面向国际中文教育的语音偏误数据集构建与应用

孙佳杰[1]，陈熙之[1]，凌锋[2]，袁丹[2]，兰韵诗[1]，王晔[1]
1. 华东师范大学数据科学与工程学院，上海 200062；
2. 华东师范大学国际汉语文化学院，上海 200062

摘要

针对国际中文教育中缺乏公开、标准化细粒度的汉语语音偏误数据集及通用模型非母语发音诊断适配有限的问题，构建面向留学生的汉语语音偏误数据集。采用融合基础声学对齐与专家诊断的十层结构化标注规范，建立“机器预标注—人工精修—专家复核”的人机协同流程，并完成数据清洗、质量复核与结构化入库，以保证样本可追溯和可训练。结果获得340份真实发音样本、12万余条声韵级时间对齐切片和数万条偏误诊断标签；在非母语口音自动语音识别和端到端语音偏误诊断任务中，基于该数据集微调的Paraformer-large与Qwen2-Audio-7B均表现出更好的任务适配效果。该数据集可为国际中文教育场景下的模型微调与智能发音反馈提供监督数据，也为垂直领域高质量数据集建设提供数据治理参考。

关键词

国际中文教育；语音偏误诊断；高质量数据集；人机协同标注；数据治理

中图分类号：TP391.41　　文献标志码：A

Construction and application of a pronunciation error dataset for international Chinese language education

SUN Jiajie[1], CHEN Xizhi[1], LING Feng[2], YUAN Dan[2], LAN Yunshi[1], WANG Ye[1]
1. School of Data Science and Engineering, East China Normal University, Shanghai 200062, China;
2. School of International Chinese Studies, East China Normal University, Shanghai 200062, China

Abstract

To address the lack of open and standardized fine-grained pronunciation error datasets for international Chinese language education and the limited adaptation of general models to non-native pronunciation diagnosis, a Chinese pronunciation error dataset for international students was constructed. A ten-layer structured annotation scheme was developed by integrating basic acoustic alignment with expert diagnostic labels, and a human-machine collaborative workflow of machine pre-annotation, manual refinement, and expert review was established. Data cleaning, quality review, and structured storage were also completed to ensure that the samples were traceable and trainable. As a result, 340 authentic speech samples, more than 120,000 time-aligned slices at the initial-final level, and tens of thousands of fine-grained diagnostic labels were obtained. To evaluate the dataset, two representative tasks were conducted: automatic speech recognition for non-native accented speech and end-to-end pronunciation error diagnosis. The results

showed that, after fine-tuning on the dataset, Paraformer-large and Qwen2-Audio-7B achieved better task adaptation. The dataset provides supervised data for model fine-tuning and intelligent pronunciation feedback in international Chinese language education, and it also offers a reference for high-quality dataset governance in vertical domains.

Key words
international Chinese language education, pronunciation error diagnosis, high-quality dataset, human-machine collaborative annotation, data governance

0 引言

国际中文教育是展示真实、立体、全面的中国的重要窗口，在促进民心相通与文明互鉴中发挥着重要作用[1]。正如习近平总书记在2024年11月15日致世界中文大会暨孔子学院成立20周年的贺信中所指出，中文承载着中华民族数千年的文明智慧，是中国贡献给世界的重要公共文化产品，支持服务国际社会开展好中文教育是中国作为母语国的责任[2]。

发音学习是语言交际的重要基础[3]。在国际中文教育场景中，语音偏误指学习者在汉语发音过程中相对于目标语音规范产生的系统性或阶段性偏离。此类偏离既包括声母、韵母等音段层面的替换、缺失和混淆，也包括声调、重音、停连、语调和节奏等超音段层面的异常。与一般口音差异相比，语音偏误往往会影响语义辨识、交际可懂度或教学目标达成，因此需要在教学中加以定位、分类并反馈给学习者。

语音偏误诊断是在检测偏误是否存在的基础上，进一步确定偏误发生的时间位置、所属语音层级、具体类型及可能成因。例如，系统既要判断某个字是否“读错了”，也要说明该偏误属于声母偏误、韵母偏误、声调偏误还是韵律偏误，并尽可能给出对应的时间戳与教学反馈。对于国际中文教师而言，发音作业批改不宜停留在“对或错”的判断上，还需要说明“错在哪里、为什么错、如何纠正”。

普通自动语音识别（automatic speech recognition，ASR）系统的目标通常是输出尽可能准确的文本，而语音偏误诊断则需要保留并解释学习者发音中的异常声学信息。处理留学生发音时，通用ASR模型可能根据上下文将非标准发音规整为正确汉字，从而在文本转写阶段掩盖实际偏误。因此，面向国际中文教育的高质量语音偏误数据集不仅需要保存音频和文本，还应提供声韵级时间对齐、偏误类型、专家诊断意见等多层级标注信息，以支撑自动语音识别、发音评测和个性化教学反馈等任务。

现有语音偏误检测技术与自动评测工具在应用中仍存在局限。较为突出的问题是，该领域缺少公开、标准化的留学生汉语语音偏误数据集[4][5]。非母语（L2）语音的声韵级或音素级诊断标注工作量较大；同时，留学生发音变体具有一定模糊性，不同标注者对错误类型和时间边界的判断容易出现分歧，标注一致性难以保障[6]。相关研究还表明[7]，通用大模型及预训练声学模型在非母语发音诊断等垂直任务中可能存在领域适配不足。未经特定领域数据微调的模型，可能难以稳定识别留学生发音中的母语迁移特征；在部分细粒度诊断实验设置中，模型表现甚至接近或低于简单基线，难以生成具有针对性的教学反馈[7]。

