非裔美国英语的自动语音识别：词汇和上下文的影响

Hamid Mojarad[1], *Kevin Tang*[1,2]

[1]Department of English Language and Linguistics, Institute of English and American Studies, Faculty of Arts and Humanities, Heinrich Heine University Düsseldorf, Germany
[2]Department of Linguistics, University of Florida, United States of America

hamid.mojarad@hhu.de, kevin.tang@hhu.de

Abstract

自动语音识别（ASR）模型经常在非裔美国英语（AAE）的语音、音系和形态句法特征方面遇到困难。本研究重点关注两个关键的 AAE 变量：辅音簇简化（CCR）和 ING 简化，探讨 CCR 和 ING 简化的存在是否会增加 ASR 误识别。随后，研究无外部语言模型（LM）的端到端 ASR 系统与有 LM 的系统相比，是否受到词汇邻居效应的影响更大，而受到上下文预测性影响更小。采用 wav2vec 2.0 对地区非裔美国语言语料库（CORAAL）进行转录，分为有和无 LM 两种情况。使用蒙特利尔强制对齐（MFA）工具通过发音扩展检测 CCR 和 ING 简化。分析表明，CCR 和 ING 对字错误率（WER）有小但显著的影响，并且在没有 LM 的 ASR 系统中词汇邻居效应更强。

Index Terms: Automatic Speech Recognition, African American English, Consonant Cluster Reduction, velar nasal fronting, ING-reduction, lexical neighborhood, contextual predictability

1. 介绍

最近，解决自动语音识别（ASR）中的种族偏见已成为一个重要关注领域。考虑到 AAE 是一种少数方言，Koenecke 等人通过语音学确认了这一问题，他们发现，在五个主要的 ASR 系统中，白人美籍说话者的平均词错误率（WER）明显低于 AAE 说话者。形态句法上的差异进一步由 Martin 和 Tang 强调，他们对“习惯性 be”这一常见的 AAE 形态句法特征的研究显示，ASR 表现中的 WER 增加。Wassink 等人对美国太平洋西北部的四种民族方言进行的研究也强调了语音学差异，研究表明 AAE 说话者的 WER 更高。

在 Wassink 等人进行的错误分类中，辅音群减少被确定为导致 AAE 相关错误的最常见特征。然而，这些错误的具体来源仍未得到深入探索。辅音群减少被定义为词尾辅音群的简化，通常涉及双辅音群中最后的停顿音的省略（例如，cold / koʊld / → [koʊl]）或三辅音群中倒数第二个辅音的省略（例如，fists / fɪsts / → [fɪsː]）。考虑到英语中双辅音群比三辅音群更为普遍，本文研究特别集中在以最终停顿音结尾的双辅音群。

在社会语言学研究中，ING-缩减指的是以-ing 结尾的单词中发音的变化，重点在于最终鼻音段 [6] 的实现。这种变化主要以两种形式表现：标准的-ing 发音为软腭鼻音 [ŋ]，以及缩减的-in 发音为齿龈鼻音 [n]。这种现象既出现在单个词素中（例如进行时的后缀-ing），也出现在较大的词形中（例如，something，during），但是单音节词（例如，thing，king）被排除在外，因为它们并不表现出这种变异。

词汇邻域密度在语音识别中起着至关重要的作用，影响着人类感知和自动语音识别系统。这一现象在研究人类语音感知时得到了验证，通常情况下，拥有许多词汇邻居的词比词汇邻居较少的词识别得更不准确且更慢。关于自动语音识别的研究表明，词汇邻域度量可以作为识别错误的预测因素，在相似上下文中有强大竞争者的词更容易被误识别。

语音简化，即词语以缩短或简化的形式发音，可能导致更容易被误识别的非词语感知。这些简化形式由于其较短的长度，往往具有更密集的词汇邻域，从而增加了准确感知的挑战 [10] 。例如，“test”的简化形式 [tɛs] 有 21 个邻近词，包括在 CLEARPOND 数据库中识别到的“guess”和“ten”等词 [10] 。这种语音简化与词汇邻域密度之间的相互作用凸显了人类听众和语音识别系统在准确感知语音，特别是在随意或快节奏的对话环境中所面临的复杂挑战 [7, 9] 。

