计算机与智能教育技术

2026年11月 第1卷第4期

# 基于人工智能的教学语音分析与教育应用研究综述

张玉涵，王 超

徐州鑫论源科技发展有限公司，江苏徐州，221116

**摘　要** | 教学录音可以用于测量停顿和基频，也能经自动转录用于识别课堂提问和师生话语行为。本文采用主题式叙述综述，梳理相关研究从声学测量、课堂观察编码到自动话语识别与形成性反馈的发展，比较统计声学特征、自监督语音表征、转录分析及多模态融合的适用条件。中文课堂自动编码、师范生多模态实训和国外课堂反馈研究表明，自动分析可辅助定位和复核具体片段，但测量误差、话语识别性能和教学反馈成效需要分别验证。结合师范生微格教学，本文讨论如何组织片段回听、人工复核和下一轮试讲，并提出按课堂或教师分组进行独立评估、以盲评检验后续表现及迁移的建议。

**关键词** | 人工智能；教学语音；声学表征；课堂话语；形成性反馈

Copyright © 2026 by the author(s).

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

https://creativecommons.org/licenses/by-nc/4.0/

## 1 引言

### 1.1 研究问题与应用范围

教师在一段讲解中停顿两秒，可能是在等待学生作答，也可能在翻找教学材料。录音可以帮助定位停顿，判断其教学作用还需要结合停顿前后的话语和课堂活动。人工智能可以分别处理这两类问题：声学算法量化录音中的可观测信号，语音识别和自然语言处理可从转录中寻找提问、追问或回应学生观点等事件。两者使用的数据和核验方式不同。

本文讨论师范生微格教学、教师课堂回看和讲课竞赛录音。师范生通过短时试讲练习特定技能，教师回看课堂时需要还原师生连续互动，竞赛录音通常用于复盘一段特定的讲授。即使分析的是同一种声音信号，三个场景所用的标签和核验条件也会不同。本文据此考察录音能提供哪些测量结果、模型能定位哪些言语行为，以及这些信息如何用于改进下一次教学。

### 1.2 文献范围与证据组织

本文采用主题式叙述综述。初稿的检索与出版信息核查在2026年11月9日至10日进行，11月11日补核在线出版条目。2026年11月17日补检国内外课堂编码、自动转录误差、师范生微格教学及多模态实

作者简介：张玉涵，徐州鑫论源科技发展有限公司技术工程师，主要从事智能化技术应用、数据处理与分析、技术方案实施及优化等工作，研究方向为智能语音分析与教育技术应用；王超，徐州鑫论源科技发展有限公司技术总监，主要负责企业技术研发规划、研发团队管理、智能算法开发及科技成果转化等工作，研究方向为人工智能、机器学习与智能评价技术。

文章引用：张玉涵, 王超. 基于人工智能的教学语音分析与教育应用研究综述[J]. 计算机与智能教育技术, 2026, 1(4): 156-165.

https://doi.org/10.68070/2vvawp06

训研究。补检用期刊官网、期刊在中国知网的门户、大学机构库、ACL Anthology和出版方页面核对出处，并从直接相关研究追查原始论文。中文补检词包括“课堂话语 自动编码”“教学基本技能 智能实训”和“师范生多模态教学反思”，英文词包括“classroom discourse”“automatic speech recognition”“talk moves”和“automated feedback”。

文献语种限于中文和英文，纳入与教学录音、师生话语识别或教师反馈直接相关的期刊论文、会议论文与语料论文，同时保留解释测量、建模或验证方法所需的经典研究，不设统一起始年份。先人工筛阅题名和摘要，再查阅可获得的方法、数据及结果信息；预印本与正式版按题名、作者或DOI合并，优先使用正式版。产品宣传、无法追溯的材料和不涉及所述任务的研究不作为应用证据。初稿未保留完整命中清单与各阶段剔除数，因此不报告筛选统计，也不把本综述作为系统综述或荟萃分析。

本文比较各研究的评价对象、输入、标签及验证方式。图1列出了每条技术路线需要的证据。同一段录音可以用于测量停顿、识别提问或提供教学反馈，各用途的有效性仍需分别检验。

| 研究任务 | AI技术路线 | 可复核输出 | 验证依据 |
| --- | --- | --- | --- |
| 声学测量 | 分帧、声学特征<br>或自监督表征 | 停顿边界、基频<br>相对强度 | 参考测量<br>时长误差、无效帧 |
| 言语行为识别 | 说话人区分、转录<br>上下文编码与融合 | 带时间戳的提问<br>和师生话语动作 | 人工编码一致性<br>逐类识别性能 |
| 形成性教学反馈 | 片段检索、指标展示<br>人工更正与讨论 | 可回听依据<br>下一轮练习目标 | 盲评对照<br>后续行为、迁移 |
| 历史奖项预测 | 录音表征<br>与分类模型 | 与既有奖项<br>对应的类别 | 独立赛事或分组验证<br>分项评分另行验证 |

