面向印度语言语音识别系统的正字法感知评估

Kaushal Santosh Bhogale[*1], *Tahir Javed*[*1], *Greeshma Susan John*[*1], *Dhruv Rathi*[2], *Akshayasree Padmanaban*[1],*Niharika Parasa*[1],*Mitesh M. Khapra*[1]

[1]AI4Bharat、WSAI、印度理工学院马德拉斯分校，印度， [2]Sarvam AI，印度
cs22d006@cse.iitm.ac.in，miteshk@dsai.iitm.ac.in

arXiv:2603.00941v1 [cs.CL] 1 Mar 2026

摘要

由于拼写变体、后缀拆分灵活性以及混合语码词中的非标准拼写，评估印度语言的自动语音识别系统颇具挑战。传统词错误率（WER）对系统性能的评价往往比人类用户的实际感受更悲观。要使评估结果更贴近现实表现，就需要捕捉允许存在的正字法变体；对于资源匮乏的印度语言而言，这极具挑战性。借助大语言模型的最新进展，我们提出了一个用于创建基准测试集的框架，以捕捉允许存在的变体。通过大量实验，我们证明，OIWER通过考虑正字法变体，降低了悲观的错误率（平均提升6.3个百分点），缩小了被夸大的模型差距（例如，Gemini与Canary之间的性能差距从18.1个百分点降至11.5个百分点），并且与此前的WER-SN等方法相比，更贴近人类感知，提升了4.9个百分点。

索引词— 语音识别、评估、低资源

1. 引言

过去几年中，印度语言自动语音识别（ASR）数据集的创建和模型的构建取得了稳步进展，旨在使其性能接近英语等高资源语言。尽管取得了这些进展，印度语言的词错误率（WER）仍显著偏高，与英语系统报告的个位数词错误率形成鲜明对比。 [1, 2] [3, 4, 5] [6]

进一步分析发现，这一差距在很大程度上并非仅由自动语音识别模型的错误造成，还源于缺乏完善的正字法标准以及针对印度语言的一致转写指南，导致词错误率数值偏高。这个问题根植于这些语言固有的丰富性：常见后缀的拼写变体灵活多样，语素可以拆分或合并，而且混合语码语境中非标准拼写十分普遍。尽管这类混合语码词在日常口语中广泛使用，但至今尚未得到全面规范。现有转写指南仅部分涵盖这些允许的变体，许多仍未编入规范。因此，转录文本中的风格不一致会引入评估噪声，使自动语音识别系统看起来不如实际准确。

以往的研究曾尝试通过多参考词错误率 [7, 8, 9, 10], 来应对这一挑战，即为同一段音频收集多个独立的转录文本。自动语音识别系统的预测结果会分别与每个参考文本进行评分，并取最低词错误率作为最终得分。尽管这种方法能有效提高稳健性，但资源消耗较大，需要投入大量人力和成本。此外，少量参考文本仍可能无法涵盖所有有效的正字法变体。另一种策略是确保全面覆盖允许的变体（见表 1）。以往沿这一方向进行的尝试，依赖人工整理[11], 从推特等社交媒体来源提取内容 [12], 或采用启发式归一化规则 [13]。然而，这些方法往往需要大量标注工作量、获取大型文本语料库，或搭建专用流水线，而这些条件对于低资源印度语言而言都难以满足。

本文提出了一种框架，用于构建正字法感知基准测试集，以系统地涵盖可能的拼写和构词变体。我们的方法通过为每个词生成详尽的候选变体列表，扩充现有基准测试集。尽管手动编制此类列表并不可行，但我们借助大语言模型（LLM）自动提出变体，仅需人工参与完善和纠正。为将这些词级变体纳入评估，我们提出了正字法感知词错误率（OIWER）指标。通过广泛的实验，我们证明，考虑正字法变体后，OIWER 能够降低偏高的错误率：对于最先进的 Canary模型 [14], ，所有语言的平均得分提高了 6.3 分，且错误率估计与人类感知更加吻合。此外，OIWER 缩小了模型之间被夸大的性能差距（例如，Gemini模型与 Canary模型之间的平均差距从 18.1 分降至 11.5 分），从而更真实地反映了进展。最后，与 WER-SN [13], 等以往基于规范化的方法相比，OIWER 与人类感知的吻合度提高了 4.9 分。代码可在此 GitHub 页面获取。