围绕数据缺乏、标注成本高和模型适配不足等问题，本研究与华东师范大学国际汉语文化学院合作，设计并实施了面向国际中文教育场景的语音偏误数据集构建流程，并开展代表性应用验证。本文的主要工作如下：

(一) 提出面向国际中文教育的十层结构化标注规范，并建立“机器预标注—人工精修—专家复核”的人机协同生产流程。该标注规范将基础声学对齐与专家诊断标注相结合：前四层用于记录句子、汉字、音节和声韵等基础时间对齐信息，后六层用于记录声母偏误、韵母偏误、声调偏误、韵律偏误和副语言现象等专家诊断信息。基于该流程，本研究构建了包含340份真实留学生发音样本的数据集，形成超过12万条声韵级时间对齐切片和数万条细粒度偏误诊断标签。

(二) 验证该数据集在两项代表性任务中的应用可行性。本文分别选取非母语口音ASR与端到端语音偏误诊断作为应用验证任务：前者利用基础标注层验证数据集对字级和句级转写的支持作用，后者利用高级诊断层验证数据集对偏误类型、时间边界和结构化反馈生成的支持作用。上述实验定位为数据集应用验证，相关结论限于本文所选代表性模型和测试集设置，并非对所有主流模型进行全面横向比较。

(三) 总结面向垂直领域高质量数据集构建的数据治理经验。本文从分层标注、人机协同、质量控制、结构化入库和数据闭环等方面梳理本研究的数据治理流程，并讨论其在医学语音、方言资源、儿童语言发展、特殊教育语音评估等涉及专家标注和时间对齐任务中的可迁移价值。

1 相关研究

1.1 现有数据集与模型的局限性

在外国留学生汉语语音偏误检测研究中，现有数据集与通用模型主要存在三方面限制。

首先，现有开源语料对细粒度偏误标注的支持不足。AISHELL系列主要服务于普通话语音识别任务[4]、SpeechOcean762等语料则更多用于非母语发音评测[5]。这类数据集通常缺少逐字、逐音素的时间戳，以及偏误类型、偏误原因等多层级诊断信息。因此，语音偏误检测与诊断(mispronunciation detection and diagnosis，MDD）模型很难准确定位局部偏误，也难以学习具体的声学差异[8]。

其次，人工标注的一致性不足。汉语二语发音中有不少介于可接受变体和偏误之间的发音，标注者对边界和类型的判断并不总是一致。针对日语母语者汉语发音语料的研究显示，两组标注者的音素标注一致率平均为80.7%，不一致标注占19.3%，标注者间的判断差异是重要原因[9]。这类分歧会把标签噪声带入数据集，并影响后续模型训练[10]。

第三，通用语音模型用于发音诊断也有局限。Whisper等模型主要面向通用识别、翻译和语言识别，并非为二语发音偏误检测专门训练[11]。已有研究表明，预训练模型迁移到二语偏误检测任务时，仍会受偏误样本稀疏、音素级标注差异和母语迁移特征影响[12]；GPT-4o Voice Mode的评测也显示，大音频语言模型在细粒度音频理解上并不稳定，结果容易受任务类型和评测方式影响[13]。自动口语评测研究进一步指出，现有多模态大语言模型(multimodal large language model，MLLM）在内容评价上较强，但在发音、流利度、重音和语调等维度仍需专门优化[14]。

1.2 自动语音识别与强制对齐的辅助作用及使用局限

在数据集构建中，自动语音识别与强制对齐常被用于为音频生成初步的文本对应与时间边界[15]。强制对齐技术，如蒙特利尔强制对齐器（Montreal Forced Aligner，MFA）通常基于隐马尔可夫模型，能够在已知文本的前提下，自动估算字、词或音素在音频中的时间戳，从而减少从零开始人工切分音频的工作量[16]。

然而，这些工具用于留学生语音时仍有误差。MFA依赖给定文本做强制对齐，若学习者出现替换、省略或明显变调，语音片段和标准音素序列之间就会拉开距离，边界可能被切错。已有研究指出，基于强制对齐的发音评分（goodness of pronunciation，GOP）方法容易受声学差异影响，出现标注或切分误差[17]。

通用ASR也不能直接当作偏误标注工具。它的目标是转成可读文本，而不是保留每一个发音偏差。对于非母语语音，音素识别本身存在不确定性，且“识别音素准确”和“只有一种正确读法”这两个前提并不总成立[18]。如果系统把有偏误的读音转成了目标汉字，文本层面就看不到原始发音问题。因此，机器预标注只能作为初稿，仍需人工复核。

1.3 语音偏误诊断方法的发展与微调数据需求

语音偏误诊断方法经历了从统计模型、深度学习网络到多模态大模型的发展。早期方法通常结合隐马尔可夫模型和后验概率评分[19]。这类方法依赖准确的音素级时间边界，时间对齐误差会直接影响评分结果。

进入深度学习阶段后，研究开始采用连接时序分类（connectionist temporal classification，CTC）、注意力机制等端到端序列结构，减少对强制对齐预处理的依赖，直接建立从语音到偏误序列的映射[20]。这类方法减少了中间步骤的误差传递，但需要较多带有偏误标签的语音数据，才能学习标准发音和错误发音之间的差异。

近年来，支持音频输入的MLLM开始被应用于语言教育任务[14]。与只能输出简单分类结果的早期模型相比，MLLM可以生成更完整的结构化诊断反馈。要让这类模型适配汉语语音偏误诊断，通常需要采用低秩自适应（low-rank adaptation，LoRA）等参数高效微调方法[21]。不过，现有语料较少同时覆盖声母、韵母、声调、韵律和专家诊断标签。受限于这类微调数据，很多基于大模型的汉语教学应用仍停留在流利度打分层面[22]。