在通过整合上下文知识或文本适应机制来提高 ASR 模型性能方面，上下文可预测性被认为是一个关键因素。最近的研究表明，采用类似于人类认知过程的这种机制可以提高转录的准确性。在与语言模型集成的 ASR 系统中，语言模型在基于上下文信息预测词语方面起着作用，特别是在面临退化的声学信号、超出词汇表的词汇或模糊的语音序列等挑战时。在这些情况下，当声学信号不清晰时，类似于人类听众，语言模型可能会优先考虑在句子上下文中具有高上下文可预测性的词语。通过动态适应上下文线索，语言模型可以提高转录准确性，特别是对于那些仅基于声学信息难以识别的词语。这种基于上下文的预测方法将 ASR 系统从纯粹依赖声学驱动的模型转变为更智能、具有上下文感知的转录工具，从而能以更高的精确性和适应性处理复杂的语言场景。

鉴于多达 20 % 的 ASR 错误可以由社会语言学的语音变量 [3] 解释，本研究集中于两种常见的非洲裔美国英语（AAE）语音特征，即 CCR 和 ING 缩减。在 Wassink 等人的基础上 [3] ，我们研究这些变量如何影响 ASR 的性能，使用更大样本的 AAE，并考察由此产生的错误是否可以通过词汇邻近效应和上下文可预测性来预测。

在这项研究中，我们提出了两个主要假设 (H1 和 H2)。在 H1 中，我们假设 AAE 中 CCR 和 ING 缩减特征的存在导致 ASR 识别错误增加和更高的 WER。基于此，我们的第二个假设 (H2) 推测将外部 LM 集成到最先进的 ASR 中将导致 lexical neighborhood 错误减少。我们的错误分析方法以这些 AAE 的常见音韵特征为中心，旨在量化外部 LM 在通过减少 lexical neighborhood 错误来提高 ASR 性能的程度。数据和代码可在 `osf.io` [1] 上获得。

2. 方法论

2.1. 语料库

区域非裔美国人语言语料库（CORAAL）[17] 是这项研究的基础数据集，提供了对区域非裔美国人语言（AAL）变体的全面记录。该语料库提供了丰富的语言资源，包括具有时间对齐正字法转录的音频记录，格式为 TextGrid，特色是在词句和词/音素对齐层面上的特定发言者层。对于

[1]https://doi.org/10.17605/OSF.IO/QN6A2

这项研究，我们特别使用了 DCA（华盛顿特区）子语料库，其中包括来自 68 位发言者（40 名男性和 28 名女性）的 74 份录音，这些发言者分为四个年龄组：ag1（19 岁以下）、ag2（20-29 岁）、ag3（30-50 岁）和 ag4（51 岁及以上），以及三个社会经济阶层（1 到 3，从最低到最高）。该数据集包括 34 小时的社会语言学访谈，总计 333,500 个词，以 WAV 格式录制（44.1 kHz，16-bit，单声道）。

为了提取 CCR 和 ING-减少特征，我们采用了受 Kendall 等人启发的强制对齐方法。他们的研究比较了社会语言学变量（ING）的人工编码与强制对齐和机器学习分类器，证明自动编码算法在分类 ING 变体方面可以接近人工编码员的表现。借鉴这一思路，我们利用蒙特利尔强制对齐器（MFA，版本 2.2.17）和卡内基梅隆大学（CMU）发音词典来自动化我们分析的特征提取过程。我们识别出数据集中缺失的词汇，基于 CMU 词典训练了一个字素到音素（g2p）模型，并生成了这些词的发音。然后，我们将这些新增词汇更新到 CMU 词典中。对于易于发生 CCR 或 ING 变异的词汇，我们在词典中同时包含了原始和简化发音（例如，“accept” 既表示为 “AH0 K S EH1 P T”，又表示为 “AH0 K S EH1 P”）。使用 MFA 的训练命令，我们在整个 DCA 音频集上开发了一个自定义声学模型。最后，我们使用这一经过训练的声学模型和更新后的 CMU 词典对完整音频集进行了对齐。