图 1 教学语音 AI 分析路线与证据层次

## 2 教学语音分析的研究对象与证据层次

### 2.1 声学参数与讲授情境

Bottalico和Astolfi在六所学校采集了40名小学教师的发声资料。在传统课堂中，背景噪声升高伴随教师发声声级和基频的变化[1]。同一教师的声学指标因而可能随环境变化。录音文件中的相对满幅强度还受麦克风增益、距离和压缩影响，其含义有别于经过校准的声压级。跨赛事按录音响度排序时，设备和录音条件的差别可能被误判为讲授表达的差别。

Thorson和Nesbitt录制了五个场所10名学前教师的课堂语音，并结合话语长度、提问和话轮转换分析基频及其轮廓[2]。按课堂活动和话语事件分析，可以查到音高变化出现在哪个环节。整段录音的均值和分位数便于批量比较，但会把导入、讲解、等待和互动合在一起。判断一次停顿是为了给学生留出思考时间，还是因为讲授中断，需要回到它出现的位置和上下文。

Zhou和Gu考察了韵律、语言及手势与学生对二语词汇教学片段的喜好之间的关系[3]。研究记录的是特定参与者的偏好。竞赛奖项和学习成效各有评价规则，这些关联能否迁移仍需检验。例如，语速与学生喜好的关系可能随学生群体、学科或任务变化，单凭声学统计难以确定通用阈值。

### 2.2 从语音事件识别到教学表现评价

Blanchard等利用自动语音识别找出真实课堂中的教师提问，再用人工标注核验[4]。Suresh等发布的TalkMoves数据以句子为单位，标注数学课中师生的讨论动作[5]。两种做法都把标签对应到录音或转录中的具体位置，研究者可以核对模型找得准不准，教师也能回看那一段互动。

国内研究中，马苗等在弗兰德斯互动分析的基础上设置课堂讲授、提问互动、不当言语、课堂管理和教师评价五类标签，并在来自27名师范生的课堂录像上研究言语行为识别[6]。其输入涉及转录与文本关系，任务单位和评估标准均有别于整段竞赛奖项预测。这类研究对竞赛音频分析的启示在于，应先根据教学目标定义行为单位与标注规则，再决定采用声学、文本或上下文特征。对于缺少学生真实互动的展示型竞赛录像，同一话语行为的发生频率也不能直接与常态课堂类比。


https://doi.org/10.68070/2vvawp06

Demszky和Hill发布了小学数学课堂的转录数据，并考察话轮标注与课堂观察评分、学生学习结果之间的关联[7]。课堂观察评分和学习结果可以为话语指标提供外部参照，研究结果仍限于相关性。对讲课竞赛而言，如果有评委的分项分数，就可以分别考察录音特征与语言表达、教学设计的关系。只有奖项等级时，模型只能学习录音与赛事奖项之间的统计关系，难以区分各评分维度的贡献。

表1概括了各类研究的分析单位、效标及其能支持的判断。特别是，行为检测依赖人工事件标签，反馈成效则需要观察教师或师范生的后续表现。

表 1 教学语音研究中常见的证据类型

| 研究类型 | 分析单位 | 效标或观察对象 | 可支持的判断 |
|---|---|---|---|
| 形成性教学反馈 | 片段、教师或训练组 | 独立评分、后续试讲 | 评估反馈是否改变目标行为 |
| 教师发声与课堂声学 | 教师、课时、工作日 | 发声与环境声学参数 | 观察声学变化及相关条件 |
| 课堂话语识别 | 话语、句子、话轮 | 提问或讨论动作的人工标注 | 判断具体言语行为能否被识别 |
| 话语与外部效标关联 | 话轮、整节课 | 课堂观察评分、学生结果 | 分析特定指标与外部效标的关系 |
| 竞赛奖项预测 | 参赛录像或音轨 | 既有竞赛奖项 | 评估特定样本和规则下的分类性能 |

### 2.3 从观察编码到自动反馈的研究演变

较早的研究已经显示出两种方向。Blanchard等[4]把任务限定为检测教师提问，每一个结果都可以回到原录音核对。孙众等[8]则从整个教学分析流程出发，讨论如何用人工智能从课堂记录中提取行为和互动证据。随着研究关注点从行为是否出现转向它在对话中起什么作用，前后话轮也进入了分析。宋宇等[9]以音视频转录文本为材料，使用卷积神经网络和双向长短期记忆网络提取课堂对话中的思维相关特征。自动标注可以用于定位和比较互动片段，学生是否形成了相应的思维能力还要通过独立任务检验。

