UDM 系列在真实生活口吃语音应用中的部署：一项临床评估框架

Eric Zhang*, Li Wei, Sarah Chen, Michael Wang
SSHealth Team, AI for Healthcare Laboratory
ericzhang@sshealthai.com

Abstract

结巴和不流畅的语音检测系统历来面临着准确性和临床可解释性之间的权衡。虽然端到端深度学习模型实现了高性能，但它们的黑盒性质限制了临床上的应用。本文探讨了无约束失语建模（UDM）系列——由伯克利开发的当前最先进的框架，它结合了模块化架构、显式的音素对齐和可解释的输出以实现现实世界的临床部署。通过涉及患者和认证言语语言病理学家（SLPs）的广泛实验，我们证明 UDM 实现了最先进的性能（F1: 0.89±0.04），同时提供了具有临床意义的可解释性分数（4.2/5.0）。我们的部署研究表明有 87% 的临床医生接受率和 34%的诊断时间减少。结果强有力地证明了 UDM 代表了一条向临床环境中人工智能辅助言语治疗的实际路径。

1 介绍

口吃和不流畅的言语几十年来一直是言语病理学和计算言语研究中的一个核心话题。不流畅现象，如重复、延长和阻塞，不仅是语言病理学家（SLPs）的关键诊断标志，也对个体的沟通能力、教育成就和生活质量产生强烈影响。全球范围内的口吃患病率约为 1%，特别是在获得专业医疗服务有限的地区影响尤为严重。

传统检测不流利言语的方法严重依赖于手工制作的声学特征（如：抖动，震颤，音高中断）和流畅度指标（如：每分钟音节数，语速）。虽然这些方法提供了一些可解释性，但它们难以在不同说话者和临床背景下进行泛化。手动设计的特征工程过程通常只捕捉到表面级别的声学属性，忽略了不流利言语模式所具有的复杂时间动态性和上下文依赖关系。

随着深度学习的兴起，端到端（E2E）模型已被广泛用于自动检测不流畅现象。这些方法通常直接在音频波形或频谱图上操作，使用 CNNs、RNNs 或 Transformers 来分类不流畅行为。尽管此类模型展示了提高的原始准确性，但它们存在三个基本限制：

1. **缺乏可解释性：**黑盒架构没有为其预测提供透明的推理，这使得临床医生不愿意在敏感的医疗场景中采用它们。最近的一项调查显示，78% 的言语语言病理学家不会信任没有明确解释的人工智能系统。

*Corresponding author

Preprint.

2. **有限可控性：**端到端模型倾向于捕捉全局相关性，但很难适应不同年龄段和严重程度的多样化的不流利模式。
3. **部署差距：**临床环境需要能够不仅提供预测，还能提供中间解释、错误分析和人工验证的白盒模型。

为了解决这些挑战，我们分析了**无约束失 fluency 模型（UDM）**系列 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] 的框架。与严格定义不流利类型或依赖僵化特征集的受限模型不同，UDM 采用了灵活、模块化的设计，可以在不预先强加严格界限的情况下表示各种各样的不流利行为。

SSHealth 团队专注于开发此类模型，以改善中国患者的生活质量，在中国，获得认证的言语语言病理学家极为有限。经过广泛的文献回顾和初步临床试验，我们确定 UDM 系列是平衡准确性、可控性和可解释性最有前景的方法之一。

1.1 贡献

本文做出了以下关键贡献：

- 我们首次全面评估了 UDM 框架在实际部署场景中的临床效果。
- 我们引入了专门为临床口吃检测系统设计的新评估指标，包括可解释性分数、适应性措施和实时性能指标。
- 我们提供了 UDM 在不同年龄组和言语流利度类型中的详细性能分析，展示了其临床有效性。
- 我们在一个大型儿科医院环境中呈现了对临床医生接受度、诊断准确性及患者结果的定量分析。

2 相关工作

2.1 传统口吃检测方法

早期的计算方法在检测不流畅性方面主要基于声学分析和规则系统。这些方法通常实现了适度的性能（F1 分数约为 0.65-0.72），但通过依赖于语言驱动特征提供了清晰的可解释性。

统计方法使用了隐马尔可夫模型（HMMs）和高斯混合模型（GMMs），以捕捉不流畅语音中的时间依赖关系。虽然这些方法改进了基于规则的系统，但仍需要大量的手工特征工程，并且难以处理说话人变异性。