2.2. 自动语音识别转录

我们采用了 wav2vec 2.0 [19] 作为我们数据转录的端到端 ASR 模型之一，特别是使用了预训练模型 facebook/wav2vec2-large-960h。此版本经过 960 小时语音训练，随后通过外部训练在 CORAAL 的 DCA 和 DCB 子语料库（整个华盛顿特区数据）上的 5-gram LM 进行增强，使用 KenLM [20] 。为了在有和没有 LM 的情况下进行转录，我们首先将我们的音频文件重新采样到 16kHz 以与 wav2vec 2.0 兼容。音频随后被分割成至少 30 秒的块，重点关注 CCR 和 ING 缩减倾向词，同时利用 CORAAL 提供的说话者发言时间框架确保没有句子在发音中被分割。分割后的音频块通过 wav2vec 2.0 进行处理，既有带 LM 也有不带 LM。随后，我们使用 Needleman-Wunsch 算法进行序列对齐，通过 Python 的 string2string [2] 库实现，将转录结果与相应的真实情况对齐。此对齐使我们能够评估转录准确性。

2.3. 后处理

为了确保对 AAE 特征进行重点分析，我们仅处理来自采访对象的发言，排除来自 DCA 数据集中采访者的发言。利用前一阶段获得的电话对齐，我们提取了每个目标词的相应转录及其相关发言。此数据用于计算 WER 并检查可能的词汇邻近效应。

根据 Luce 的定义，我们将一个单词视为词汇邻居，如果它可以通过在任何位置进行一个音素替换、删除或添加从目标单词派生而来。为了识别这些邻居，我们使用了 Levenshtein 算法。我们的过程首先涉及检查自动语音识别（ASR）转录的单词是否存在于 CMU 词典中。对于未找到的单词，我们使用 MFA 的 g2p 命令生成发音，然后将新条目更新到 CMU 词典中。然后，我们计算语音的 Levenshtein 距离，以汇编每个目标单词的词汇邻居列表。我们使用每个目标单词的 MFA_Status 来确定它是以其原始形式还是简化形式被 MFA 检测到的。如果是原始的，我们根据单词的完整发音生成邻居列表。否则（对于简化形式检测），我们基于单词的简化发音获得邻居列表。最终，如果转录的单词出现在这些邻居中，我们将转录错误归因于词汇邻域效应。

[2] https://github.com/stanfordnlp/string2string

在假设 H1 中，将 WER 作为因变量进行分析，MFA_Status、AgeGroup 和 Gender 作为固定效应变量。MFA_Status 作为一个二元因子，表示发音是否被 MFA 检测为原始或简化。

为了处理数据中潜在的非独立性，特别是频繁出现的目标词汇和个体说话人特征的影响，采用了线性混合效应回归，其中将 Target_Word 和 Speaker_Id 作为随机效应。此外，对于 Target_Word 随机效应，我们包括了 MFA_Status、AgeGroup 和 Gender 的随机斜率。这意味着发音风格、年龄组和性别对 WER 的影响被允许对不同的目标词有所不同。同样，对于 Speaker_Id 随机效应，我们包括了 MFA_Status 的随机斜率，这允许发音风格的影响在不同的说话人之间有所变化。分析在三个数据集（总体、仅 CCR 和仅 ING）上进行，并使用两种 ASR 类型（有 LM 和无 LM）。WER 的描述性统计在图 1 中进行了可视化。

使用 R 软件（版本 4.4.1）中 `lmerTest` 包的 `lmer` 函数拟合了

2.3.1. 统计程序

线性混合效应模型。该模型指定 WER 是由 MFA_Status、AgeGroup 和 Gender 的固定效应预测的，其中对 Target_Word 的预测具有随机截距和斜率，对 Speaker_Id 的 MFA_Status 具有随机截距和斜率。这些分类变量进行了对比编码。MFA_Status 被编码为 −0.5 表示 “原始发音” 和 0.5 表示 “简化发音”。Gender 被编码为 −0.5 表示男性和 0.5 表示女性。AgeGroup 进行了赫尔默特编码，以将每个水平与之前水平的平均值进行比较。