话语功能的细分提高了反馈的解释力，也使研究间的指标更难直接合并。TalkMoves的567份中小学数学课堂转录以人工转录、说话人切分和句级标注为基础[5]。其编码以促进数学讨论的行为为目标。胡艺龄等[10]则针对高中信息技术优质课，在语义内容与语用功能两个维度建立自动编码，不同维度报告的准确率分别为72.47%和81.35%。这两个研究的课程、语种、编码对象与评价条件不同，适合比较框架能回答什么问题，不宜用准确率比较谁更适合所有课堂。比如“你是怎么想到的”可以在一种框架中编码为追问理由，在另一种框架中与知识领域标签并列；需要预先规定连续追问按句、按话轮还是按互动事件计数。

叶鹏等[11]使用Whisper和GPT-4依次完成转录、清洗和分类，并报告了“启动—回应—评价”自动编码与人工编码的一致性。这项结果适用于该研究使用的课堂材料和编码规则。识别开放式提问、同伴互评，或判断某句话由谁说出，都需要另行检验。以“教师评价”为例，复述学生答案、指出错误和继续追问都可能归入这一类。如果反馈要区分这三种行为，标注手册就应给出可辨认的例子，并说明边界。

## 3 教学语音的人工智能分析方法

### 3.1 手工统计特征与测量口径

GeMAPS定义了一组声学参数，并通过统一的提取方式减少研究间的实现差异[12]。常用参数包括基频、能量、频谱和时序。特征集规模较小时，计算方法更容易记录，也便于逐项检查特征是否携带设备、片段长度或说话人信息。只写“使用音高与停顿”仍可能无法复现，报告中还应交代单位、算法、参数和聚合窗口。

McFee等介绍了librosa的音频处理功能[13]。处理讲课录音时，可以先统一采样率和声道，再分帧提取特征。Mauch和Dixon提出的pYIN利用概率阈值分布处理基频候选[14]。用这一方法计算整段均值和波动时，结果会受搜索范围、帧长、无声帧判定及异常值处理影响。比较不同说话人时，可同时保留绝对音高和相对个人基线的变化。

录音切分方式会改变停顿次数和发声比例。按能量设阈值容易实现，但其他说话人发出的声音、粉笔声或设备声也可能被判为非静音。按说话人或语音活动切分需要额外标注和核验。报告结果时，应交代录音首尾静音、段内停顿和等待学生回答各自如何计入。除了整段统计，也可以按固定窗口或教学事件核对差异集中在哪些片段。

### 3.2 自监督语音表征

Baevski等提出的wav2vec 2.0利用掩蔽和对比学习，从语音信号中学习包含上下文的信息[15]。相比整段统计量，这种表征可以保留更多局部变化。原研究主要检验语音识别性能。用于讲课奖项分类时，仍要决定如何切分长录音、汇总片段嵌入，以及哪些参数需要微调。样本较少时，可以冻结编码器、只训练简单的分类头作为对照，并固定数据划分和特征口径。

### 3.3 自动转录与多模态分析

Radford等以大规模弱监督语料训练Whisper，并比较了不同语音识别条件下的表现[16]。在教育应用中，自动转录把录音转换为可查找的文字，但语音识别与课堂行为识别应分别评估。例如，整体字错率较低时，术语、短问句和多人重叠处仍可能误识别；这些片段恰好会影响对提问或回应的判断。评估时可以同时报告中文字符错误率和目标行为的精确率、召回率，并用人工转录与自动转录作对照，分辨错误来自语音识别还是下游模型。

语义与视觉信息能够补充音频单独缺失的证据。转录可以区分问题类型、术语与解释内容，但识别误差可能集中在口音、术语或多人重叠场景。图像可以定位板书、手势及教学媒介。将三种输入放到同一时间轴后，才能比较每种模态的独立增益与缺失时的表现。竞赛录像中的剪辑、配乐、机位与服装也可能成为非教学线索。多模态融合需要在外部赛事上检查这些线索是否造成虚高性能。

## 4 课堂话语识别与多模态分析：可比性与迁移

### 4.1 上下文、说话人和识别误差

转录回答了“说了什么”，要给课堂言语行为编码，还得知道是谁说的、在回应谁，以及前后发生了什么。只看“对，还有吗”这一句，模型可能将它统一判为评价。结合学生上一句的回答，才有依据区分确认答案与邀请补充。说话人角色既应作为分析输入，也需要单独核验。人工转录语料上的分类效果主要反映编码器的能力，自动处理还会遇到短回应漏转、学生话语误归教师等问题。分析小组讨论时，宜分别记录遗漏、归属错误和下游误标。