2. 相关研究

为应对拼写变体和方言差异， [7] 针对方言阿拉伯语自动语音识别提出了多参考词错误率（MR-WER），之后又将其扩展到其他方言 [8]。类似地， [15] 利用基于格的参考文本，为日语提出了“宽松”的字符错误率（CER）； [10]则将多参考方案应用于英语的风格变体。然而，收集多份转录文本仍然成本高昂。SCLITE 等基于人工规则的方法 [11]支持全局映射文件，用于指定允许的替换项。不过，对于印度语言而言，定义此类映射尤其困难，因为这些语言的拼写形式存在大量差异。许多混合语码词有五种或更多拼写形式，因此穷举所有形式并不现实。WERd [12] 通过从社交媒体语料库中挖掘变体实现了自动化，但大多数印度语言都缺少此类资源。

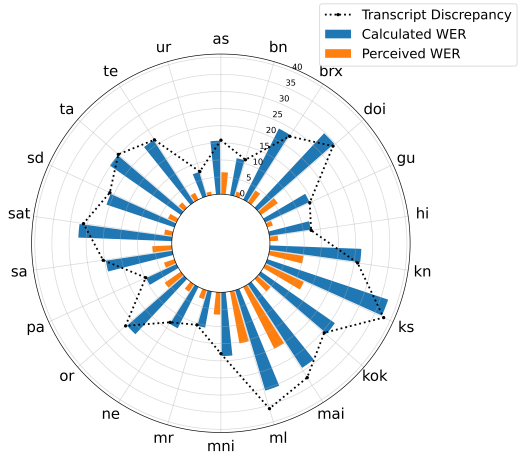

图1。对于印度语言，词错误率反映出虚高的数值，远高于感知到的错误（橙色和蓝色）。此外，两份有效转录文本之间存在较大差异（黑色）。

在印度语境下，[16] 指出，词错误率会对马拉雅拉姆语等形态丰富语言中的有效拼写变体过度惩罚。文本规范化 [17] 可以统一自动语音识别的输出；然而，[18] 报告称，此类方法往往会因删除关键元音符号而扭曲转录文本，导致词错误率得分虚低。作为替代方案，[19] 提出使用字符错误率（CER）；但这一指标可能会高估性能，因为即使只有一个字符错误，在印度语系键盘上也可能需要重新输入整个单词。类似地，poWER [20] 利用基于 WX 的发音方案，为语码转换同音词提出了一套框架；[13] 则通过拼写与分词规范化，为四种印度语言生成变体。这两种方法都高度依赖字素到音素转换器和转写工具等辅助资源。对于低资源印度语言，这类资源往往不完整或不一致。

3. 印度语言中虚高的词错误率

在提出解决方案之前，我们首先刻画因正字法变体而导致词错误率虚高的问题。为确定其范围与影响，我们评估了一个现有模型，以证明考虑正字法变体的问题真实且显著。为此，我们评估了一个在MahaDhwani [5] 数据集上微调的Canary [14] 模型，并使用涵盖四个不同语系、共22种印度语言的IndicVoices [2] 基准。在初始评估中，词错误率是对照原始人工生成的转录文本计算的，结果始终呈现较高的错误率（图1中的蓝色）。随后，我们请经验丰富的转录员对模型输出进行后编辑，并将这些人工校正的转录文本视为另一种黄金标准。当对照这一新参考重新计算时，词错误率大幅下降（图1中的橙色），因为转录员保留了模型预测结果中存在的有效正字法变体。理想情况下，对原始转录文本的忠实重建应使两份人工参考（即原始参考与后编辑参考）之间的词错误率为零。然而，在所有语言上取平均后，原始转录文本与后编辑转录文本之间的差异为20.4词错误率（图1中的黑色）。

