arXiv:2301.05562v1 [eess.AS] 13 Jan 2023

通过自发语音识别多语种阿尔茨海默型痴呆：信号处理大挑战

Saturnino Luz[1], Fasih Haider[1], Davida Fromm[2], IoulietaLazarou[3], Ioannis Kompatsiaris[3] 和 Brian MacWhinney[2]

[1] 爱丁堡大学爱丁堡医学院 Usher Institute，英国[2] 卡内基梅隆大学心理学系，美国[3] 希腊塞尔米-塞萨洛尼基 CERTH 信息技术研究所

摘要

这项信号处理大挑战（SPGC）聚焦于一个具有重要社会意义和医疗价值的复杂自动预测问题，即阿尔茨海默型痴呆（AD）检测。挑战邀请参与者运用信号处理和机器学习方法，基于自发语音数据构建预测模型。该挑战旨在评估基于一种语言（英语）语音构建的预测模型在另一种语言（希腊语）上的泛化能力。据我们所知，尚无研究考察过多语种阿尔茨海默病检测中的语音信号声学特征。我们的基线系统采用传统机器学习算法，并使用声学特征的主动数据表示，在阿尔茨海默病检测中达到 73.91% 的准确率，在认知评分预测中达到 4.95 的均方根误差。

索引词——阿尔茨海默型痴呆检测、语音处理、语音生物标志物。

1. 引言

痴呆是一类神经退行性疾病，其认知功能会长期且通常逐渐下降。痴呆研究领域一直在寻求经济有效且准确的神经退行性变生物标志物，基于语音的“数字生物标志物”因此成为一种颇具前景的可能途径。尽管信号处理和机器学习领域对通过语音进行认知障碍检测的自动化方法十分关注， [1], 但大多数已提出的方法并未探究哪些语音特征能够跨语言泛化和迁移，以用于阿尔茨海默病预测；据我们所知，也尚无研究考察多语言阿尔茨海默病检测中的语音声学特征。本 SPGC“ADReSS-M：通过自发语音识别多语言阿尔茨海默型痴呆”旨在解决这一问题，设定了相关预测任务：参赛者使用英语语音数据训练模型，并在希腊语口语数据上评估模型表现。提交至本次挑战的模型侧重于语音信号的声学或语言特征，这些特征的预测能力能够在不同语言间保持。

本次 SPGC 旨在提供一个平台，供各方交流和讨论如何应用信号处理与机器学习方法进行多语言阿尔茨海默病识别，并推动围绕机器学习架构、新型信号处理特征、特征选择与提取方法，以及其他相关主题的讨论，以满足日益壮大的研究群体的兴趣；这些研究人员致力于探究语音与痴呆之间的联系。

2. 预测任务

ADReSS-M挑战赛包含以下任务：（1）分类任务，即模型旨在区分健康对照语音与阿尔茨海默病/轻度认知障碍语音；（2）MMSE评分预测（回归）任务，即根据语音数据创建模型，以推断说话者的简易精神状态检查（MMSE）评分。参与者可以选择完成其中一项或两项任务。主办方提供了训练集，并在论文提交截止日期前两周提供了测试集，供参与者测试模型。每位参与者在每项任务中最多可提交五组结果进行评分。所有尝试均须一并提交。

2.1. 数据集

这些 SPGC 数据集可通过 Dementia-Bank[1], 按需获取。训练数据集由自发语音样本组成，这些样本对应于认知正常受试者和经阿尔茨海默病诊断的患者对波士顿失语症诊断检查中的偷饼干图片进行描述时的音频录音[2]。参与者均为英语使用者。测试集由希腊语中对另一幅图片的自发语音描述组成。录音以这些语言之一录制。参与者最初只能访问训练数据（英语）以及部分希腊语样本数据（8 段录音），用于开发。

[1]https://dementia.talkbank.org/

希腊语录音通过一幅图片评估参与者的言语流畅度和情绪，参与者在观看图片时对其进行描述。评估者首先向参与者展示一幅图片，画面中一只狮子躺在沙漠里，一边进食，一边陪着一只幼崽。随后，评估者要求参与者用几句话描述图片。

训练数据集在年龄和性别方面进行了平衡，以消除潜在的混杂和偏差。由于我们采用倾向评分匹配方法，因此无需对教育程度进行调整，因为教育程度与年龄和性别相关，而年龄和性别已足以作为可接受的调整变量（见[3, 第348页-352]）。我们根据逻辑回归估算的概率分数检查数据集的匹配情况，该概率表示在给定协变量年龄和性别的条件下，某个样本被判定为阿尔茨海默病患者的概率，并选取匹配的样本。所有协变量的标准化均值差均低于0.1，协变量平方项及其双向交互作用的所有标准化均值差均低于0.15，表明这些协变量达到了良好的平衡。