Cao等[17]在真实课堂小组协作录音中比较了两类转录系统及其对下游任务的影响。同样的录音经自动转录后，句级语义解析、在任务状态识别和讨论动作识别受损程度不同。这提示研究者除报告总体字词错误率外，还应检查与目标行为有关的词、话轮和说话人，并用同一编码器比较人工转录与自动转录下的性能。对于无法判别的重叠语音，宜把拒判比例与分类表现同时报告；只在清晰片段上计算指标，会使教师误以为整堂课均有同等覆盖。

在另一项研究中，Cao等[18]从课堂数学话语数据迁移到数学辅导数据SAGA22，比较了上下文长度、说话人身份与训练策略。课堂数据的补充训练对目标场景有帮助，但研究使用了目标场景的标注及验证。将课堂语料用于微格教学时同样需要抽样复核；尤其不能把缺少真实学生的模拟讲授视为互动识别的等价场景。测试切分宜至少按课堂或教师分组，否则同一教师的惯用表达、相邻话轮和录音条件可能同时出现在两侧。

### 4.2 多模态证据与人工复核

多模态研究的关键在于每种输入补足哪一项证据。例如“看这里，为什么会变成这样”提供了一个提问位置，但板书或课件才能确定“这里”指什么；学生是否回答则需要后续话轮，可能还要查看视频。若任务只是定位语速变化，加入图像未必改善声学测量。若任务是解释提问的指向，音频、转录和画面须先按时间对齐，然后在同一数据划分上对比单模态和融合输入，并报告缺失模态或时间偏移时的表现。整节课的总分不能说明模型是在找相应行为，还是只识别了画面布置、设备或录制规范。

生成式模型可以依据编码手册给出候选标签和理由。用于反馈时，每一条理由都应链接到实际片段或转录句，缺乏上下文应标记待确认。可同时记录提示词、模型版本、解码设置与更正状态，用同一验证集重复运行以检查输出稳定性。人为修订也需保留依据：例如“尚未识别到回答”可能是学生音频太弱，不应自动写成“学生没有作答”。这类设计把识别失败与课堂观察结论区分开来，便于学科导师处理边界片段。

## 5 教学应用的实证进展

### 5.1 从课后观察到有指导的实训

孙硕等[19]提出包含四个维度、18项指标的师范生教学基本技能观测框架，并将智能分析用于实训。与普通课堂话语分类相比，这项研究把语言表达放在教学基本技能中考察，并将分析结果用于练习。比较此类实训系统时，需要看清它们评价了哪些技能、怎样采集资料、如何安排练习，以及用什么指标判断变化。实训中的使用记录能够说明工具怎样进入练习，各项自动指标的准确性以及真实课堂中的长期表现仍需分别验证。

吴忠明等[20]从试讲视频提取多模态教学技能数


https://doi.org/10.68070/2vvawp06

据，以反思仪表盘组织证据，并对17名师范生使用前后的教学反思文本进行比较。研究观察到反思文本在若干维度的广度和深度有所改善。观察对象是反思文本，训练者能否在后续课堂稳定调整具体教学行为仍是另一个问题。穆肃等[21]将AI诊断和教师指导组合起来，考察39名师范生的教学基本技能及效能感，给出了对照自动诊断和人工支持如何配合的应用实例。两项工作提示，反馈工具评价不宜只看是否有人查看报告，还需要预先说明目标究竟是提高反思质量、改变试讲行为，还是帮助师范生独立发现问题。

### 5.2 自动反馈能改变哪些教学行为

Demszky等在大型在线编程课中检验M-Powering Teachers的自动反馈，研究用随机邮件提醒促进教师查看反馈，发现教师更常接续和利用学生的观点[22]。在线授课录音和教学组织有其特点，因而该结果支持特定环境下的反馈干预。在线下课堂的随机试验中，Demszky等发现自动反馈增加了数学和科学教师用来追问学生思路的聚焦性问题，未观察到其他被测教学实践的改善[23]。两项研究针对的是可以定义和观察的教学行为，结果不宜扩展为一般教学质量提升。

Malamut等访谈了9名教学指导者，并分析25段指导对话[24]。指导者用系统标出的片段追问、举例或比较，也会质疑系统对学生话语的识别。研究使用人工转录作为输入，因此可以观察机器反馈如何进入指导对话，自动转录在类似教室里的可用性仍需检验。结合前述试验，判断识别是否准确要核对录音和标注，教师是否使用反馈要看指导或练习记录，行为是否改变则要观察后续教学。