基于对这些数据的分析，我们发现，这些令人悲观的词错误率数值背后主要有三个因素：（1）预测转录文本与参考转录文本之间允许存在拼写变体，（2）少数情况下存在人工转录错误，以及（3）确实存在音频歧义：除拼写差异之外，在缺乏更多上下文的情况下，可能存在多种有效的理解方式。在本研究中，我们聚焦于第一个因素，并提出一种在计算词错误率时考虑正字法变体的方法。

4. 创建正字法感知（OI）基准测试集的框架

本节提出一种创建正字法感知基准测试集的方案，以系统地涵盖拼写变体。

4.1. 识别变体类型

首先，我们采用以语言学为基础的方法，识别印度语言中可能出现的不同正字法变体。为此，我们咨询各语言的专家，并确定下文列出的变体类别。随后，我们提出一种人在回路的方法来生成变体。

马特拉和变音符号变体。马特拉是变音符号或元音符号，可改变辅音以表示不同的元音音值，并用于所有印度文字。为了进行稳健的评估，我们考虑元音和变音符号的所有替代形式，包括印地语中的努克塔（置于某些辅音下方的变音点）以及泰米尔语中的普利（用于标示纯辅音）。

借词的拼写变体。若干印度语言拥有来自英语、梵语、波斯语或其他语言的借词，这些借词往往有多种公认拼写。我们考虑此类单词的所有合理形式，包括转写和本地化版本。

复合词的拆分/合并。我们需要考虑一个复合词能否分解为其组成部分，以及两个独立的单词能否合并为一个单词，作为有效变体。这是马拉雅拉姆语等黏着语中常见的变体类型。

语音变体。我们考虑因地区口音、方言或发音而产生的拼写变体。

连字变体。部分印度文字允许将连续辅音连接成连字，或保持分开。我们同时考虑有连字和无连字的变体。

连声。连声规则支配印度语言中位于单词或语素边界处的语音合并与变换，并根据语音条件改变字母。这些规则有助于将复杂单词拆分为有意义的成分，并将较小单位合并为一个整体。我们考虑对单词应用连声规则后可能出现的所有变体。

逆文本规范化（ITN）。包含数字、日期、货币、单位及其他符号标记的口语话语，可以有多种合乎规范的书写形式。我们会考虑所有此类允许的逆文本规范化变体。

4.2. 大语言模型辅助生成单词变体

在许多情况下，单词可能的变体取决于句子的上下文，尤其是有关单词合并与拆分的规则。因此，我们不孤立地为单词生成变体，而是将音频的完整转录文本纳入考量，以生成上下文感知变体。

目标。给定一段音频转录文本 $T = (T_1, T_2, ..., T_K)$，它表示由 K 个单词组成的有序序列，目标是为给定的话语 T 生成一组变体 V。这些变体

表 1。原始转录文本示例（T）以及为不同单词生成的变体集合（V）（以蓝色标示）。虽然本研究聚焦于印度语言，但我们也加入了一个英语示例，以便更广泛的读者理解。

	“我能找到下载雇员公积金组织通行证的说明吗？T - 我账户号码五六八四九的对账单，*56849]*在蓝色目录中？”
“我能找到下载 [雇员公积金组织，雇员公积金V组织] [存折，存折] 我账户号码[五六八四九的对账单， 56849] 在蓝色 [的，彩色的] [目录，目录]”	