如图 1 所示，结果的描述性统计表明，根据中位数值，减少的发音导致更高的 WER，但仅限于 CCR。然而，在所有数据集中——总体、CCR 和 ING，在表 1 中——MFA_Status 在统计上对 WER 有显著的正效应。无论是否包含语言模型（LM），这一效应都持续存在，表明这些变化对 ASR 系统构成了一致的挑战。虽然影响在统计上显著，$\hat{\beta}$ 值范围从 0.021 到 0.040 的相对较小效应量表明其对识别准确率的影响是中等而非严重。AgeGroup 似乎对 ASR 性能有显著影响，当第二年龄组与第一组比较时（没有 LM 的 ING 数据集中 $\hat{\beta}$ 值最高）。然而，性别对 CCR/ING 容易出错的词的 ASR 性能没有显著影响。

在第二个假设中，Neighborhood_Status 被分析为因变量，ASR_Type 作为固定效应变量。Neighborhood_Status 被编码为二进制变量（参考级别：Neighbor_Error），并对 ASR_Type 应用了对比编码（without_LM：-0.5，with_LM：0.5）。使用混合效应逻辑回归，其中 Target_Word 和 Speaker_Id 作为随机效应。此外，对于两个随机效应，我们包含了 ASR_Type 的随机斜率，以允许 ASR 类型对 Neighborhood_Status 的影响在不同的目标词和个人说话者之间变化。

使用 R 中 `lme4` 包的 `glmer` 函数拟合了一个逻辑混合效应模型。选择了 logit 链接函数，因为 Neighborhood_Status 变量是二进制的（Non_Neighbor_Error 与 Neighbor_Error）。该模型应用于合并了两种 ASR 类型的数据集。为了实施该模型，我们过滤掉了目标词正确转录的 ASR 词，以仅获取错误。

在没有语言模型的自动语音识别（ASR）中，我们观察到我们的目标词在 12,734 个总错误转录中有 7.9 %（1,006）的邻域错误。然而，随着语言模型的整合，邻域错误的数量大幅下降到在 8,283 个总错误识别中的 3.3 %（277）。回归模型证实了这一描述性发现，显示跨所有数据集的显著效果，表明语言模型的使用影响了词汇邻域错误。当比较有和没有语言模型的 ASR 时，ASR_Type 显示出一致且显著的正向效果（p 至 < 0.001 ）。这种效果在 ING 数据集（$\hat{\beta}$ ：−2.1879）中最强，其次是整体数据集（$\hat{\beta}$ ：−1.2875）

Table 1: 跨数据集和 *ASR* 类型的固定效应总结

Effect	Overall Dataset		CCR Dataset		ING Dataset	
	Without LM	With LM	Without LM	With LM	Without LM	With LM
MFA_Status	0.021 (0.006) ***	0.030 (0.006) ***	0.026 (0.007) ***	0.031 (0.007) ***	0.040 (0.012) **	0.030 (0.012) *
Age Group (2 vs. 1)	-0.158 (0.042) ***	-0.108 (0.035) **	-0.146 (0.042) ***	-0.102 (0.034) **	-0.173 (0.046) ***	-0.131 (0.039) **
Age Group (3 vs. 2,1)	-0.041 (0.030)	-0.035 (0.025)	-0.038 (0.030)	-0.032 (0.024)	-0.035 (0.033)	-0.029 (0.027)
Age Group (4 vs. 3, 2, 1)	0.033 (0.029)	0.037 (0.024)	0.038 (0.028)	0.033 (0.023)	0.030 (0.031)	0.035 (0.026)
Gender (Female vs. Male)	-0.013 (0.033)	-0.028 (0.027)	-0.011 (0.033)	-0.029 (0.027)	-0.003 (0.037)	-0.013 (0.031)

Note: Values are presented as: Estimate (Standard Error). Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.