### 5.3 对照不同应用研究的结果

国内智能实训与国外课堂话语反馈研究可以互作参照，但对象并不相同。微格课堂时间短，任务目标可事先约定，适于对照同一技能的多轮表现；真实课堂中的教师面对不同班级、教材与学生回答，同一行为出现多少次也可能取决于可用机会。比较干预效果时，宜把计数换成有定义的机会比例，例如在存在学生回应的提问中，有多少次教师追问了理由，并同时报告分子和分母。这样既能避免把教师发言频率当作互动质量，也能识别低样本量导致的大幅波动。

附录A保留一个小样本竞赛分类案例，用来说明开发评估的选择偏差。该项目预测历史奖项，样本和任务与上述公开课堂研究不同。即便开发集准确率较高，也不能据此推断真实课堂评价或微格教学训练有效。对于共享教师、课次或赛事来源的音频，应先按来源划分，再在训练侧拟合标准化、特征筛选和分类模型；独立测试只用于最终评估。确需在小样本中比较多个配置时，嵌套交叉验证可把内层选择与外层评估分开[25][26]，但不能代替对其他学校或录制条件的外部验证。

## 6 微格教学反馈的设计与评价

### 6.1 用片段组织反馈

在微格教学中，可让师范生先选一个待改进的行为，例如“提问后是否给学生留出作答时间”。系统给出片段位置和计算依据，教师回听后修正误识别，导师再根据问题类型、学生回应和教学目标讨论是否需要调整。对于没有真实学生互动的模拟讲授，只呈现可核对的节奏和表达信息，不把安静时长解释为学生思考。保存和共享真实课堂录音前，应明确授权、可访问范围和保留期限；报告中优先用不含身份信息的片段标记。

微格教学反馈界面可按录音、片段和修订记录组织。录音层显示任务目标、课题、日期与处理版本；片段层显示起止时间、播放入口、说话人及其待确认状态、转录上下文、停顿时长和基频变化等实际测得指标，并提供单位、计算窗口和异常提示。声学指标与行为标签应分别呈现；无法可靠识别说话人时，不展示师生话语比例。人工更正需保留原输出、修改内容、修改人和时间，区分“未复核”“已确认”“已修正”和“无法判断”，以免将系统输出自动当成人工结论。

师范生先选定本次训练目标，再上传获授权的录音。回听系统标出的片段后，可以修正转录和标签，记录这样判断的理由，并确定下一次试讲要调整的地方。导师查看已复核片段后给出建议，师范生在相同任务要求下再次试讲，以便比较前后变化。上述流程及界面字段仍需在微格课堂试用和验证。

### 6.2 分别检验测量、识别和训练成效

声学测量的验证应使用同一录音的人工标记或独立参考实现，报告停顿边界的时间误差、持续时长的平均绝对误差，以及基频估计中的无效帧比例。比较前须统一采样率、帧长、阈值和聚合规则，并分别检查安静、噪声和远距离录音；不能把某一条件下的误差直接推广到其他设备。若只比较教师自身前后表现，还需固定录音位置及增益，避免把采集变化误认为训练效果。

行为识别应由至少两名经过编码培训的标注者独立标注一部分资料，说明分歧处理规则并报告一致性；随后在未参与调参的参赛者或课堂上计算各类精确率、召回率、宏平均F1和混淆矩阵。将人工

转录与自动转录送入同一个下游模型，可估计转录误差的影响。对少见行为应同时报告支持样本数，对无法识别的重叠话语保留拒判，不以删去难例后的准确率代表完整课堂性能。

教学反馈的对照研究可比较“常规回看”与“回看加自动反馈”，使两组拥有相近的练习时间和导师支持，并按班级或指导组处理可能的相互影响。主要指标可预先设为独立评分者对下一次试讲中目标行为的评分，次要指标包括证据定位耗时、错误发现率、反馈采纳情况和学生解释修改理由的质量。评分者宜不知道分组，报告前后变化、组间差异及不确定区间，并增加延迟或新课题任务检验迁移。工具满意度和点击次数只能说明使用体验，不能单独证明教学能力改善。

## 7 结语

已有研究在中文课堂自动编码、师范生多模态实训和真实课堂反馈中给出了具体做法。三类任务所用的数据、标签和效标不同：停顿和基频应当能够复算，话语标签应当能够回到原片段核对，反馈是否有效则要看教师接下来怎样做。微格教学研究可以从一项明确的行为入手，用模型定位片段，经人工复核后给出练习建议。之后用独立评分判断行为有无变化，再安排延迟任务，检验变化能否迁移。

## 参考文献

[1] BOTTALICO P, ASTOLFI A. Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms[J]. Journal of the Acoustical Society of America, 2012, 131(4): 2817-2827.