2.2 深度学习方法

深度神经网络的引入标志着口吃检测研究的一个重要转变。基于卷积神经网络的方法在基准数据集上实现了 0.81 的 F1 分数，这比传统方法有了显著的进步。

基于变压器的模型在口吃检测中的序列建模方面显示出前景，实现了在多个基准数据集上的最先进性能。然而，这些模型在其决策过程方面仍然很大程度上不透明。

2.3 临床需求与部署挑战

尽管技术有所进步，研究系统与临床应用之间的差距仍然显著。临床采纳的关键要求包括：

- **透明性：**临床医生必须理解决策是如何作出的

- **可靠性：** 系统必须在不同的患者群体中保持一致的性能
- **适应性：** 模型应适应不同的年龄和严重程度级别
- **集成：** 工具必须适应现有的临床工作流程

我们的 UDM 框架通过其模块化、可解释的架构直接满足这些需求。

3 方法

UDM 框架遵循一个模块化且可解释的流水线，结合了序列学习和基于显式对齐的推理。该架构由五个主要组件协同工作，以提供准确的检测和具有临床意义的解释。

3.1 多尺度特征提取

原始语音信号首先被转换成多尺度的声学表示，这些表示捕捉了精细的发音动态和更广泛的韵律线索：

$$\boldsymbol{F}_{mel} = \text{MelSpectrogram}(\boldsymbol{x}, n_{fft} = 2048, hop = 256) \tag{1}$$

$$\boldsymbol{F}_{pitch} = \text{PitchTracker}(\boldsymbol{x}, f_{min} = 75, f_{max} = 500) \tag{2}$$

$$\boldsymbol{F}_{energy} = \text{EnergyContour}(\boldsymbol{x}, win\_size = 1024) \tag{3}$$

$$\boldsymbol{F}_{mfcc} = \text{MFCC}(\boldsymbol{F}_{mel}, n_{coef} = 13) \tag{4}$$

$$\boldsymbol{F}_{combined} = \text{Concat}([\boldsymbol{F}_{mel}, \boldsymbol{F}_{pitch}, \boldsymbol{F}_{energy}, \boldsymbol{F}_{mfcc}]) \tag{5}$$

3.2 音素对齐模块

UDM 中的一个关键创新是显式的音素对齐阶段，它提供了可解释的中间表示形式：

$$\boldsymbol{P} = \text{PhonemeEncoder}(\boldsymbol{F}_{combined}) \tag{6}$$

$$\boldsymbol{\alpha} = \text{CTCAlignment}(\boldsymbol{P}, \boldsymbol{T}_{expected}) \tag{7}$$

$$\boldsymbol{A} = \text{AttentionRefinement}(\boldsymbol{\alpha}, \boldsymbol{P}) \tag{8}$$

对齐模块明确跟踪四种音素级别的错误：

- **插入：** 表示重复的额外音素
- **删除：** 不完整单词中缺失的音素
- **代换：** 发音扭曲在块中
- **延长：** 扩展时长对齐

3.3 时间模式分析

时间分析模块捕捉跨越多个时间尺度的动态模式：

$$\boldsymbol{H}_{local} = \text{LocalLSTM}(\boldsymbol{A}, hidden\_size = 256) \tag{9}$$

$$\boldsymbol{H}_{global} = \text{GlobalTransformer}(\boldsymbol{A}, n\_heads = 8, n\_layers = 6) \tag{10}$$

$$\boldsymbol{H}_{multi} = \text{FusionLayer}([\boldsymbol{H}_{local}, \boldsymbol{H}_{global}]) \tag{11}$$

3.4 无约束口吃分类器

UDM 的核心是一个分类模块，该模块操作对齐的音素段。

$$\boldsymbol{C}_{canonical} = \text{CanonicalClassifier}(\boldsymbol{H}_{multi}) \tag{12}$$

$$\boldsymbol{C}_{open} = \text{OpenSetClassifier}(\boldsymbol{H}_{multi}) \tag{13}$$

$$\boldsymbol{P}_{final} = \text{WeightedCombination}([\boldsymbol{C}_{canonical}, \boldsymbol{C}_{open}]) \tag{14}$$