V表示为由 L 组成的有序序列，其中每个集合V_i对应于特定单词子序列（T_s、……、T_e）的一组变体，其中 s和 e分别是该子序列在 T中的起始和结束索引。我们允许将一个单词子序列视为单个变体，以涵盖单词合并的情况。请注意，按照设计 $L \leq K$。表 1 展示了T和 V的示例。**生成**。我们提示文本大语言模型为转录文本 T中的所有单词生成变体。为确保一致性，我们向大语言模型提供第 4.1 节所列的印度语言变体预定义类别，并为每个类别提供具有代表性的示例作为指南。此外，我们还指定 V的输出格式，并指示大语言模型以集合序列的形式生成变体，每个单词对应的变体之间用逗号分隔。**后编辑**。我们请人工专家审核生成的变体，并允许标注员完善与每个单词对应的集合。他们可以添加遗漏的变体或删除错误生成的变体，以确保准确性。我们使用 LabelStudio [21] 开发了一个定制网页界面，以便高效开展人工审核，让用户能够轻松添加或删除元素。

4.3. 正字法感知词错误率

我们修改了词错误率算法，以考虑单词变体。词错误率通过计算真实标注 T与模型预测 $P = (P_1, P_2, \ldots P_N)$之间的最小编辑距离来计算。我们通过将预测词与参考文本中的每一种可能变体进行比较，计算预测转录文本（P）与参考变体（V）之间的最小编辑距离。遵循 SCLITE工具包 [11], ，我们使用动态规划（DP）实现该算法。DP 表记录变体 V前 k 个片段与预测文本 P前 j 个单词对齐时的最小编辑距离。通过 DP 递推遍历所有对齐方式，算法能够确定 P与 V之间的最优匹配路径。我们将这一指标称为正字法感知词错误率（OIWER）。

5. 实验设置

正字法指导的 IndicVoices 基准的创建。我们将该框架应用于 IndicVoices [2] 基准，使用 Gemini-2.5-Pro 生成单词变体。我们从这 22 种语言中招募了 61 名母语转录员，并为他们提供指南以及每种类型的 5 个示例变体。表 2 展示了所创建话语、单词和变体的统计数据。不同语言中每个单词的变体数量介于 1.3 至 3.2 之间，克什米尔语等低资源语言的数值更高。此外，后编辑正字法变体每个话语仅需 1.2 分钟，而手动转录需要 2.4 分钟，这使得我们的方法相比多参考词错误率方法更加省时且更为全面。

表 2。正字法信息指导的 IndicVoices 基准的统计数据（utts：话语数，vars：变体数）

语言	# 话语数	# 单词数	# 变体数	语言	# 话语数	# 单词数	#变体数
as	3.0K	40.2K	89.1K	**曼尼普尔语**	1.7K	15.1K	27.5K
bn	2.2K	30.2K	61.0K	**mr**	2.3K	32.3K	76.0K
brx	1.8K	17.5K	28.5K	**ne**	2.5K	33.4K	73.3K
doi	0.5K	6.9K	18.7K	**or**	1.6K	19.6K	50.9K
gu	2.2K	34.0K	66.4K	**pa**	1.5K	24.2K	57.0K
hi	2.2K	39.7K	79.7K	**sa**	1.8K	19.0K	41.6K
kn	1.2K	12.8K	25.7K	**sat**	2.9K	34.8K	83.4K
ks	0.7K	7.3K	22.2K	**sd**	2.2K	33.6K	94.0K
kok	0.8K	8.3K	23.0K	**泰米尔语**	2.7K	31.5K	97.4K
五月	2.5K	33.1K	90.7K	**te**	2.6K	29.1K	92.7K
ml	1.2K	13.7K	24.9K	**ur**	0.2K	3.4K	4.6K

对比的自动语音识别系统。我们的研究考察了以下自动语音识别系统，涵盖开源模型以及面向印度语言的公开可用 API。

Canary [14] -C: 我们在 MahaDhwani [5] 数据集的伪标注转录文本上微调了 Canary-1B-flash [14] 模型。该数据集包含来自 22 种印度语言的 20 万小时数据。

IndicConformer [5] -IC: 由 AI4Bharat 发布的开源多语言 Conformer 模型，支持 22 种印度语言。*Saarika:v2*： 由 SarvamAI 发布的公开语音转文本 API，支持 10 种印度语言。