[2] THORSON J, NESBITT K. Melodies of learning: A prosodic analysis of preschool teachers’ language patterns in the classroom[C]//Proceedings of Speech Prosody 2024. 2024: 180-184.

[3] ZHOU J, GU Y. Unraveling students’ liking of teachers: The impact of multimodal cues during L2 English vocabulary teaching[C]//Proceedings of Speech Prosody 2024. 2024: 881-885.

[4] BLANCHARD N, DONNELLY P, OLNEY A M, et al. Identifying teacher questions using automatic speech recognition in classrooms[C]//Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles: Association for Computational Linguistics, 2016: 191-201.

[5] SURESH A, JACOBS J, HARTY C, et al. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves[C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2022: 4654-4662. https://aclanthology.org/2022.lrec-1.497/.

[6] 马苗, 康璐, 李雨桐, 等. 数据驱动的教师言语行为识别新方法[J]. 武汉大学学报（理学版）, 2025, 71(5): 643-651.

[7] DEMSZKY D, HILL H. The NCTE transcripts: A dataset of elementary math classroom transcripts[C]//Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications. Toronto: Association for Computational Linguistics, 2023: 528-538.

[8] 孙众, 吕恺悦, 骆力明, 等. 基于人工智能的课堂教学分析[J]. 中国电化教育, 2020(10): 15-23.

[9] 宋宇, 许昌良, 朱佳, 等. 面向思维培养：基于精准标注技术的智能化课堂教学分析及应用[J]. 华东师范大学学报（教育科学版）, 2023, 41(8): 79-89.

[10] 胡艺龄, 汪沁婵, 吴忭, 等. 语义、语用“双管齐下”：优质信息技术课堂互动话语自动化分析及特征发现[J]. 现代远距离教育, 2025(1): 77-87.

[11] 叶鹏, 肖吴, 何炬, 等. 生成式人工智能在课堂师生对话编码中的应用[J]. 中部高教研究, 2025, 1(2): 73-80.

[12] EYBEN F, SCHERER K R, SCHULLER B W, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[13] MCFEE B, RAFFEL C, LIANG D, et al. librosa: Audio and music signal analysis in Python[C]//Proceedings of the 14th Python in Science Conference. 2015: 18-24.

[14] MAUCH M, DIXON S. pYIN: A fundamental frequency estimator using probabilistic threshold distributions[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing. 2014: 659-663.

[15] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[C]//Advances in Neural Information Processing Systems. 2020, 33: 12449-12460.

[16] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th


https://doi.org/10.68070/2vvawp06

International Conference on Machine Learning. PMLR, 2023, 202: 28492-28518. https://proceedings.mlr.press/v202/radford23a.html.
[17] CAO J, GANESH A, CAI J, et al. A comparative analysis of automatic speech recognition errors in small group classroom discourse[C]//Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization. New York: ACM, 2023: 250-262.
[18] CAO J, SURESH A, JACOBS J, et al. Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse[C]//Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics, 2025: 7671-7684. https://aclanthology.org/2025.coling-main.513/.
[19] 孙硕, 胡小勇, 穆肃, 等. 师范生教学基本技能智能实训模型及应用研究[J]. 电化教育研究, 2024, 45(6): 113-120.
[20] 吴忠明, 刘明, 苏逸飞, 等. 教育数据叙事赋能师范生教学反思的实证研究[J]. 现代教育技术, 2025, 35(8): 57-66.
[21] 穆肃, 张誉尹, 陈孝然. 从智能到智慧：人智协同反馈能让教学基本技能训练提速增效吗？[J]. 现代远距离教育, 2025(3): 45-53.
[22] DEMSZKY D, LIU J, HILL H C, et al. Can automated feedback improve teachers’ uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course[J]. Educational Evaluation and Policy Analysis, 2024, 46(3): 483-505.
[23] DEMSZKY D, LIU J, HILL H C, et al. Automated feedback improves teachers’ questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement[J]. Computers & Education, 2025, 227: 105183.
[24] MALAMUT J, DEMSZKY D, BYWATER C, et al. Facilitating evidence-based instructional coaching with automated feedback on teacher discourse[J/OL]. Journal of Mathematics Teacher Education, 2026（在线发表：2026-10-28）[2026-11-17].
[25] CAWLEY G C, TALBOT N L C. On over-fitting in model selection and subsequent selection bias in performance evaluation[J]. Journal of Machine Learning Research, 2010, 11: 2079-2107. https://www.jmlr.org/papers/v11/cawley10a.html.
[26] VARMA S, SIMON R. Bias in error estimation when using cross-validation for model selection[J]. BMC Bioinformatics, 2006, 7: 91.