2.2. 评估

分类任务采用准确率、特异度、灵敏度和 F_1 分数进行评估。回归任务（MMSE预测）使用的指标为决定系数和均方根误差。提交结果的排名依据分类任务（任务1）的准确率分数，以及MMSE分数回归任务（任务2）的RMSE分数。

3. 基线模型

首先，我们使用 ffmpeg 的 EBU R128 扫描器滤镜对音频文件的音量进行了归一化处理。随后，对音频应用了长度为 1 秒且无重叠的滑动窗口，并在这些帧上提取 eGeMAPS 特征。eGeMAPS 特征集 [4] 是一组用于检测发声过程中生理变化的基础声学特征。该特征集包含F0半音、响度、谱通量、梅尔频率倒谱系数、基频微扰、闪烁、F1值、F2、F3、阿尔法比、哈马伯格指数和V0斜率特征，以及这些特征最常用的统计函数，每帧共计 88 个特征。基于 eGeMAPS 特征，我们应用主动数据表示方法（ADR） [5]，为每段音频录音生成帧级声学表示。此前，ADR 方法已被用于生成大规模时间序列数据表示。该方法采用自组织映射对原始声学特征进行聚类，然后在这些聚类上计算二阶特征，以提取新的特征 [5]。这种方法完全自动化，因为无需向算法提供语音分段或说话人分离信息。

对于任务 1，我们采用了带有核平滑估计的朴素贝叶斯分类器。特征提取的主动数据表示通过网格搜索进行优化（$C = 5$, 10,15,20,25）。使用 15+2 主动数据表示、年龄和性别作为每段录音的特征，我们在样本数据和测试数据上分别取得了 75.00% 和 73.91% 的准确率。在测试集上，特异度为 79.2%，精确率为 75%，灵敏度为 68.2%，F_1 为 71.4%。特征与训练音频的比例为 19:237。

对于简易精神状态检查回归任务（任务 2），我们采用了支持向量机（SVM）模型，使用径向基函数核、盒约束为 1，并采用序列最小优化求解器。我们通过网格搜索（$C = 5$, 10,15,20,25）优化了用于特征提取的主动数据表示（ADR）。该模型使用每段录音的 25+2 ADR、年龄和性别特征，在样本数据和测试数据上分别取得了 3.887（$r = 0.273$）和 4.955（$r = 0.348$）的均方根误差（RMSE）。特征与训练音频录音的比例也为 29:237。

4. 结论

自发语音分析有望为语音技术带来新的应用，使其能够长期、无干扰地监测认知健康。本 SPGC 聚焦于利用自发语音进行阿尔茨海默病识别，探索了神经心理学和临床评估方法之外的阿尔茨海默病检测与认知评估途径。此外，SPGC 提供的多语言环境使研究人员能够考察可能在不同语言间具有普适性的特征，从而拓展模型的适用范围。为符合阿尔茨海默病预测评估的目标，ADReSS-M挑战赛提供了统计匹配的数据集，以减轻阿尔茨海默病检测方法评估中常被忽视的常见偏差，包括同一参与者的语音反复出现、音频质量差异，以及性别、年龄和教育水平分布不均衡。我们希望这能成为未来多语言阿尔茨海默病评估研究的基准。

5. 参考文献

[1] S. de la Fuente Garcia、C. Ritchie 和 S. Luz，《人工智能、语音与语言处理方法在阿尔茨海默病监测中的应用：系统综述》，《阿尔茨海默病杂志》，第 78 卷，第 4 期，2020 年。[2] J. Becker、F. Boller、O. Lopez、J. Saxton 和 K. Mc- Gonigle，《阿尔茨海默病自然史：研究队列描述与诊断准确率》，《神经病学档案》，第 51 卷，第 6 期，第 585–594 页，1994 年。[3] J. Pearl，《因果关系：模型、推理与推断》，剑桥大学出版社，第 2 版，2009 年。[4] F. Eyben 等，《面向语音研究与情感计算的日内瓦精简声学参数集》，IEEE 情感计算汇刊，第 7 卷，第 2 期，2016 年。[5] F. Haider、S. de la Fuente 和 S. Luz，《评估用于检测自发语音中阿尔茨海默型痴呆的副语言声学特征》，IEEE 信号处理专题精选期刊，第 14 卷，第 2 期，第 272–281 页，2020 年。