因此，面向留学生汉语语音偏误诊断，仍需要从数据收集、时间对齐和专家标注环节入手，构建具有多层级诊断标签的数据集，为后续模型微调和教学反馈生成提供支持。

2 汉语语音偏误数据集的构建与质量控制

用于特定领域的模型调优，需要专门的领域数据集。在调研当前多模态大模型批改留学生语音作业的能力后，本研究与华东师范大学国际汉语文化学院合作，通过“机器预标注-人工精修-专家复核”的流程，构建了汉语语音偏误数据集。

2.1 细粒度结构化标注规范设计

汉语二语习得中的语音偏误具有多维特征，且常伴随交叉表现。为满足细粒度

诊断需求，本研究与华东师范大学国际汉语文化学院共同制定了十层结构化标注规范，以记录具体的偏误类型并定位音频时间戳。

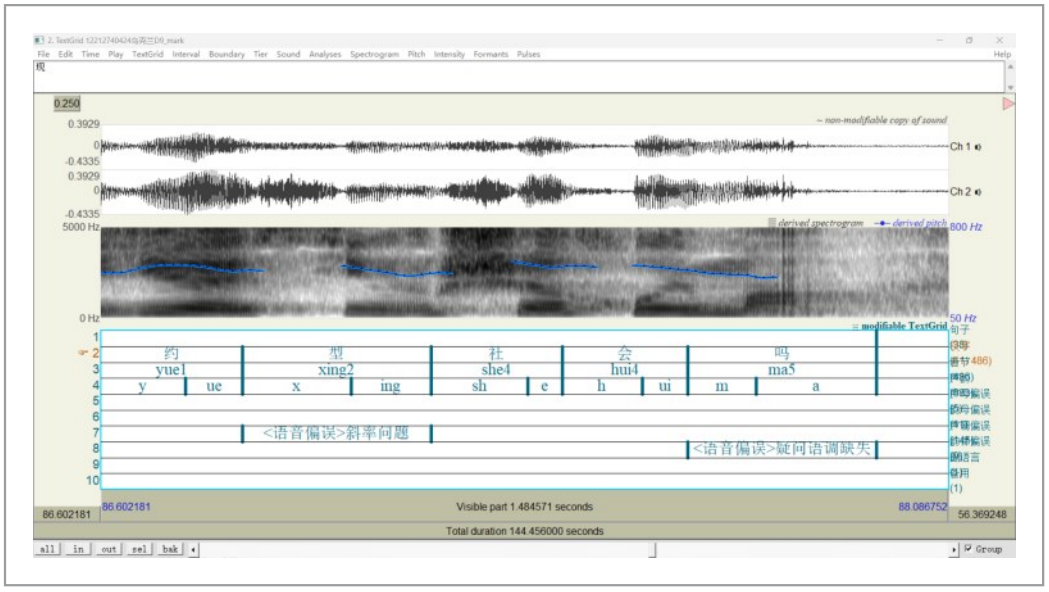

图1 TextGrid十层结构化标注结果示例

整个标注体系自上而下划分为两大模块：第一部分为前四层基础标注层（句子、汉字、音节、声韵层），主要解决语音片段与文本在时间轴上的对应问题。其中，句子层确定句法边界，音节与声韵层进行音节结构拆分。第二部分为后六层高级诊断层（声母偏误、韵母偏误、声调偏误、韵律偏误、副语言处理及备用层），由专家依据汉语发音教学理论进行标注，要求在切分区间内记录具体的语音偏误类型（如标出“n-l混”等）。

2.2 基于人机协同的生产流程构建

如果前四层的切分工作完全交由人工逐字逐句完成，工作量巨大且容易因疲劳导致边界标记的不一致。因此，本研究设计了人机协同的生产流程，具体框架如图2所示。

在数据预处理阶段，系统将收集到的作业录音统一转换为16000Hz的单声道WAV格式。为避免首尾静音段对后续声学算法造成干扰，系统引入基于能量门限的语音活动检测（voice activity detection，VAD）[23]，自动识别并切除音频首尾的静音部分。同时，鉴于作业录制环境的背景噪声问题，预处理流程采用基于谱减法的noisereduce库对音频进行降噪[24]。

在人工智能辅助预标注环节，系统通过ASR模型将学生的音频转换为文本序列，随后使用MFA执行强制对齐。MFA基于文本与声学特征，计算出每个音节、音素在音频中的起始与结束时间戳，自动生成前四层的初步TextGrid标注文件。

在需要人工精修的高级诊断层阶段，

图2 数据集人机协同流水线及质量控制全流程架构图

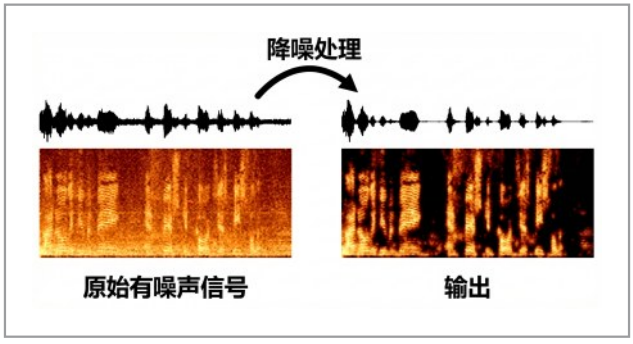

图3 noisereduce降噪前后谱图对比图

由华东师范大学国际汉语文化学院的研究生组成标注团队。标注人员使用跨平台语音分析软件Praat①，在机器预生成的基础上对时间边界进行校正，并重点完成对后六层的专业偏误判断。由此，计算机处理了基础的时间切分工作，而具备专业背景的学生则将精力集中于高级语言偏误分析。