Meta MMS-1B-all [22]： Meta 推出的开源大规模多语言语音模型，支持 14 种印度语言。

Azure： Microsoft Azure 提供的批量转录 API（版本：2024-06-01），支持 10 种印度语言。

Gemini-2.5-Pro [23]： 谷歌的 Gemini-2.5-Pro 模型，通过传入以下提示词使用：“Transcribe the audio in the <lang> language”，其中 <lang> 是语言名称。

6. 结果与讨论

6.1. 使用 OIWER 评估自动语音识别模型

表3显示，在所有模型和语言中，整体 OIWER 均低于词错误率。这一结果符合预期，也与我们作为自动语音识别从业者对模型质量的感知，以及下游用户对感知质量的反馈一致，如图1所示。接下来我们观察模型之间的比较结果：词错误率表明，Gemini 平均比 Canary 差 18.1 个百分点。然而，OIWER 显示二者之间的差距较小，仅为 11.5 个百分点，从而更准确地反映了模型性能。这一点在泰米尔语中尤为明显：使用词错误率时，差距为 17.4 个百分点，而使用 OIWER 后降至 4.3 个百分点。这表明词错误率可能会夸大性能差距，进而造成进展显著的假象。Azure 的模型也呈现出类似趋势，证实与标准词错误率相比，OIWER 能够减轻过度惩罚。

此外，我们观察到不同语言的评估结果存在差异。对于马拉雅拉姆语和泰米尔语，在所有模型上的平均值中，词错误率与 OIWER 的差值分别为 13.9 和 14.4。这表明变体对这两种语言的影响更大。另一方面，对于正字法标准化程度较高的语言，如古吉拉特语和马拉地语，平均差值较小，分别为 5.4 和 5.3。我们还观察到，尽管印地语的拼写已标准化，其平均差值仍为 5.8，这归因于大量借词导致了更多变体。

表3。自动语音识别系统在词错误率和 OIWER 上的比较。绿色的深浅表示词错误率与 OIWER 之间的差异。受篇幅限制，我们展示排名前11的语言的结果，并在“其他”列中报告其余11种语言的平均值。

	bn		gu		hi		kn		ml		mr		or		pa		ta		te		ur		其他（平均值）	
	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER	词错误率	OIWER		
-C	11.2	7.7	14.6	9.7	13.2	8.9	24.3	16.6	30.1	16.0	11.7	8.1	19.7	11.4	10.7	7.9	27.0	17.2	22.9	13.7	5.4	3.0	23.2	17.0
-IC	13.4	9.6	16.9	11.6	14.8	10.2	27.0	18.9	32.2	18.8	14.2	10.3	20.6	12.1	12.5	9.8	30.1	19.8	25.5	16.5	6.4	3.9	23.5	17.7
	22.2	12.8	18.6	12.3	19.5	11.6	40.9	27.2	37.0	20.8	19.1	12.8	30.3	16.2	21.2	12.7	44.4	21.5	33.4	19.4	10.2	6.3	49.7	35.1
	25.2	13.8	23.3	17.4	19.9	13.3	37.0	23.3	40.3	24.0	21.3	15.1	31.9	19.3	21.8	15.1	44.1	23.5	33.5	20.6	-	-	-	-
	34.8	21.0	49.1	43.8	22.2	15.4	53.4	43.6	53.5	37.1	35.9	29.3	38.1	25.7	38.9	31.6	61.4	43.6	52.4	40.5	18.1	13.0	-	-
	47.2	40.7	48.3	43.6	43.2	38.8	67.6	62.6	68.5	61.7	53.6	48.6	54.7	45.1	54.3	48.9	79.1	74.0	68.5	62.0	26.5	22.7	-	-