# 附录A 竞赛录音分类原型的资料与局限

## A.1 样本背景及开发结果

本附录依据Make_prize_audio_agent交付README、当前features.py和候选方案原图编制。已知训练集106条、原测试集25条，交付标签定义为0＝一等奖、1＝二等奖、2＝三等奖。随附素材含题为“三等奖-第六届全国青年教师教学竞赛决赛”的网页，但没有逐条样本清单或链接可证明它与131条数据的对应关系。因此，全部样本的赛事名称及届次、学科与学段、录音时长分布、各奖项人数或比例均无法核实，亦无法检查参赛者是否跨集合重复。这些缺项限制了对样本代表性和分类误差构成的判断。

表A1汇总交付结果。模型A为PCA（7维）与收缩线性判别分析，模型B为PCA（8维）与随机森林；硬门控按随机森林输出中第一、第二高类别值之差选择分支。回溯材料指出门控参数曾参考原测试集比较，故该集合应视作开发比较数据，不能作为独立外部验证。每个测试样本对应4个百分点；模型A与融合均命中24条也不足以判断是否错在同一样本。未进行重新训练或逐条复算。

表 A1 交付说明记载的历史分类结果

| 交付模型 | 训练集（106条） | 原测试集（25条） | 原测试命中数 |
|---|---|---|---|
| PCA7＋LDA | 49.06% | 96.00% | 24/25 |
| PCA8＋随机森林 | 97.17% | 84.00% | 21/25 |
| 硬门控融合 | 84.91% | 96.00% | 24/25 |

注：样本量及正确率来自项目交付 README；命中数按正确率与原测试样本量换算。

图A1将候选原图的左右面板按同一配置名配对，避免独立排序造成误读。全特征LDA和PCA7后的LDA原测试正确率均为0.960，训练正确率分别为0.519和0.491。高训练正确率也可伴随较低的原测试正确率。图中差异是排查样本划分、候选选择与特征口径的线索；缺少样本级记录时，不能确定差异

原因。

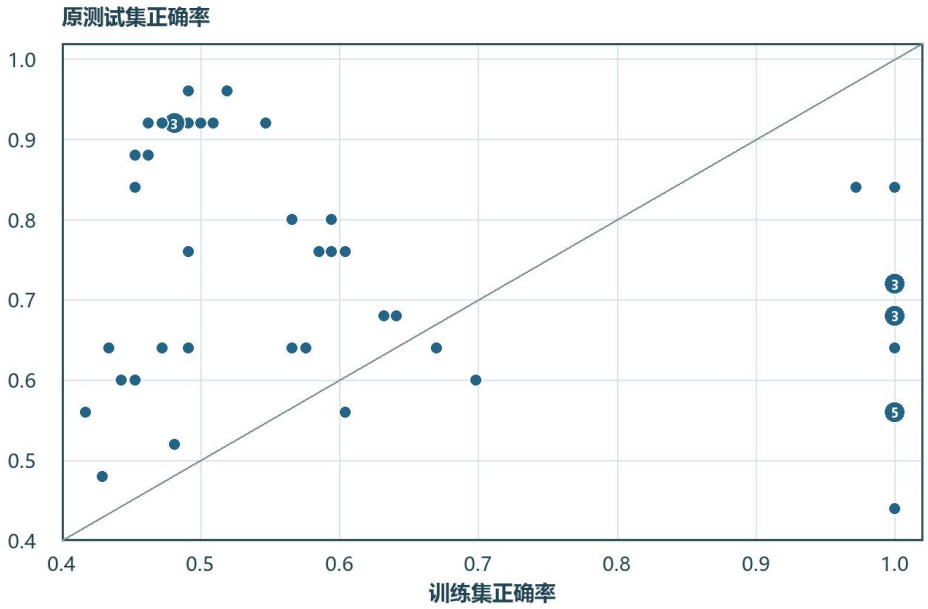



图 A1 50个候选配置的训练与原测试正确率

注：数据手工转录自原始双面板图，按配置名配对。每个点表示一对正确率，重合点标配置数；灰线表示训练与原测试正确率相等。

## A.2 当前音频接口的34列特征定义

表A2按当前重写的提取器输出顺序列出各项。交付README将第17列写为频谱重心50%分位、第33列写为25%分位；当前代码正好相反，需用原训练输入核对后方可确认哪一列应送入模型。接口统一为16 kHz单声道，帧长2048个采样点、帧移512个采样点，相对非静音阈值30 dB，pYIN基频搜索范围80—400 Hz。下表仅说明部署代码；原始训练提取脚本和配对数据缺失，无法确认两者逐列一致。