2.3 质量验证与数据集成

为确保标注数据的质量，国际中文教育相关院系的教师团队对初稿进行了人工交叉复核。复核人员通过逐条审听，重点校验声韵边界的切分精度与偏误归类的合理性。为量化标注质量，研究抽取数据子集计算了标注者间一致性（inter-annotator agreement, IAA）[10]，其Cohen's Kappa系数达0.81，表明该数据集在复杂发音偏误的主观诊断上具备高度可靠性。此外，对于录音质量过低、重度口音无法辨识或层级间时间戳冲突的无效样本，系统均予以剔除。经过上述复核与清洗，最终保留了整体准确率在95%以上的数据。其中，准确率根据专家抽检样本中标签与复核结果一致的比例计算。

在数据集成方面，为了将TextGrid文件中的离散标注用于后续的模型训练，本研究使用Parselmouth声学解析库[25]对标注文件进行结构化处理。脚本通过Parselmouth读取TextGrid文件中的十个层级对象（item），遍历并提取所有存在人工记录的有效标注区间（interval）。系统将区间内的文本内容（text）提取后存入数据库偏误记录表的label_content

①https://www.fon.hum.uva.nl/praat/

字段，并将其对应的层级名称属性（name）存入 label_level 字段。这种提取方式将分散的 TextGrid 文件转换为结构化数据库记录，方便后续检索与使用。

2.4 数据集统计分析

经过一个学期的收集与复核，本研究最终整理了 340 份留学生语音作业样本。系统将其存储于关系型数据库中，主要由“语音作业实体表”与“层级标注记录表”构成。前者存储作业的音频路径、ASR 文本、学生外键等元数据；后者则通过作业主键关联，存储具体的毫秒级起始时间戳与偏误诊断内容。

在地域特征与母语背景上，该批留学生的来源涵盖了东亚、东南亚、欧美等主要生源地，其中以韩国等东亚国家的母语者占比最高，这与2018年来华留学生的宏观人口统计分布趋势基本一致[26]。在语言水平分布上，样本以汉语水平考试（Hanyu Shuiping Kaoshi，HSK）4 级及以上的中高级学习者为主，也符合来华留学生相关统计研究的结论[27]。

从标注数据量来看，前四层基础切片采用算法预标注与人工微调相结合的方式生成，仅声韵层就包含了超过12万条的精细对齐数据。后六层则由专家进行人工诊断，记录了大量具体的汉语语音偏误样本（如声调偏误占比最高，达9456条，其次为韵母与声母偏误）。

为直观呈现数据分布特征，本研究将留学生的母语构成、HSK 等级水平与核心十层数据标注量级进行了综合可视化呈现（详见图 4）。

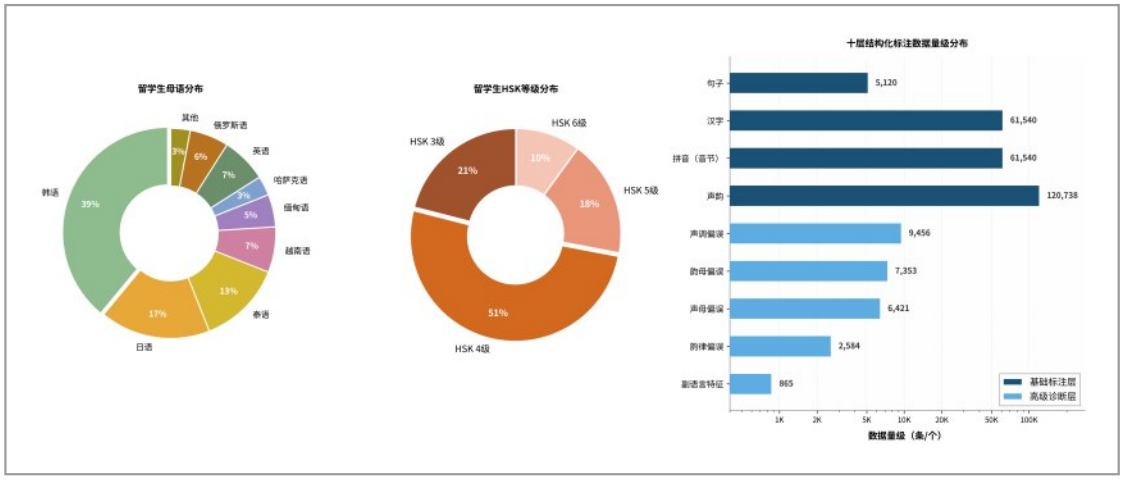

图4　语料库分布特征综合视图

通过上述的数据收集、标注及结构化处理流程，本研究获得了一个具备一定代表性、且经过专业规范复核的留学生汉语语音偏误数据集。该数据集能够为后续不同模型在特定领域的调优应用提供数据支持。

3 数据集特定领域应用验证

为验证所构建语音偏误数据集的任务适用性，本文设置了两个代表性应用任务。十层标注体系具有明确的层级分工：前四层基础标注层记录句子、汉字、音节和声韵等时间对齐信息，可用于支撑非母语口音ASR等转写任务；后六层高级诊断层记录声母、韵母、声调、韵律和副语言等专家偏误判断，可用于支撑语音偏误诊断与教学反馈生成任务。基于上述分工，本文分别选取Paraformer-large和Qwen2-Audio-7B作为两项任务的主要模型，前者用于考察基础对齐数据对非母语口音识别的支持作用，后者用于考察专家诊断数据对结构化偏误诊断的支持作用。

3.1 实验目标、模型选型与实验边界

本文实验的目的，是检验语音偏误数据集能否支撑具体下游任务，并非对主流语音模型进行全面横向评测。实验按照十层标注的功能分工设计：前四层提供句子、汉字、音节和声韵层面的时间对齐信息，用于非母语口音ASR；后六层提供声母、韵母、声调、韵律和副语言等专家诊断标签，用于端到端语音偏误诊断。