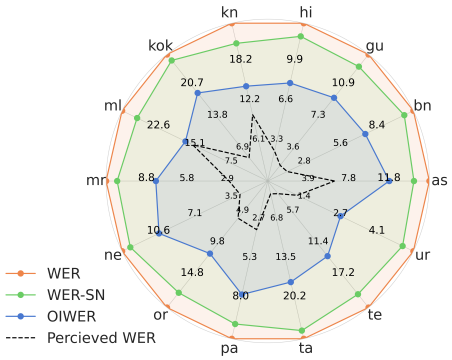

图2。 OIWER 与各语言的感知词错误率最为吻合 语种。

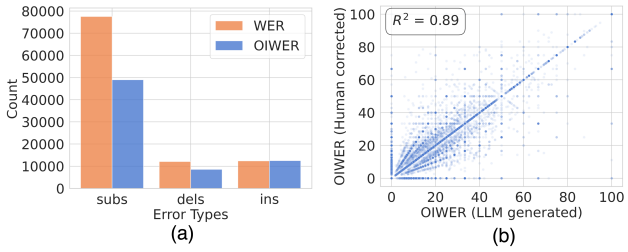

图 3。（a）OIWER 消除了错误替换。（b）使用大语言模型生成变体的 OIWER 与人工校正变体呈现出较强的相关性。

6.2. OIWER 是否能更好地反映人类感知的词错误率?

我们将本方法与 [13], 进行比较；该方法引入了拼写与分词规范化（WER-SN），以应对印度语言中的正字法变体。虽然他们的工作依赖于内部开发的字素转音素（G2P）系统，但我们采用开源的 eSpeak-ng，以确保结果可复现。图 2 展示了 Canary（-C）在词错误率、WER-SN 和 OIWER 指标下的结果。WER-SN 减少了拼写变体带来的惩罚，相较于标准词错误率，平均提升了 1.5 个点。然而，OIWER 提供了更准确的性能评分；与 WER-SN 相比，它使评分与感知词错误率之间的差距平均缩小了 4.9 个点。感知词错误率（由人工感知）通过让人工对模型输出进行后编辑，再将后编辑后的转录文本与模型输出进行比较来计算（如第 3 节所述）。进一步分析发现，WER-SN 未能捕捉借词的拼写变体，也遗漏了若干口语词形。这表明，尽管基于规范化的方法能够捕捉某些合理的变体，OIWER 对印度语言正字法多样性的呈现更为全面。理想情况下，OIWER 与感知词错误率应当接近。我们确实在马拉雅拉姆语中观察到了这一点，两者相差 1.5 个词错误率点。然而，各语言的平均差距仍为 6.9 个词错误率点，泰卢固语和孔卡尼语的差距尤其大。经检查，我们发现这一差距源于音频中固有的歧义，若无更多上下文便无法消除（第 3 节提到的三个因素之一）。

6.3. 变体能否减少替换错误?

我们计算了所有语言中 WER 和 OIWER 的替换、插入和删除错误总数，如图 3(a) 所示。分析表明，尽管删除和插入错误的数量相近，但 OIWER 中的替换错误显著减少。具体而言，引入变体后，替换错误减少了 28.5K（所有语言合计）。这一减少表明，纳入变体能够减少错误替换，验证了我们方法的有效性。

6.4. 大语言模型生成变体能否作为人工校正变体的替代方案?

我们为 IndicVoices 基准测试集创建了两种结合正字法信息的变体：一种仅采用大语言模型建议的变体（OIWER LLM），另一种采用经人工转录员验证的变体（OIWER Human）。随后，我们分别计算两个基准测试集中每条话语的 OIWER，以探究能否通过使用大语言模型生成的变体来减少对人工干预的依赖，从而降低时间和成本。本分析涵盖所有语种的 40.3K 条话语。图 3(b) 中的散点图显示，两个基准测试集中的 OIWER 值具有较强的相关性，决定系数为 0.89。这表明，大语言模型生成变体可以作为人工校正变体的可靠替代方案。

7. 结论

我们提出了一个全面的框架，用于对印度语言的自动语音识别系统进行正字法知情（OI）评估，并纳入语言专家核实的全部正字法变体。我们引入了正字法知情词错误率（OIWER），以计入这些变体，从而更稳健、准确地反映人类感知对模型性能的评估。我们的分析揭示了若干关键发现：（i）在所有模型和语言中，OIWER 的得分始终低于词错误率（WER），因此能够更真实地评估自动语音识别模型；（ii）OIWER 减少了替换错误带来的过度惩罚，并且更符合人类感知；（iii）大语言模型生成变体可作为人工校正变体的可靠替代方案，在保持评估质量的同时节省时间和成本。