规范分类器处理定义良好的类别：

- 发音重复（例如，“b-b-b-球”）
- 音节重复（例如，“ba-ba-球”）
- 单词重复（例如，“the-the-the 球”）
- 延拓（例如，“baaaall”）
- 块（无声停顿伴有构音紧张）

3.5 可解释性特征

UDM 输出专门设计用于临床医生验证：

- **视觉对齐图：** 时间对齐的可视化显示异常时间或音素错误
- **特征归因：** 基于梯度的归因分数，显示哪些特征对预测有贡献
- **置信分数：** 临床决策中精确校准的不确定性估计
- **阈值控制：** 不同诊断目标的可调灵敏度阈值

4 实验

4.1 数据集

我们使用北京儿童医院的临床数据进行了全面评估。表 1 概述了数据集的特点。

表 1: 北京临床数据集特征

特征特性	值
Total Speakers	507
Total Hours	78.9
Age Range	4-67 years
Language	Mandarin Chinese
Annotation Level	Multi-level (frame, phoneme, word)
Recording Environment	Clinical setting
SLP Annotators	4 certified professionals
Inter-annotator Agreement	0.87ś0.05

该数据集代表了最大规模的临床注释中文言语不流畅数据集合，记录于常规临床评估过程中，并获得了知情同意和机构审查委员会的批准。

4.2 基线模型

我们将 UDM 与最先进的方法进行了比较：

4.2.1 端到端模型

- **CNN-RNN 混合模型：** 带有 LSTM 序列建模的卷积特征提取模块
- **Transformer 端到端：** 自监督变换器针对不流畅检测的微调
- **Wav2Vec2-口吃：** 用于检测口吃的微调 Wav2Vec2 模型

4.2.2 传统方法

- **SVM-声学：** 带手工设计特征的支持向量机
- **随机森林：** 具有时域和谱特征的集成方法
- **HMM-高斯混合模型：** 高斯发射的隐马尔可夫模型

4.3 评估指标

我们开发了全面的评估指标包括：

- **检测性能：** 精确率, 召回率, F1 分数, 平衡准确性
- **对齐错误率（AER）：** 音素到帧映射的质量
- **可解释性得分：** 临床医生评定的有用性（1-5 分量表）
- **实时因子（RTF）：** 处理时间相对于音频时长的比例
- **临床一致性和** 与金标准诊断的科恩卡帕值

4.4 结果

4.4.1 整体性能

表 2 提供了所有模型的综合比较。

表 2: 北京临床数据集上的性能比较

模型	F1 分数	精度	回忆	AER (%)	插值。得分
CNN-RNN Hybrid	0.82±0.06	0.79±0.07	0.85±0.05	23.4±3.2	2.1±0.4
Transformer-E2E	0.85±0.04	0.83±0.05	0.87±0.04	19.7±2.8	2.3±0.3
Wav2Vec2-Stutter	0.87±0.03	0.86±0.04	0.88±0.03	18.1±2.1	2.4±0.5
SVM-Acoustic	0.73±0.08	0.71±0.09	0.75±0.07	31.2±4.1	3.8±0.6
Random Forest	0.75±0.07	0.74±0.08	0.76±0.06	29.8±3.7	3.9±0.5
HMM-GMM	0.69±0.09	0.67±0.10	0.72±0.08	35.6±4.8	3.2±0.7
UDM（我们的）	**0.89±0.04**	**0.88±0.04**	**0.90±0.03**	**15.3±1.8**	**4.2±0.3**

UDM 在所有指标上均达到最高性能，F1 分数比最佳基线高出 2-4%，同时保持出色的可解释性。

4.4.2 年龄组分析

表 3 显示了 UDM 在不同发展阶段的表现。

表 3: 不同年龄组的 UDM 性能

年龄组	N	F1 分数	精度	回忆
Early Childhood (3-6)	89	0.86ś0.06	0.84ś0.07	0.88ś0.05
School Age (7-12)	156	0.89ś0.04	0.88ś0.04	0.90ś0.04
Adolescent (13-18)	134	0.91ś0.03	0.90ś0.04	0.92ś0.03
Young Adult (19-30)	78	0.90ś0.04	0.89ś0.04	0.91ś0.03
Middle Age (31-50)	39	0.89ś0.04	0.88ś0.05	0.90ś0.04
Older Adult (51+)	11	0.87ś0.05	0.86ś0.06	0.88ś0.05
总体	**507**	**0.89ś0.04**	**0.88ś0.05**	**0.90ś0.04**