ASR任务以Paraformer-large为主要微调模型，并设置Whisper-large-v3-turbo零样本结果和未微调Paraformer-large作为对照。这样设计是为了比较通用模型、领域模型和经本文数据微调后的模型在同一测试集上的表现，重点观察基础标注层能否改善非母语口音转写。

语音偏误诊断任务选用Qwen2-Audio-7B，是因为该任务需要同时输出偏误类型、时间边界和JSON结构化结果。Wav2Vec2.0、HuBERT等自监督语音表征模型更适合作为特征提取或下游分类模型的基础，若直接用于本文任务，还需要另设CTC解码器、分类头、时间戳预测或格式约束模块。此时实验差异会同时来自模型结构和任务适配模块，难以单独归因于本文数据集。因此，本文暂不报告这两类模型的定量结果，而将其作为后续在统一适配框架下扩展的基线。

本文实验结论仅限于所选模型、数据划分和测试集设置。表1中列出相关模型的角色和任务适配关系。

3.2 实验环境配置

本研究的实验环境基于单NVIDIA A100 80GB GPU，操作系统采用Ubuntu 24.04 LTS，深度学习框架为PyTorch 2.0及以上版本（配合CUDA 12.x工具包）。在模型微调阶段，自动语音识别任务使用funasr 1.2.6框架，多模态大模型微调任务则使用ms-swift 3.8.0框架。

3.3 应用1：基于基础标注层的非母语口音ASR微调

3.3.1 任务定义与实验设置

受母语音系迁移和个体口音差异影响，留学生汉语发音在声母、韵母、声调及韵律等方面可能与普通话规范发音存在偏离，通用语音识别模型在转写这类语音时容易出现识别错误。为考察本文基础标注层对非母语口音ASR任务的支持作用，本节选取Paraformer-large作为主要微调模型。该模型属于非自回归端到端ASR模型，适合完成汉字序列转写，并可用于字级和句级识别效果评估。同时，本文保留Whisper-large-v3-turbo作为通用多语种ASR的零样本对照，用于观察通用模型在本文测试集上的表现。

在数据划分方面，本节从前四层基础

表1 主要语音模型的技术路线与本文任务适配关系

模型	技术路线	典型用途	与本文任务的关系	是否纳入本文定量实验
Whisper-large-v3-turbo	多语种生成式ASR	通用语音转写	作为通用零样本ASR对照模型	是
Paraformer-large	非自回归端到端ASR	中文语音识别	作为ASR任务主要微调模型	是
Wav2Vec2.0	自监督语音表征模型	语音表征学习、ASR微调、发音诊断	需额外构建CTC解码器或分类头后适配本文任务	否，列为后续工作
HuBERT	自监督语音表征模型	语音表征学习、音素级诊断、下游分类	需额外构建任务专用诊断模块	否，列为后续工作
Qwen2-Audio-7B	音频语言多模态模型	音频理解、语音问答、结构化生成	作为端到端语音偏误诊断主要微调模型	是

本表用于说明不同模型的技术路线及其与本文任务的适配关系。本文只报告已在统一数据划分下完成推理或微调验证的模型结果。

标注层（句子、汉字、音节和声韵层）中提取3087条语音片段。其中，254份作业共2557条片段作为训练集，52份作业共320条片段作为验证集，另取34份作业共210条片段作为独立测试集。评估指标采用字错误率（character error rate，CER）和句错误率（sentence error rate，SER）。CER根据识别结果相对于参考文本的插入、删除和替换字符总数占参考文本字符总数的比例计算，用于衡量字符级识别错误。SER统计至少包含一处识别错误的句子占总句数的比例，用于反映整句层面的识别情况。

3.3.2 结果评估与分析

测试集评估结果见表2。使用本文基础对齐样本进行微调后，Paraformer-large在非母语口音测试集上的字错误率（CER）由6.1%降至4.2%，相对下降31.1%；句错误率（SER）由8.3%降至6.7%。Whisper-large-v3-turbo在零样本设定下的CER和SER分别为7.8%和12.4%，均高于未微调Paraformer-large的对应结果。

表2 非母语口音ASR任务的代表性应用验证结果

模型	CER (%)	SER (%)
Whisper-large-v3-turbo (零样本基线)	7.8	12.4
Paraformer-large (未微调基线)	6.1	8.3
Paraformer-large (基于本文数据集微调)	**4.2**	**6.7**

在本文测试集与实验设置下，Whisper-large-v3-turbo的零样本识别误差高于未微调和微调后的Paraformer-large。该结果可能与模型训练语料、解码方式以及中文非母语口音场景的领域差异有关。需要说明的是，本文未进一步分析Whisper的错误类型，也未将Wav2Vec 2.0、HuBERT等模型纳入同一测试集进行横向比较，因此上述结果不应被解释为对通用语音模型整体能力的判断。

就数据集应用验证而言，基于本文基础标注层微调后的Paraformer-large在CER和SER上均有所下降，说明前四层基础对齐数据能够为非母语口音ASR提供有效监督信号。该结果初步验证了本文数据集在中文非母语口音识别任务中的应用价值。

3.4 应用2：基于高级诊断层的端到端语音偏误诊断

3.4.1 任务定义与实验设置

在语音偏误诊断任务中，系统不仅需要识别语音内容，还需要判断偏误类别、定位偏误时间边界，并输出可被程序解析的结构化诊断结果。传统“自动语音识别+文本推理”级联架构容易在文本转换阶段丢失细微声学特征，因此本文选取支持音频输入的Qwen2-Audio-7B作为端到端诊断模型。该模型能够结合音频和文本提示生成结构化输出，适合与本文后六层高级诊断标注对接。