8. 参考文献

[1] Tahir Javed、Kaushal Santosh Bhogale、Abhigyan Raman、Anoop Kunchukuttan、Pratyush Kumar 和 Mitesh M. Khapra，《IndicSUPERB：印度语言语音处理通用性能基准》，*CoRR*，第 abs/2208.11761 卷，2022 年。

[2]Tahir Javed、Janki Nawale、Eldho Ittan George、Sakshi Joshi、Kaushal Santosh Bhogale、Deovrat Mehendale、Ishvin- der Virender Sethi、Aparna Ananthanarayanan、Hafsah Faquih、Pratiti Palit、Sneha Ravishankar、Saranya Sukumaran、Tripura Panchagnula、Sunjay Murali、Kunal Sharad Gandhi、Ambu- javalli R、Manickam K. M、C. Venkata Vaijayanthi、Krish- nan Srinivasa Raghavan Karunganni、Pratyush Kumar与Mitesh M. Khapra，“Indicvoices：面向构建包容性的印度语言多语言语音数据集”，载于计算语言学协会研究成果，*ACL 2024*，*16*，*2024*，*11* 曼谷，泰国及线上会议，*8*月 -Lun-Wei Ku、Andre Martins与Vivek Srikumar编，2024，第10740–10782页，计算语言学协会。[3]Tahir Javed、Sumanth Doddapaneni、Abhigyan Raman、Kaushal Santosh Bhogale、Gowtham Ramesh、Anoop Kunchukuttan、Pratyush Kumar与Mitesh M Khapra，“面向为下十亿用户构建自动语音识别系统”，载于*AAAI*人工智能会议论文集，2022，第36卷，第10813–10821页。[4] Kaushal Santosh Bhogale、Sai Sundaresan、Abhigyan Raman、Tahir Javed、Mitesh M. Khapra与Pratyush Kumar，“Vistaar：面向印度语言自动语音识别的多样化基准测试集与训练集”，载于国际语音通信协会第*24*届年会，*Interspeech 2023*，爱尔兰都柏林，*2023*年*8*月*20-24*日，Naomi Harte、Julie Carson-Berndsen与Gareth Jones编，2023，第4384–4388页，ISCA。[5] AI4Bharat，“IndicConformer”，2024，访问日期：2025-02-19。[6] Vaibhav Srivastav、Somshubra Majumdar、Nithin Koluguri、Adel Moumen、Sanchit Gandhi等，“开放自动语音识别排行榜”，https://huggingface.co/spaces/hf-audio/open_asr_leaderboard，2023。[7] Ah med M. Ali、Walid Magdy、Peter Bell与Steve Renals，“用于评估无正字法规则语言自动语音识别的多参考词错误率”，载于*2015*年*IEEE*自动语音识别与理解研讨会，*ASRU 2015*，美国亚利桑那州斯科茨代尔，*2015*年*12*月*13-17*日。2015，第576–580页，IEEE。[8] Ahmed Ali、Salam Khalifa与Nizar Habash，“面向抗变异性的方言语音评估”，载于国际语音通信协会第*20*届年会，*Interspeech 2019*，*19*，*2019*，*15*2019，336奥地利格拉茨，*9*月 -Gernot Kubin与Zdravko Kacic编，第–340页，ISCA。[9] Annika Heuser、Tyler Kendall、Miguel Del Rio、Quinten Mc- Namara、Nishchal Bhandari、Corey Miller与Mig¨uel Jett´e，“对人工与自动语音识别生成的非裔美国人英语转录文本中风格差异的量化”，*CoRR*，第abs/2409.03059卷，2024。