表现保持在各年龄段的一致性，青少年显示出最高的准确性。

4.4.3 不流畅类型分析

表 4 提供了按口吃类别详细分解的数据。

表 4: 按口吃类型分解的性能表现

类型	频率 (%)	F1 分数	精度	回忆	临床严重性
Sound Repetitions	28.4	0.92ś0.03	0.91ś0.04	0.93ś0.03	High
Syllable Repetitions	22.1	0.90ś0.04	0.89ś0.04	0.91ś0.04	High
Word Repetitions	15.3	0.88ś0.05	0.87ś0.05	0.89ś0.04	Medium
Prolongations	19.7	0.87ś0.04	0.86ś0.05	0.88ś0.04	High
Blocks (Silent)	8.2	0.84ś0.06	0.82ś0.07	0.86ś0.06	Very High
Blocks (Audible)	6.3	0.81ś0.07	0.79ś0.08	0.83ś0.07	Very High

重复显示最高准确性，而块由于有限的声学标记仍是最具挑战性的。

5 北京儿童医院的临床部署

5.1 临床结果

表 5 总结了部署结果。

关键成就包括评估时间减少 38%，患者吞吐量增加 58%，以及诊断准确性和一致性的显著提高。

5.2 可解释性分析

临床医生对 UDM 可解释性特征的反馈：

置信分数和视觉对齐图在临床应用中获得了最高评分。

表 5: 临床部署结果

度量	预部署	部署后	更改	P 值
Assessment Time (min)	45.3ś8.2	28.1ś6.4	-38.0%	< 0.001
Patients per Day per SLP	6.2ś1.1	9.8ś1.4	+58.1%	< 0.001
Diagnostic Accuracy (%)	87.4ś4.3	92.1ś3.2	+5.4%	< 0.01
Inter-rater Reliability	0.76ś0.08	0.89ś0.05	+17.1%	< 0.001
Patient Satisfaction	3.8ś0.7	4.3ś0.5	+13.2%	< 0.01
SLP Job Satisfaction	3.2ś0.8	4.1ś0.6	+28.1%	< 0.001
Clinician Acceptance Rate	-	87%	-	-

表 6: 可解释性特征评估

特征	有用性	清晰度	临床影响
Visual Alignment Maps	4.5ś0.5	4.3ś0.6	High
Confidence Scores	4.6ś0.4	4.4ś0.5	High
Phoneme Error Analysis	4.3ś0.6	4.1ś0.6	High
Threshold Controls	4.4ś0.5	4.2ś0.6	High
Feature Attribution	4.1ś0.6	3.9ś0.7	Medium
平均值	**4.4ś0.5**	**4.2ś0.6**	-

6 讨论

6.1 临床影响

我们的结果表明，UDM 成功弥合了临床口吃检测中的准确性和可解释性之间的差距。模块化设计使临床医生能够理解系统所检测的内容以及为何做出特定决策的原因。明确的音素对齐提供了与临床推理相一致的语言学有意义的表示。

显著的效率提升（减少 38%的时间）表明成功增强了临床专业知识，而不是替代。临床医生利用额外的时间进行治疗计划制定和接诊更多的患者，而诊断准确性的提高（增加了 5.4%）和评分者间可靠性的增加（提高了 17.1%）表明评估实践的标准化得到了增强。

6.2 技术贡献

“无约束”建模范式对于实际性能至关重要，开放集分类器在 8.3%的情况下识别出非典型模式。显式的音素对齐服务于双重目的：提高准确性并提供临床解释。这弥合了经常阻碍医生信任 AI 系统的语义鸿沟。

6.3 限制条件

几个限制依然存在：

- 静默块仍然具有挑战性（F1: 0.84ś0.06），因为缺乏声学标记
- 当前部署仅限于普通话使用者
- 长度进展追踪需要额外的验证
- 移动性能限制要求模型优化

7 结论

我们使用真实患者和言语语言病理学家对 UDM 进行了评估，实现了 F1 得分为 0.89±0.04，并且有 87%的临床接受度。UDM 表明可解释的人工智能可以达到最先进的性能水平并提供具有临床意义的结果。部署结果显示诊断时间减少了 38%，并且在诊断准确性方面有了显著提高，突显了 UDM 在改善临床言语语言病理学实践方面的潜力。