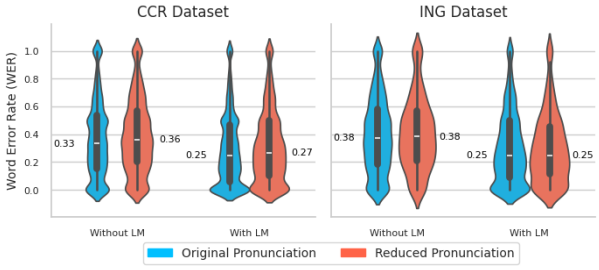

Figure 1: *CCR* 和 *ING* 目标词的 *MFA* 状态 *WER*

和 CCR 数据集 ($\hat{\beta}$: −0.8954) 。

3. 讨论

我们的研究揭示了当面对 CCR 和 ING-reduction，这些常见的 AAE 变化时，ASR 系统的性能有显著的洞察。各数据集中 MFA_Status 的一致积极作用表明，AAE 特征对 ASR 误识别有显著影响。即使在引入外部 LM 以提供更多背景以帮助 ASR 生成更准确的转录时，这种影响仍然显著。因此，这强烈支持了我们的第一个假设，即 CCR 和 ING-reduction 变化的存在导致 ASR 误识别的增加。

在这一发现的基础上,我们在图 1 中详细列出的 WERs 与先前关于 AAE 数据集上 wav2vec 2.0 模型的报告相当。例如，Johnson 等人 [24] 报告说，当在 AAE 语音上使用 wav2vec 2.0 时，故事复述的 WER 为 39 %，图片描述任务的 WER 为 30 %。同样，Chang 等人 [25] 发现 wav2vec 2.0 对 CORAAL 数据集的转录的 WER 为 52.8 %，并强调了包含更多音位和形态句法 AAE 特征的语句显示出更高的错误率。这些发现与我们的结果一致，并突显了识别 AAE 语音的挑战，强调了进一步的模型适应需求，以改善方言多样性处理。

观察到的与年龄相关的对自动语音识别（ASR）性能的影响突出了说非洲裔美国英语（AAE）社群中语言使用的代际差异。较年轻的说话者在各数据集中的字错误率（WER）都高于年长说话者，这表明元音连包音略以及 ING 音的减少变化为当前的 ASR 系统带来了额外的挑战。这个发现与 ASR 性能的广泛理解形成对比，通常显示儿童 [26] 和年长说话者 [27] 的 WER 较高，这通常是由于发音变化和语速较慢等因素。在我们的案例中，ag1 组更准确地描述为青少年或少年而不是儿童，因为这些说话者是 1968 年至 1969 年间录制的，出生日期范围是 1891 年到 1958 年 [17] 。这意味着在录制时，最年轻的说话者至少有 10 岁。

与之前几项报告中关于 ASR 性能存在性别差异的研究不同，我们的研究未发现性别对识别准确率有显著影响。此发现与现有文献的结果不一致，因为许多研究显示出混合结果，部分研究倾向于男性讲话者的优势 [28] ，另一些则表明女性讲话者表现更好 [1, 29] 。这一发现表明，至少在本研究的范围内，基于性别的差异可能在 AAE 讲话者的 ASR 性能中并不发挥实质性作用。

在我们的第二个假设中，我们认为将外部语言模型集成到语音识别模型中将减少由词汇邻近效应引起的错误。我们在第 3.2.3 节的研究结果强烈支持了这一点。换句话说，尽管端到端语音识别模型通常因能够消除对单独语言模型的需求而受到推广，但我们的结果与最近的研究一致，这些研究强调了语言模型在提高语音识别表现中持续的重要性。如图 1 所示，整合语言模型显著降低了 CCR 和 ING 数据集的词错误率。这种减少可以归因于语言模型提供上下文预测能力，从而减轻词汇邻近效应。

此外，研究发现，在没有语言模型的 ASR 系统中，非邻错误明显比邻错误更频繁。这表明可能还有其他因素导致这些错误，例如训练数据的总体数量有限、训练数据与测试数据在声学上的不匹配 [1] ，以及我们尚未考虑的其他方言特征 [2] 。