[10]Quinten McNamara、Miguel ´Angel del R´ıo Fern´andez、Nishchal Bhandari、Martin Ratajczak、Danny Chen、Corey Miller 和 Mig¨uel Jett´e，《使用多份参考转录文本对自动语音识别进行风格无关的评估》，*CoRR*，第 abs/2412.07937 卷，2024 年。[11] Ronald Rosenfeld 和 Philip Clarkson，《Cmu-cambridge 统计语言建模工具包 v2》，1997 年。[12] Ahmed Ali、Preslav Nakov、Peter Bell 和 Steve Renals，《WERD：使用社交文本拼写变体评估方言语音识别》，载于*2017* 年 *IEEE* 自动语音识别与理解研讨会 *(ASRU)*。IEEE，2017 年，第 141–148 页。[13] Satarupa Guha 、 Rahul Ambavat 、Ankur Gupta、Manish Gupta 和 Rupeshkumar Mehta，《无监督、语言无关的词错误率标准化》，*arXiv* 预印本 *arXiv:2303.05046*，2023 年。[14] Piotr ˙Zelasko、Kunal Dhawan、Daniel Galvez、Krishna C Puvvada、Ankita Pasad、Nithin Rao Koluguri、Ke Hu、Vitaly Lavrukhin、Jagadeesh Balam 和 Boris Ginsburg，《编码器-解码器语音模型的训练与推理效率》，*arXiv* 预印本 *arXiv:2503.05931*，2025 年。[15] Shigeki Karita、Richard Sproat 和 Haruko Ishikawa，《宽松评估日语语音识别：对自然产生的拼写不一致进行建模》，*CoRR*，第 abs/2306.04530 卷，2023 年。[16] Kavya Manohar、A. R. Jayan 和 Rajeev Rajan，《使用子词词元进行语言建模，改进针对形态复杂的马拉雅拉姆语的语音识别系统》，*EURASIP J. Audio Speech Music. Process.*，第 2023 卷，第 1 期，第 47 页，2023 年。[17] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever，《通过大规模弱监督实现稳健的语音识别》，载于机器学习国际会议，*ICML 2023*，*2023* 年 *7* 月 *23–29* 日，美国夏威夷州檀香山，由 Andreas Krause、Emma Brunskill、Kyunghyun Cho、Barbara Engelhardt、Sivan Sabato 和 Jonathan Scarlett 编，2023 年，第 202 卷，机器学习研究论文集，第 28492–28518 页，PMLR。[18] Kavya Manohar 和 Leena G. Pillai，《规范化过程中失去了什么？探究多语言自动语音识别模型评估中的陷阱》，*CoRR*，第 abs/2409.02449 卷，2024 年。[19]Thennal D. K、Jesin James、Deepa P. Gopinath 和 Muhammed Ashraf K，《倡议在多语言自动语音识别评估中采用字符错误率》，*CoRR*，第 abs/2410 07400 卷，2024 年。。[20] Brij Mohan Lal Srivastava 和 Sunayana Sitaram，《代码转换语音识别中的同音词识别与合并》，载于*Interspeech2018*，2018 年，第 1943–1947 页。[21] Maxim Tkachenko、Mikhail Malyuk、Andrey Holmanyuk 和 Nikolai Liubimov，《Label Studio：数据标注软件》，2020–2025 年，开源软件，网址：https://github.com/HumanSignal/label-studio。[22] Vineel Pratap、Andros Tjandra、Bowen Shi、Paden Tomasello、Arun Babu、Sayani Kundu、Ali Elkahky、Zhaoheng Ni、Apoorv Vyas、Maryam Fazel-Zarandi、Alexei Baevski、Yossi Adi、Xi- aohui Zhang、Wei-Ning Hsu、Alexis Conneau 和 Michael Auli，《将语音技术扩展至 1, 种语言》，*J. Mach. Learn.Res.*，第 25 卷，第 97:1–97:52 页，2024 年。[23] Rohan Anil 等，《Gemini：一系列能力强大的多模态模型》，*CoRR*，第 abs/2312.11805 卷，2023 年。