表 A2 当前音频特征列与定义

| 字段 | 列号 | 计算口径与单位 |
|---|---|---|
| 录音总时长 | 1 | 采样点数除以采样率（s） |
| 静音/停顿总时长 | 2 | 总时长减去非静音段时长之和（s） |
| 连续发声片段数 | 3 | 相对幅度阈值识别的非静音段数（段） |
| 单段发声中位数时长 | 4 | 非静音段时长的中位数（s） |
| 停顿检测阈值 | 5 | 帧长除以采样率（s）；与实际幅度阈值30 dB 区分 |
| 平均停顿时长 | 6 | 相邻非静音段间隔平均值，不计首尾（s） |
| 最短停顿时长 | 7 | 相邻非静音段间隔最小值（s） |
| 较长停顿次数 | 8 | 不短于1.0 s 的段间隔数（次） |
| 有声帧占比 | 9 | 非静音时长除以总时长乘100（%）；与第19列相同 |
| F0标准差 | 10 | 非缺失基频的总体标准差（Hz） |
| F0全局最大值 | 11 | 非缺失基频最大值（Hz） |
| F0主体最大值 | 12 | 非缺失基频第95百分位（Hz） |


https://doi.org/10.68070/2vvawp06

续表 A2

| 字段 | 列号 | 计算口径与单位 |
|---|---|---|
| F0中位相对变化率 | 13 | 滤除无效帧后的相邻基频绝对变化率中位数（%） |
| 语音段 RMS 强度 | 14 | 拼接非静音段样本的均方根幅值转分贝（dBFS） |
| 信噪比 | 15 | 非静音与背景均方根幅值比转分贝，为近似值（dB） |
| 频谱重心均值 | 16 | 逐帧频谱重心均值（Hz） |
| 频谱重心25%分位 | 17 | 逐帧频谱重心第25百分位（Hz） |
| 有效发声时长 | 18 | 非静音段时长之和（s） |
| 发声占比 | 19 | 非静音时长除以总时长乘100（%） |
| 单段发声平均时长 | 20 | 非静音段时长均值（s） |
| 最长连续发声片段 | 21 | 非静音段时长最大值（s） |
| 停顿总次数 | 22 | 段间隔数组元素数；不足两段时占位零也会计入（次） |
| 停顿时长中位数 | 23 | 段间隔时长中位数（s） |
| 最长停顿时长 | 24 | 段间隔时长最大值（s） |
| 较长停顿平均时长 | 25 | 不短于1.0 s 的段间隔平均值（s） |
| F0均值 | 26 | 非缺失基频均值（Hz） |
| F0全局最小值 | 27 | 非缺失基频最小值（Hz） |
| F0主体最小值 | 28 | 非缺失基频第5百分位（Hz） |
| F0平均相对变化率 | 29 | 滤除无效帧后的相邻基频绝对变化率均值（%） |
| 整体 RMS 强度 | 30 | 逐帧均方根幅值取均值后转分贝（dBFS） |
| 背景噪声 RMS | 31 | 按背景样本能量或代码回退值计算（dBFS） |
| 动态范围 | 32 | 第95与第5百分位帧均方根幅值的分贝差；非正时置零（dB） |
| 频谱重心50%分位 | 33 | 逐帧频谱重心第50百分位（Hz） |
| 频谱重心75%分位 | 34 | 逐帧频谱重心第75百分位（Hz） |

注：F0指基频，RMS 指均方根幅值。非静音段按相对幅度阈值划分，包含其他人声或环境声的可能，不能直接解释为教师实际发声。

# Artificial Intelligence for Teaching Speech Analysis and Educational Applications: A Review

Zhang Yuhan, Wang Chao

Xuzhou Xinlunyuan Technology Development Co., Ltd., Xuzhou 221116

**Abstract:** This narrative review traces the development of AI-assisted teaching speech analysis from acoustic measurement and classroom observation coding to discourse recognition and formative feedback. It compares acoustic statistics, self-supervised speech representations, transcript analysis, and multimodal approaches. Studies on Chinese classroom discourse, multimodal training for preservice teachers, and automated feedback in other settings show how these tools can help users locate and review episodes. Measurement error, recognition performance, and educational effects require separate validation. For microteaching, the review discusses timestamped evidence, human verification, and follow-up practice, and proposes grouped evaluation and blinded assessment of subsequent performance and transfer.

**Keywords:** artificial intelligence; teaching speech; acoustic representation; classroom discourse; formative feedback


https://doi.org/10.68070/2vvawp06