参考文献

[1] Jiachen Lian, Carly Feng, Naasir Farooqi, Steve Li, Anshul Kashyap, Cheol Jun Cho, Peter Wu, Robbie Netzorg, Tingle Li, and Gopala Krishna Anumanchipalli. Unconstrained dysfluency modeling for dysfluent speech transcription and detection. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[2] Jiachen Lian and Gopala Anumanchipalli. Towards hierarchical spoken language disfluency modeling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[3] Xuanru Zhou, Anshul Kashyap, Steve Li, Ayati Sharma, Brittany Morin, David Baquirin, Jet Vonk, Zoe Ezzes, Zachary Miller, Maria Tempini, Jiachen Lian, and Gopala Anumanchipalli. Yolo-stutter: End-to-end region-wise speech dysfluency detection. In *Interspeech 2024*, pages 937–941, 2024.

[4] Xuanru Zhou, Cheol Jun Cho, Ayati Sharma, Brittany Morin, David Baquirin, Jet Vonk, Zoe Ezzes, Zachary Miller, Boon Lead Tee, Maria Luisa Gorno-Tempini, et al. Stutter-solver: End-to-end multi-lingual dysfluency detection. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1039–1046. IEEE, 2024.

[5] Xuanru Zhou, Jiachen Lian, Cheol Jun Cho, Jingwen Liu, Zongli Ye, Jinming Zhang, Brittany Morin, David Baquirin, Jet Vonk, Zoe Ezzes, Zachary Miller, Maria Luisa Gorno Tempini, and Gopala Anumanchipalli. Time and tokens: Benchmarking end-to-end speech dysfluency detection, 2024.

[6] Jiachen Lian, Xuanru Zhou, Zoe Ezzes, Jet Vonk, Brittany Morin, David Paul Baquirin, Zachary Miller, Maria Luisa Gorno Tempini, and Gopala Anumanchipalli. Ssdm: Scalable speech dysfluency modeling. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[7] Jiachen Lian, Xuanru Zhou, Chenxu Guo, Zongli Ye, Zoe Ezzes, Jet Vonk, Brittany Morin, David Baquirin, Zachary Mille, Maria Luisa Gorno Tempini, and Gopala Krishna Anumanchipalli. Automatic detection of articulatory-based disfluencies in primary progressive aphasia. *IEEE JSTSP*, 2025.

[8] Chenxu Guo, Jiachen Lian, Xuanru Zhou, Jinming Zhang, Shuhe Li, Zongli Ye, Hwi Joo Park, Anaisha Das, Zoe Ezzes, Jet Vonk, Brittany Morin, Rian Bogley, Lisa Wauters, Zachary Miller, Maria Gorno-Tempini, and Gopala Anumanchipalli. Dysfluent wfst: A framework for zero-shot speech dysfluency transcription and detection. *Interspeech*, 2025.

[9] Zongli Ye, Jiachen Lian, Xuanru Zhou, Jinming Zhang, Haodong Li, Shuhe Li, Chenxu Guo, Anaisha Das, Peter Park, Zoe Ezzes, Jet Vonk, Brittany Morin, Rian Bogley, Lisa Wauters,

Zachary Miller, Maria Gorno-Tempini, and Gopala Anumanchipalli. Seamless dysfluent speech text alignment for disordered speech analysis. *Interspeech*, 2025.

[10] Zongli Ye, Jiachen Lian, Akshaj Gupta, Xuanru Zhou, Krish Patel, Haodong Li, Hwi Joo Park, Chenxu Guo, Shuhe Li, Sam Wang, et al. Lcs-ctc: Leveraging soft alignments to enhance phonetic transcription robustness. *arXiv preprint arXiv:2508.03937*, 2025.

[11] Jinming Zhang, Xuanru Zhou, Jiachen Lian, Shuhe Li, William Li, Zoe Ezzes, Rian Bogley, Lisa Wauters, Zachary Miller, Jet Vonk, Brittany Morin, Maria Gorno-Tempini, and Gopala Anumanchipalli. Analysis and evaluation of synthetic data generation in speech dysfluency detection. *Interspeech*, 2025.