成器的局限性，它往往无法准确转写儿童语音。选择语义上最接近的前 300 条话语有助于排除无关示例，而声学重排序步骤则通过优先选取与测试话语声学相似的样本，进一步优化上下文，不受词汇相似性的影响。

在全部四个数据集上，TICL+的表现始终优于零样本和TICL。在 MyST 上观察到最大的相对提升（53.3%）；该数据集包含会话语音，说话者年龄和背景噪声的变异性很高。这表明，声学相似性可能有助于识别说话者及环境条件相近的示例，从而实现更稳健的上下文对齐。

在 OGI 和 ENNI 上观察到的提升较小，这两个数据集主要包含朗读语音或结构化语音，在语义检索步骤中，词汇重叠本身已能提供有力的指导。在这些情况下，声学过滤器可能通过考虑说话人之间的发音差异和发育变异性发挥作用。尽管 RSR 也包含朗读语句，其表现仍有所提升，这表明即使词汇多样性有限，声学重排序仍然有效，这可能是因为它能够识别发育阶段相似的说话人。

总体而言，这些结果表明，声学相似性为词汇相似性提供了互补信息，使模型能够更好地捕捉儿童语音中固有的说话人间和说话人内变异。所有语料库上的性能提升证明，TICL+ 适用于低资源且随发育而变化的语音领域。

表 1。*Phi-4-MM*的儿童语音结果。↓词错误率%。

方法	k	MyST	OGI	ENNI	RSR
零样本	0	12.81	16.17	14.37	20.06
TICL	1	17.27	9.55	17.57	18.92
	2	11.77	8.94	14.07	18.92
	3	11.69	8.75	13.54	18.90
	4	11.81	8.52	13.75	19.54
	Δ_{rel}	8.7%	47.3%	5.8%	5.8%
TICL+	1	11.48	8.84	14.83	12.89
	2	**10.17**	7.97	12.01	12.26
	3	10.52	7.78	**11.52**	**12.19**
	4	10.57	**7.55**	**11.52**	12.75
	Δ_{rel}	**20.6%**	**53.3%**	**19.8%**	**39.2%**

4. 结论

在这项工作中，我们在 TICL 流程中引入了声学重排序步骤，构建了 TICL+ 框架，从而提升了 SICL 在儿童语音识别中的表现。所提出的方法同时利用声学相似性和语义相似性，构建更有效的上下文示例。在四个儿童语音语料库上的实验表明，该方法取得了显著的性能提升：与零样本性能相比，词错误率相对降低最高达 53.3%；与基线 TICL 相比，最高降低 37.62%。这些发现凸显了在选择上下文示例时纳入多种因素的价值，并为构建更稳健、可扩展的儿童语音自动语音识别系统铺平了道路。

5. 致谢

本研究得到了美国国家科学基金会资助，项目编号为 # 2229873。本研究通过高级网络基础设施协调生态系统：服务与支持（ACCESS）计划提供的分配额度 beiq-delta-gpu，使用了国家超级计算应用中心的 Delta 系统。该计划由美国国家科学基金会资助，项目编号为 #2138259、#2138286、# 2138307、#2137603 和 #2138296。

6. 参考文献

[1] 布法罗大学，纽约州立大学，“该项目：面向特殊教育的国家人工智能研究院——技术”，https://www.buffalo.edu/ai4exceptionaled/technology.html，2025年，访问日期：2025-10-30。[2] Dancheng Liu 等，“使用自动语音识别对言语障碍儿童进行自动筛查：机遇与挑战”，载于*AAAI* 研讨会系列论文集，2024年，第4卷。[3] Laura L Koenig、Jorge C Lucero 和 Elizabeth Perlman，“儿童擦音中的语音产生变异性以及