我们发现的一个关键影响是，在训练过程中标注语音变化可以通过明确捕捉 AAE 的声学变化来提高 ASR 的准确性。例如，已经进行了一些关于 AAE 自动特征标注的工作（参见 [31] 及其引用）。这样的标注将有助于 ASR 系统更好地考虑系统性的语音差异，如 CCR 和 ING 缩减，从而提高准确性并减少对代表性不足的语音群体的偏见。

我们研究的几个局限性可以在未来解决。首先，由于时间限制，我们未能通过人工编码来评估 MFA 检测的 CCR 和 ING-缩减变量。就像 Kendall 等人在 CORAAL 中对 ING-缩减所做的那样，这种比较可以增强我们的 CCR 结果的普遍性。其次，我们使用了 Large-960h wav2vec 2.0 模型，由于其与外部语言模型的兼容性，以验证第二个假设；然而，评估拥有更多训练时长的模型可能会降低错误率，并提供更清晰的词汇邻域错误情况。这一要求也限制了我们在测试第一个假设时的模型选择，因为加入更多的 ASR 模型可以提高研究的普遍性。最后，我们没有明确测试 CCR/ING-缩减对 ASR 性能的影响是否受词汇邻域密度增加的影响。相反，我们依赖于单词长度和邻域大小之间的公认关系。

4. 结论

本研究考察了 ASR 系统的性能，重点关注 CCR 和 ING-reduction，这两种是 AAE 中的常见音韵变化。我们的研究结果强调了 ASR 系统在转录方言言语时面临的持续挑战，即使使用了像 wav2vec 2.0 这样的先进架构和 LM 的集成。首先，我们的结果证实了 CCR 和 ING-reduction 变化显著导致 ASR 误识别，强烈支持了我们的初始假设。为了进一步探索这一点，我们分析了性别和年龄对 ASR 性能的影响。虽然性别未显示出显著影响，但在 19 岁以下的 AAE 说话者中，年龄是一个关键因素，导致更高的 ASR 误识别率。其次，在所有数据集中，带有 LM 的 ASR 在减少邻域错误方面始终优于没有 LM 的 ASR。这验证了我们的第二个假设，并突出 LM 利用上下文可预测性、减少发音相似词语之间的混淆并提高转录准确性的能力。

5. 致谢[3]

这项工作是 HM 博士研究的一部分。构思、方法学、形式分析、撰写–原始草稿/审阅/编辑: HM, KT；数据整理、调查、资金获取、项目管理、软件、验证: HM；资源、监督: KT。

6. References

[1] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," Proceedings of the National Academy of Sciences , vol. 117, no. 14, pp. 7684–7689, 2020.

[2] J. L. Martin and K. Tang, "Understanding racial disparities in automatic speech recognition: The case of habitual "be" ," in Interspeech 2020 , 2020, pp. 626–630.

[3] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," Speech Communication , vol. 140, pp. 50–70, 2022.

[4] E. R. Thomas and G. Bailey, "Segmental phonology of African American English," in The Oxford Handbook of African American Language . Oxford University Press, 07 2015.

[5] R. Gregová, "A comparative analysis of consonant clusters in English and in Slovak," Bulletin of the Transilvania University of Brasov. Series IV: Philology and Cultural Studies , pp. 79–84, 2010.

[6] T. Kendall, C. Vaughn, C. Farrington, K. Gunter, J. McLean, C. Tacata, and S. Arnson, "Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ING)," Frontiers in Artificial Intelligence , vol. 4, 2021.

[7] P. A. Luce and D. B. Pisoni, "Recognizing spoken words: The neighborhood activation model," Ear and Hearing , vol. 19, no. 1, pp. 1–36, 1998.

[8] P. Jyothi and K. Livescu, "Revisiting word neighborhoods for speech recognition," in Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM , Ö. Çetinoğlu, J. Heinz, A. Maletti, and J. Riggle, Eds. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1–9.