实验选取未经微调的Qwen2-Audio-7B作为零样本基线，从后六层过滤出约3000条包含明确专家偏误标签的样本。其中，272份作业（约2580个片段）用于训练，另外68份作业（420个片段）作为独立测试集。训练主要采用LoRA（Low-Rank Adaptation）微调策略：冻结音频编码器参数，仅对投影层及语言模型的注意力与多层感知机模块进行参数更新。

3.4.2 结果评估与分析

本研究将测试集数据输入微调前后的Qwen2-Audio模型，从时间戳精度、偏误检测性能与系统可靠性三个维度进行评估，结果如表3所示。

表3 端到端语音偏误诊断任务的代表性应用验证结果

评估维度	评估指标	微调前	微调后	提升幅度
时间戳精度	MAE (ms) ↓	283	**156**	44.9%
	IoU ↑	0.82	**0.91**	11.0%
偏误检测性能	声母偏误 (Macro-F1)	0.72	**0.86**	19.4%
	韵母偏误 (Macro-F1)	0.68	**0.83**	22.1%
	声调偏误 (Macro-F1)	0.61	**0.89**	45.9%
	韵律偏误 (Macro-F1)	0.54	**0.78**	44.4%
	副语言 (Macro-F1)	0.71	**0.84**	18.3%
综合可靠性	JSON合规率 (%)	87.2	**93.4**	6.2%
	人工评估-准确性	3.4/5	**4.5/5**	32.4%
	人工评估-实用性	3.1/5	**4.3/5**	38.7%

时间戳精度通过平均绝对误差（mean absolute error，MAE）和交并比（intersection over union，IoU）衡量。其中，平均绝对误差用于计算预测时间点与真实时间点之间的平均绝对偏差，交并比用于计算预测时间段与真实时间段之间的重合比例。微调前，模型判断偏误时间边界的MAE为283ms，IoU为0.82。利用高级诊断层数据微调后，MAE缩小至156ms，IoU提升至0.91。这说明模型通过学习音素对齐信号，能够将声学异常与偏误事件在时间轴上更准确地对应起来。

在偏误检测方面，本研究采用宏平均F1值（Macro-F1）作为核心评估指标。针对真实场景中偏误类别分布极不均衡的问题（如声调偏误远多于副语言偏误），Macro-F1通过对各类别均等加权的机制，有效防止了整体得分被高频错误主导，从而真实反映模型对长尾偏误的检测能力。微调前，受预训练语言分布影响，模型对留学生特有的声调与韵律偏误不够敏感，各项F1值仅处于0.54至0.72之间。微调后，模型在声母、韵母、声调、韵律及副语言偏误上的F1值分别提升至0.86、0.83、0.89、0.78与0.84。

同时，为验证模型输出的综合可靠性，实验测试了其返回结构化诊断结果的稳定性。微调后，模型输出的JSON格式合规率从87.2%提升至93.4%，这确保了后续分析程序能够自动解析这些数据。此外，为进一步验证模型输出在教学场景中的可用性，本文随机抽取50个测试样本，由3位具有国际中文教育背景的教师进行盲评，并从标注准确性与教学实用性两个维度采用5分制评分。结果显示，微调后模型在标注准确性与教学实用性上的平均得分分别达到4.5分与4.3分。这些结果表明，在本文测试集与实验设置下，利用高级诊断层专家标注数据进行微调后，Qwen2-Audio-7B在时间定位、偏误分类和结构化输出稳定性方面均表现出更好的任务适配能力。该结果初步说明，本文后六层高级诊断标注能够为端到端语音偏误诊断模型提供有效监督信号。

综上，应用1和应用2分别从基础对齐标注和专家诊断标注两个角度验证了本文数据集的可用性。需要说明的是，本文实验定位为代表性应用验证，相关结论仅限于本文所选模型、数据划分和测试集设置。后续研究将进一步引入Wav2Vec2.0、HuBERT及更多传统MDD方法，在统一实验条件下开展更系统的横向比较。

4 数据治理经验与可迁移方法

本研究的数据集建设流程不只适用于国际中文教育。对需要专家标注、时间对齐、多层级标签管理和持续更新的音频数据任务，可复用的环节包括分层标注、人机协同、质量复核、结构化入库和迭代更新。以下结合本研究的数据采集、标注、复核和入库过程进行说明。

第一，先区分基础对齐层和专家诊断层。本文将十层TextGrid标注体系划分为前四层基础标注层和后六层高级诊断层。基础标注层记录句子、汉字、音节和声韵等时间对齐信息，可由ASR、VAD、MFA等工具先行处理；高级诊断层记录声母偏误、韵母偏误、声调偏误、韵律偏误和副语言现象等需要专业判断的标签。这样的分层方式把时间切分和专家诊断分开，减少了专家在基础切分环节的重复劳动，也便于后续按任务需要抽取不同粒度的数据。

第二，采用机器预标注、人工精修、

专家复核的生产流程。机器负责格式转换、降噪、语音活动检测、初步识别和时间对齐等批量处理工作；人工标注者在预标注结果上修正边界并补充高阶标签；专家负责争议样本复核和质量抽检。医学语音、方言资源、儿童语言发展、特殊教育语音评估和口语考试评分等任务，也常需要在自动处理结果上加入专家判断。类似的数据治理问题也出现在语音安全研究中。深度伪造音频生成与鉴伪任务需要记录真实音频与伪造音频的来源、生成方式、声学特征和测试场景，并通过规范标注与跨场景评估保证检测结果可靠[28]。这与本文所采用的样本追溯、分层标注和质量复核流程具有可比性。

第三，质量控制应落到可复核的指标和流程上。对于主观性较强的专家标注任务，只报告数据规模并不足以说明数据质量，还需要交代复核流程、一致性检验、无效样本剔除规则和时间戳冲突检查方法[10]。本文通过交叉复核、标注者间一致性计算、低质量录音剔除和层级间时间冲突检查控制数据噪声，使后续使用者能够判断标注结果的可靠性。