成人：函数型数据分析的结果，《美国声学学会杂志》，第124卷，第3170页，2008年5月，第3158页。

[4] Laura L Koenig 和 Jorge C Lucero，“女性与儿童的塞音浊化和口内压力曲线”，《美国声学学会会刊》，第123卷，第2期，第1077–1088页，2008年。[5]Sungbok Lee、Alexandros Potamianos 和 Shrikanth Narayanan，“儿童语音的声学特征：时间和频谱参数的发展变化”，《美国声学学会会刊》，第105卷，第3期，第1455–1468页，1999年。[6]Sungbok Lee、Alexandros Potamianos 和 Shrikanth Narayanan，“儿童语音分析：时长、音高和共振峰”，载于第五届欧洲语音通信与技术会议，1997年。[7] Houri K Vorperian 和 Ray D Kent，“儿童元音声学空间的发展：声学和解剖数据综述”，2007年。[8] Bruce L Smith，“儿童语音中时长与时间变异性的关系”，《美国声学学会会刊》，第91卷，第4期，第2165–2174页，1992年。[9] Ahmed Adel Attia、Jing Liu、Wei Ai、Dorottya Demszky 和 Carol Espy-Wilson，“Kid-whisper：弥合儿童与成人自动语音识别性能差距的探索”，载于*AAAI/ACM*人工智能、伦理与社会会议论文集，2024年，第7卷，第74–80页。[10] Rishabh Jain、Andrei Barcovschi、Mariam Yahayah Yiwere、Peter Corcoran 和 Horia Cucu，“将 Whisper 模型适应用于儿童语音识别”，载于*INTERSPEECH*，2023年。[11] Rishabh Jain、Andrei Barcovschi、Mariam Yiwere、Dan Bigioi、Peter Corcoran 和 Horia Cucu，“基于 wav2vec2 的实验研究：利用自监督学习方法改进儿童语音识别”，*IEEE Access*，2023年。[12] Ruchao Fan 和 Abeer Alwan，“草稿：一种减少自监督学习领域偏移的新框架及其在儿童自动语音识别中的应用”，2022年。[13] Ruchao Fan、Yunzheng Zhu、Jinhan Wang 和 Abeer Alwan，“迈向更好的自监督模型领域适应：儿童自动语音识别案例研究”，*IEEE*信号处理精选主题期刊，第16卷，第6期，第1242–1252页，2022年。[14] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah 等，“语言模型是少样本学习者”，载于神经信息处理系统进展，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，2020年，第33卷，第1877–1901页，Curran Associates, Inc.[15] Qingxiu Dong、Lei Li、Damai Dai 等，“上下文学习综述”，载于*2024*年自然语言处理实证方法会议论文集，Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen 编，美国佛罗里达州迈阿密，2024年11月，第1107–1128页，计算语言学协会。[16]Siyin Wang、Chao-Han Yang、Ji Wu 和 Chao Zhang，“Whisper 能否执行基于语音的上下文学习？”，载于*ICASSP 2024 - 2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2024年，第13421–13425页。

[17] Nathan Roll、Calbert Graham、Yuka Tatsumi 等，“上下文学习通过以类似人类的方式适应说话人和语言变体来提升语音识别”，2025年。[18] Jiaming Zhou、Shiwan Zhao、Jiabei He、Hui Wang、Wenjia Zeng、Yong Chen、Haoqin Sun、Aobo Kong 和 Yong Qin，“M2R-Whisper：用于增强 Whisper 的多阶段、多尺度检索增强”，载于*ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2025年，第1–5页。[19] Zihao Zhao、Eric Wallace、Shi Feng 等，“使用前先校准：提升语言模型的少样本性能”，载于第*38*届国际机器学习会议论文集，Marina Meila 和 Tong Zhang 编，2021年7月18–24日，机器学习研究论文集第139卷，第12697–12706页，PMLR。[20] Zhao Yang、Yuanzhe Zhang、Dianbo Sui 等，“利用两阶段行列式点过程为上下文学习选择代表性示例”，载于*2023*年自然语言处理实证方法会议论文集，Houda Bouamor、Juan Pino 和 Kalika Bali 编，新加坡，2023年12月，第5443–5456页，计算语言学协会。[21] Rishabh Agarwal、Avi Singh、Lei Zhang 等，“多示例上下文学习”，载于神经信息处理系统进展，A. Globerson、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. Tomczak 和 C. Zhang 编，2024年，第37卷，第76930–76966页，Curran Associates, Inc.[22] Haolong Zheng、Yekaterina Yegorova 和 Mark Hasegawa-Johnson，“TICL：用于语音上下文学习的文本嵌入 K 近邻方法，解锁大型多模态模型的语音识别能力”，*arXiv*预印本 *arXiv:2509.13395*，2025年。[23] 微软，：Abdelrahman Abouelenin、Atabak Ashfaq、Adam Atkinson 等，“Phi-4-mini 技术报告：通过 LoRA 混合实现紧凑而强大的多模态语言模型”，2025年。[24] Yuan Gong、Sameer Khurana、Leonid Karlinsky 和 James Glass，“Whisper-AT：抗噪自动语音识别器也是强大的通用音频事件标注器”，载于*INTERSPEECH 2023*。2023年8月，interspeech2023，ISCA。[25] Sameer Pradhan、Ronald Cole 和 Wayne Ward，“我的科学导师（MyST）——大型儿童对话语音语料库”，载于*2024*年计算语言学、语言资源与评测联合国际会议论文集（*LREC-COLING 2024*），2024年，第12040–12045页。[26] Khaldoun Shobaki、John-Paul Hosom 和 Ronald Cole，“OGI 儿童语音语料库及识别器”，2000年10月，第258–261页。[27] Phyllis Schneider、Denyse Hayward、Rita Vis Dub´e 等，“使用埃德蒙顿叙事规范工具看图讲故事”，《言语语言病理学与听力学杂志》，第30卷，第4期，第224页，2006年。[28] Dancheng Liu 和 Jinjun Xiong，“FASA：一种用于提取高质量对齐儿童语音数据的灵活自动语音对齐器”，2024年。[29]AI4ExceptionalEd，“雷德蒙句子复述（RSR）”，https://huggingface.co/datasets/ai4exceptionaled/Redmond-Sentence-Recall，Hugging Face 数据集；访问日期：2025年9月8日。