[9] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," Speech Communication , vol. 52, no. 3, pp. 181–200, 2010.

[10] V. Marian, J. Bartolotti, S. Chabal, and A. Shook, "Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities," PLOS ONE , vol. 7, no. 8, pp. 1–11, 08 2012.

[11] K. Huang, A. Zhang, Z. Yang, P. Guo, B. Mu, T. Xu, and L. Xie, "Contextualized end-to-end speech recognition with contextual phrase prediction network," in Interspeech 2023 , 2023, pp. 4933–4937.

[12] N. Manh Tien Anh and T. Ho Sy, "Improving speech recognition with prompt-based contextualized ASR and LLM-based re-predictor," in Interspeech 2024 , 2024, pp. 737–741.

[13] J. Fox and N. Delworth, "Improving contextual recognition of rare words with an alternate spelling prediction model," in Interspeech 2022 , 2022, pp. 3914–3918.

[14] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," Transactions of the Association for Computational Linguistics , vol. 12, pp. 157–173, 2024.

[3] https://credit.niso.org/

[15] R. G. Podlubny, T. M. Nearey, G. Kondrak, and B. V. Tucker, "Assessing the importance of several acoustic properties to the perception of spontaneous speech," The Journal of the Acoustical Society of America , vol. 143, no. 4, pp. 2255–2268, 04 2018.

[16] Y. Tang and A. K. H. Tung, "Contextualized speech recognition: Rethinking second-pass rescoring with generative large language models," in Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24 , K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 6478–6485.

[17] T. Kendall and C. Farrington, "The Corpus of Regional African American Language," 2023, publisher: The Online Resources for African American Language Project. [Online]. Available: https://oraal.uoregon.edu/coraal

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in Proc. Interspeech 2017 , 2017, pp. 498–502.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in Proceedings of the 34th International Conference on Neural Information Processing Systems , ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[20] K. Heafield, "KenLM: Faster and smaller language model queries," in Proceedings of the Sixth Workshop on Statistical Machine Translation , C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, Eds. Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 187–197.

[21] P. A. Luce, "Neighbourhoods of words in the mental lexicon," Indiana University, Bloomington, IN, Tech. Rep. Technical Report No. 6, 1986.

[22] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in Soviet Physics Doklady , vol. 10, no. 8, 1966, pp. 707–710.

[23] R Core Team, R: A Language and Environment for Statistical Computing , R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[24] A. Johnson, C. Chance, K. Stiemke, H. Veeramani, N. B. Shankar, and A. Alwan, "An analysis of large language models for African American English speaking children's oral language assessment," Journal of Black Excellence in Engineering, Science, & Technology , vol. 1, dec 4 2023.

[25] K. Chang, Y.-H. Chou, J. Shi, H.-M. Chen, N. Holliday, O. Scharenborg, and D. R. Mortensen, "Self-supervised speech representations still struggle with African American Vernacular English," in Interspeech 2024 , 2024, pp. 4643–4647.

[26] P. Gurunath Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," Computer Speech & Language , vol. 63, p. 101077, 2020.

[27] S. Kwon, S.-J. Kim, and J. Y. Choeh, "Preprocessing for elderly speech recognition of smart devices," Computer Speech & Language , vol. 36, pp. 110–121, 2016.

[28] R. Tatman and C. Kasten, "Effects of talker dialect, gender and race on accuracy of Bing Speech and YouTube automatic captions," in Interspeech 2017 , 2017, pp. 934–938.

[29] C. Harris, C. Mgbahurike, N. Kumar, and D. Yang, "Modeling gender and dialect bias in automatic speech recognition," in Findings of the Association for Computational Linguistics: EMNLP 2024 , Y. Al-Onaizan,

M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15 166–15 184.

[30] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in Proceedings of the 31st International Conference on Machine Learning , ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1764–1772.

[31] R. Porwal, A. Rozet, J. Gowda, P. Houck, K. Tang, and S. Moeller, "Analysis of LLM as a grammatical feature tagger for African American English," in Findings of the Association for Computational Linguistics: NAACL 2025 , L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 7744–7756.