第四，将非结构化标注文件转换为可检索、可训练的数据记录。TextGrid文件便于语音学专家编辑和校验[25]，但不适合直接用于模型训练、批量查询和持续维护。本文使用Parselmouth解析TextGrid层级结构，将起止时间、层级名称、标签内容和作业元数据写入关系型数据库，使数据能够按偏误类型、母语背景、学习水平和任务来源等维度检索与复用。类似思路可用于视频行为标注、医学影像报告对齐和课堂交互数据管理等多模态数据任务。

第五，数据集需要支持后续更新。领域数据集不宜停留在一次性静态资源层面，而应预留新样本采集、模型错误回流、专家复核和版本更新的入口。本研究流程可扩展为教学平台采集、模型辅助诊断、教师修正、数据回流和模型再训练的连续过程，使数据积累、模型适配和教学反馈能够相互校正。

总体来看，本研究的数据集建设经验可归纳为分层标注、专家复核、结构化入库和持续更新四个环节。对于其他依赖专家标注和多层级数据管理的语音数据任务，上述环节可作为流程设计和质量控制的参考。

5 结束语

本文面向国际中文教育中留学生语音偏误诊断数据不足的问题，构建了包含340份真实发音样本的汉语语音偏误数据集。研究制定了融合基础声学对齐与高级专家诊断的十层结构化标注规范，并通过“机器预标注—人工精修—专家复核”的人机协同流程，形成了超过12万条声韵级时间对齐切片和数万条细粒度偏误诊断标签。

为初步验证数据集的应用价值，本文选取非母语口音ASR和端到端语音偏误诊断两个代表性任务进行实验。结果显示，基于基础标注层微调Paraformer-large后，模型在本文测试集上的CER和SER均有所下降；基于高级诊断层微调Qwen2-Audio-7B后，模型在时间定位、偏误分类、JSON合规性和人工评估等方面表现出更好的任务适配能力。上述结果表明，本文数据集能够为国际中文教育场景中的模型微调和智能化发音反馈提供数据基础。

本文仍存在一定局限。实验部分主要用于验证数据集在两个代表性下游任务中的可用性，尚未覆盖Wav2Vec2.0、HuBERT等自监督语音表征模型，也未系

统比较更多传统语音偏误诊断方法。因此，本文结果不应被理解为对所有主流模型的全面横向比较。后续研究将扩大样本规模，丰富母语背景和偏误类型，并在统一数据划分下补充更多模型基线。本文形成的分层标注、人机协同复核、结构化入库和数据回流做法，也可供医学语音、方言资源、儿童语言评估等依赖专家标注和时间对齐的场景参考。

参考文献：

[1] 吴应辉. 国际中文教育发展的新定位——基于2024世界中文大会领导人讲话的思考[J]. 昆明学院学报, 2025, 47(1): 1-8. DOI: 10.14091/j.cnki.kmxyxb.2025.01.001. Wu Y H. New Positioning for the Development of International Chinese Language Education-Reflections Based on the Speeches of Leaders at the 2024 World Chinese Language Conference[J]. Journal of Kunming University, 2025, 47(1): 1-8.

[2] 习近平向2024世界中文大会致贺信[N]. 人民日报, 2024-11-16(02).
Xi Jinping Sends a Congratulatory Letter to the 2024 World Chinese Language Conference[N]. People's Daily, 2024-11-16(02).

[3] JIAO B. Phonetic Learning Strategies for Chinese as a Second Language[J]. Journal of Education and Educational Research, 2024, 8(2): 199-201. DOI: 10.54097/851k6918.

[4] BU H, DU J, NA X, et al. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline[C]// 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, South Korea: IEEE, 2017: 1-5. DOI: 10.1109/ICSDA.2017.8384449.

[5] ZHANG J, ZHANG Z, WANG Y, et al. speechocean762: An Open-Source Non-native English Speech Corpus For Pronunciation Assessment[C]//Proceedings of Interspeech 2021. Brno, Czech Republic: ISCA, 2021: 3710 - 3714. DOI: 10.21437/Interspeech.2021-1259.

[6] WANG W, ZHANG J. Factors predicting human performance in error annotation for non-native speech corpus[J]. Speech Communication, 2023, 149: 38-46. DOI: 10.1016/j.specom.2023.03.001.

[7] FANG Y, PENG J, LI X, et al. Low-Resource Domain Adaptation for Speech LLMs via Text-Only Fine-Tuning[A/OL]. arXiv, 2025. https://arxiv.org/abs/2506.05671. DOI: 10.48550/ARXIV.2506.05671.

[8] PENG L, GAO Y, BAO R, et al. End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning[J]. Applied Sciences, 2023, 13(11): 6793. DOI: 10.3390/app13116793.

[9] CAO W, WANG D, ZHANG J, et al. Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training[C]//Proceedings of Interspeech 2010. Makuhari, Chiba, Japan: ISCA, 2010: 1922-1925. DOI: 10.21437/Interspeech.2010-553.

[10] SYLOLYPAVAN A, SLEEMAN D, WU H, et al. The impact of inconsistent human annotations on AI driven clinical decision making[J]. npj Digital Medicine, 2023, 6: 26. DOI: 10.1038/s41746-023-00773-3.

[11] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the

40th International Conference on Machine Learning. Honolulu, Hawaii, USA: PMLR, 2023: 28492-28518.

[12] ALRASHOUDI N, AL-KHALIFA H, ALOTAIBI Y. Improving mispronunciation detection and diagnosis for non-native learners of the Arabic language [J]. Discover Computing, 2025, 28, Article 1. DOI: 10.1007/s10791-024-09489-8.

[13] LIN Y X, YANG C K, CHEN W C, et al. A Preliminary Exploration with GPT-4o Voice Mode[A/OL]. arXiv, 2025[2026-06-09]. DOI:10.48550/arXiv.2502.09940.

[14] FANG Y H, LO T H, SUNG Y T, et al. Beyond Modality Limitations: A Unified MLLM Approach to Automated Speaking Assessment with Effective Curriculum Learning[C]//2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Honolulu, Hawaii, USA: IEEE, 2025. DOI: 10.48550/arXiv.2508.12591.

[15] MCAULIFFE M, SOCOLOF M, MIHUC S, et al. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi [C]//Proceedings of Interspeech 2017. Stockholm, Sweden: ISCA, 2017: 498-502. DOI: 10.21437/Interspeech. 2017-1386.

[16] WILLIAMS S, FOULKES P, HUGHES V. Analysis of forced aligner performance on L2 English speech[J]. Speech Communication, 2024, 158: 103042. DOI: 10.1016/j.specom.2024.103042.

[17] PARIKH A K, TEJEDOR-GARCIA C, CUCCHIARINI C, et al. Enhancing GOP in CTC-Based Mispronunciation Detection with Phonological Knowledge[C]//Proceedings of Interspeech 2025. Rotterdam, Netherlands: ISCA, 2025: 5068-5072. DOI: 10.21437/Interspeech.2025-829.

[18] KORZEKWA D, LORENZO-TRUEBA J, ZAPOROWSKI S, et al. Mispronunciation Detection in Non-native (L2) English with Uncertainty Modeling[A/OL]. arXiv, 2021[2026-06-09]. DOI:10.48550/arXiv.2101.06396.

[19] WITT S M, YOUNG S J. Phone-level pronunciation scoring and assessment for interactive language learning[J]. Speech Communication, 2000, 30(2-3): 95-108. DOI: 10.1016/S0167-6393(99)00044-8.

[20] FENG Y, FU G, CHEN Q, et al. SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis[C/OL]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 3492-3496. http://dx. doi. org/10.1109/icassp40776.2020.9052975. DOI: 10.1109/ICASSP40776.2020.9052975.

[21] HU E J, SHEN Y, WALLIS P, et al. LoRA: low-rank adaptation of large language models[C/OL]//International Conference on Learning Representations. 2022[2026-06-10]. https://openreview.net/forum?id=nZeVKeeFYf9.

[22] ZHU C, WUMAIER A, WEI D, et al. Pronunciation error detection model based on feature fusion[J]. Speech Communication, 2024, 156: 103009. DOI:10.1016/j.specom.2023.103009.

[23] RABINER L R, SAMBUR M R. An algorithm for determining the endpoints of isolated utterances[J]. The Bell System Technical Journal, 1975, 54(2): 297-315. DOI: 10.1002/j. 1538-7305.1975. tb02840.x.

[24] SAINBURG T, ZOREA A. Domain general noise reduction for time series signals with Noisereduce[J]. Scientific Reports, 2025, 15: 30905. DOI: 10.1038/

s41598-025-13108-x.

[25] JADOUL Y, THOMPSON B, DE BOER B. Introducing Parselmouth: A Python interface to Praat[J]. Journal of Phonetics, 2018, 71: 1-15. DOI: 10.1016/j.wocn.2018.07.001.

[26] 中华人民共和国教育部. 2018年来华留学统计 [EB/OL]. 2019-04-12[2026-06-10]. https://www. moe. gov. cn/jyb_xwfb/gzdt_gzdt/s5987/201904/t20190412_377692. html. Ministry of Education of the People's Republic of China. Statistics on international students in China in 2018[EB/OL]. 2019-04-12[2026-06-10]. https://www. moe. gov. cn/jyb_xwfb/gzdt_gzdt/s5987/201904/t20190412_377692.html.

[27] 中华人民共和国教育部. 教外〔2018〕50号. 教育部关于印发《来华留学生高等教育质量规范（试行）》的通知[S/OL]. 2018-09-03[2026-06-10]. Ministry of Education of the People's Republic of China. Document No. 50 [2018]. Notice of the Ministry of Education on issuing the Quality Standards for Higher Education for International Students in China (Trial)[S/OL]. 2018-09-03 [2026-06-10]. https://www.moe.gov.cn/srcsite/A20/moe_850/201810/t20181012_351302.html. https://www. moe. gov. cn/srcsite/A20/moe_850/201810/t20181012_351302.html.

[28] 曾志平, 张旭龙, 瞿晓阳, 等. 深度伪造音频生成与鉴伪技术综述[J]. 大数据, 2025, 11(5): 130-151. DOI: 10.11959/j. issn. 2096-0271.2025064.Zeng Z P, Zhang X L, Qu X Y, et al. Survey of deep fake audio generation and detection techniques[J]. BIG DATA RESEARCH, 2025, 11(5): 130-151.

作者简介

孙佳杰（2001–），男，硕士研究生，华东师范大学数据科学与工程学院，主要研究方向为智能教育。

陈熙之（2000–），男，硕士，华东师范大学数据科学与工程学院，主要研究方向为大模型应用。

凌锋（1976–），男，博士，华东师范大学国际汉语文化学院，副教授，主要研究方向为实验语音学、汉语方言学。

袁丹（1982–），女，博士，华东师范大学国际汉语文化学院，副教授，主要研究方向为实验语音学、方言学、社会语言学以及二语习得。

兰韵诗（1994–），女，博士，华东师范大学数据科学与工程学院，副教授，主要研究方向为知识图谱，智能问答以及其他与自然语言处理相关的任务。

王晔（1977–），男，博士，华东师范大学数据科学与工程学院，专任研究员，主要研究方向为Web数据管理，海量数据挖掘，分布式系统。

收稿日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: