# 国家标准《信息技术 藏语语音识别数据标注规范》（征求意见稿）编制说明

## 一、工作简况

### （一）任务来源

根据《国家标准委关于下达2026年第六批推荐性国家标准计划及相关标准外文版计划的通知》（国标委发〔2026〕40号），由全国信息技术标准化技术委员会提出并归口的推荐性国家标准《信息技术 藏语语音识别数据标注规范》制定项目（计划号为20263257-T-469）正式立项。

**注：**根据标准实际内容范围，计划名称《信息技术 藏语语音识别数据标注规范》拟申请于报批阶段变更为《信息技术 藏语语音识别数据标注》。

本标准由西藏大学主办，起草单位包括：西藏大学、电子科技大学、中国电子技术标准化研究院、青海师范大学、中国科学院声学研究所。

本标准主要起草人：尼玛扎西、于永斌、仁增多杰、黄姗姗、陈晓研、洛桑嘎登、王向向、头旦才让、班玛宝、仁青东主、赵庆卫、群诺、范满平、朱宇雷、冯箫、盛天颖、张子玥、道吉扎西、拉毛吉、钱木吉、蔡郁青、王超、李臣波、郑智文、刘羽童、张瑾、才让卓玛。

标准起草组人员组成和分工见表1。

**表1 起草组人员组成和职责分工表**

| 序号 | 起草人姓名 | 所在单位 | 工作内容 |
|---|---|---|---|
| 1 | 尼玛扎西 | 西藏大学 | 总体技术负责人 |
| 2 | 于永斌 | 电子科技大学 | 总体组织与协调 |
| 3 | 仁增多杰 | 西藏大学 | 技术内容编制 |
| 4 | 黄姗姗 | 中国电子技术标准化研究院 | 组织与协调 |
| 5 | 陈晓研 | 中国电子技术标准化研究院 | 组织与协调 |
| 6 | 洛桑嘎登 | 西藏大学 | 技术内容编制 |
| 7 | 王向向 | 电子科技大学 | 技术内容编制 |

| | | | |
|---|---|---|---|
| 8 | 头旦才让 | 青海师范大学 | 藏语语言学专业支撑 |
| 9 | 班玛宝 | 西藏大学 | 技术内容编制 |
| 10 | 仁青东主 | 西藏大学 | 技术内容编制 |
| 11 | 赵庆卫 | 中国科学院声学研究所 | 技术内容编制 |
| 12 | 群诺 | 西藏大学 | 藏语语言学专业支撑 |
| 13 | 范满平 | 电子科技大学 | 技术内容编制 |
| 14 | 朱宇雷 | 西藏大学 | 技术内容编制 |
| 15 | 冯箫 | 电子科技大学 | 试验验证与数据支撑 |
| 16 | 盛天颖 | 电子科技大学 | 试验验证与数据支撑 |
| 17 | 张子玥 | 电子科技大学 | 试验验证与数据支撑 |
| 18 | 道吉扎西 | 西藏大学 | 藏语语言学专业支撑 |
| 19 | 拉毛吉 | 西藏大学 | 技术内容编制 |
| 20 | 钱木吉 | 西藏大学 | 技术内容编制 |
| 21 | 蔡郁青 | 电子科技大学 | 技术内容编制 |
| 22 | 王超 | 青海师范大学 | 试验验证与数据支撑 |
| 23 | 李臣波 | 电子科技大学 | 试验验证与数据支撑 |
| 24 | 郑智文 | 电子科技大学 | 试验验证与数据支撑 |
| 25 | 刘羽童 | 电子科技大学 | 试验验证与数据支撑 |
| 26 | 张瑾 | 电子科技大学 | 试验验证与数据支撑 |
| 27 | 才让卓玛 | 西藏大学 | 文稿校核人员 |

**（二）制定背景**

藏语作为我国藏族人民的主要语言，其语音标注规范的制定对于推动藏语语音技术发展、传承保护少数民族语言文字具有重要意义。

当前，藏族地区经济社会正处于转型升级的关键阶段，信息化技术的普及与应用成为提升生活质量、促进产业发展的重要手段。从全球来看，语音技术产业正在快速发展，人工智能和大数据技术的进步为语音识别、语音合成等领域提供了强有力的支持。我国在语音技术领域已具备较强的研发能力和产业基础，市场规模持续扩大。作为少数民族语言信息化的一部分，藏语语音技术具有独特的市场需求和发展潜力，尤其在教育、文化旅游和智能服务等领域，展现出广阔的应

用前景。制定《信息技术 藏语语音识别数据标注》国家标准，能够为藏语语音识别、语音合成及相关应用提供统一的技术基础，满足藏族群众在教育、医疗、公共服务等领域的实际需求，推动藏族地区经济社会和信息产业的协同发展。

我国政府高度重视少数民族语言文字的保护与发展，出台了一系列相关法律法规和政策规划。例如，《中华人民共和国民族区域自治法》明确规定保障少数民族使用和发展本民族语言文字的权利；《国家通用语言文字法》强调促进少数民族语言文字的规范化、标准化和信息化；此外，《西藏自治区国民经济和社会发展第十五个五年规划纲要》中也提出要加强少数民族语言文字教育和保护传承，实施藏语言大模型建设应用工程。制定《信息技术 藏语语音识别数据标注》国家标准，符合国家关于少数民族语言文字保护的政策导向，是落实相关法律法规的具体实践，有助于推动藏语的信息化进程，传承和弘扬藏族优秀传统文化。

**（三）起草过程**

2024年11月：起草核心人员编制完成标准草案预研稿。

2024年12月~2025年2月：全国信息技术标准化技术委员会字符集与编码分技术委员会（SAC/TC28/SC2）协助修改完善标准草案，准备全套标准立项材料。

2025年3月20日：参加全国信息技术标准化技术委员会组织召开的立项内审会，同意立项。会后，起草核心人员按内审意见完善标准草案。

2025年7月11日：经全国信息技术标准化技术委员会下属藏文信息技术国家标准工作组审查，同意以此方案提出立项。

2025年10月：按照申请标准立项的要求，起草核心人员编制完善标准草案，由中国电子技术标准化研究院提出立项。

2026年6月27日：国家标准化管理委员会下达本标准制定计划。

2026年7月：西藏大学牵头成立本标准编制工作组，成员来自高校和科研院所。起草组成员信息导入“标准智能研制系统”，分配系统账号。

2026年8月：编制形成工作组讨论稿（第1稿），经全国信息技术标准化技术委员会组织专家审查，同意开题，标准名称拟修改为《信息技术 藏语语音识别数据标注》（于报批前提交更名申请）。会后，根据审查专家意见，就涉及数据安全的技术内容征求全国网络安全标准化技术委员会SAC/TC 260有关专家建议。

2026年9月：由起草组以线上会议形式组织召开标准编制讨论会，对标准草

案进一步修改完善。起草组成员按照任务分工，通过“标准智能研制系统”对标准文本各章节协同编写，确保文本表述清晰、严谨，格式符合GB/T 1.1要求。

2026年9月底：形成征求意见稿初稿。

2026年10月14日：全国信标委组织召开内审会，审议通过标准征求意见稿初稿，并提出修改意见。

2026年10月至11月：由起草组以线上会议形式组织召开标准编制讨论会，根据内审意见，完善形成征求意见稿和征求意见稿编制说明。

## 二、国家标准编制原则、主要内容及其确定依据

### （一）标准编制原则

**科学性与实用性统一原则：**以藏语语音识别技术研发和产业落地的实际需求为核心，充分适配藏语卫藏、安多、康巴三大方言的语音特征，保证标注规则清晰可执行，标注结果准确一致，能够为小资源藏语语音识别模型开发提供高质量的标准化数据基础。

**体系衔接兼容原则：**编制过程与现有藏语文信息技术领域国家标准、藏语文信息技术标准体系整体要求保持协调，统一数据标注规则与已有的藏文信息处理、语音数据采集相关标准的技术要求，完善藏语人工智能领域标准体系架构。

**通用与场景适配结合原则：**标注规则既满足通用藏语语音识别数据的标注要求，同时覆盖新闻、生活服务、日常对话等多类实际应用场景，兼顾不同场景下藏语口语、书面语语音的标注需求，提升标准的适用范围。

**安全合规原则：**遵循国家数据安全、个人信息保护相关规范要求，明确数据标注全流程的安全管理规则，防范标注过程中的数据安全风险，保障藏语语音数据的合法合规使用。

### （二）标准主要内容

本标准规定了信息技术领域藏语语音数据标注的要求。本标准适用于藏语卫藏方言、康巴方言和安多方言的自然语言处理、语音翻译、同声传译、实时翻译等语音处理过程。

本标准的主要技术内容涵盖了藏语语音数据标注的全过程，包括流程设计、前期准备和任务执行，具体如下：

（1）藏语语音数据标注流程

- 藏语语音数据标注流程：确立了藏语语音数据从采集到标注的整体框架，确保标注工作的系统性和连贯性。详细规定了标注的具体步骤，包括任务分配、标注执行和质量检查等，为标注工作提供清晰的操作路径。

（2）标注任务前期准备要求

- 待标注数据的采集及确定：明确了待标注数据的来源、采集方法和质量要求，确保数据输入的可靠性和一致性。

- 语音数据标注任务：定义了标注任务的类型，包括语音时间边界定位和文件标注等。提供了任务概述及具体要求，作为标注工作的基础指导。

- 标注规则：详细描述了各层级的标注要求。全局（GLOBAL）层：标注全局信息，如录音环境或文件属性。说话人（SPEAKER）层：标注说话人相关信息，如身份或特征。内容（CONTENT）层：标注语音内容的具体细节。

- 标注需求变更流程：规定了当数据需求方提出需求变更时应遵循的流程，确保需求管理的有效性。

- 标注环境要求：规定了标注工具的选择标准和标注场地的环境要求，确保工作条件符合技术需求。

（3）标注任务执行过程的要求

- 标注任务创建要求：包括任务包和任务账户的创建方法，确保任务管理的规范性。

- 标注任务分发要求：规定了任务分配的具体流程，提高执行效率。

除上述主要内容外，本标准另设资料性附录“标注人员管理要求”，明确了标注人员的选拔范围、培训要求及能力档案管理，确保人员素质达标。

**（三）标准主要内容的确定依据**

**国家语言文字及信息技术相关标准体系：**标准主要内容的设置，应以国家现行的语言文字规范、信息技术类基础标准和语音技术相关标准为重要依据，参考其在术语定义、数据结构、编码规则、元数据描述等方面的通用要求，确保本标准在国家标准体系中的一致性和衔接性。

**现有藏语语音资源建设与工程实践经验：**充分总结科研机构、高等院校、企

业和行业在藏语语音识别、语音合成、语音标注等项目中的实践经验，吸收已有语料库建设中行之有效的标注方案、质量控制方法和工具使用经验，作为确定标注层次、标注粒度和操作流程的重要参考。

**数据质量控制与评测要求：**以构建高质量可重复利用的藏语语音语料资源为目标，结合数据质量评价指标、标注一致性评估方法和抽检机制要求，明确本标准需要覆盖的质量控制相关内容，包括标注规范性要求、检查流程及误差记录等。

**藏语信息化与智能应用需求：**围绕藏语语音识别、语音合成、语音检索、人机交互等重点应用场景的功能需求，分析系统性能瓶颈与数据需求特征，确定与应用效果紧密相关的标注项目和精度要求，突出工程实用性和面向产业应用的导向。

**信息安全、隐私保护及相关法规：**在遵循国家关于数据安全、个人信息保护、民族语言文字工作等法律法规和政策要求的前提下，确定与说话人信息、语料采集环境、数据发布与共享相关的标注范围和敏感信息处理原则，确保标准在实际使用中的合规性和安全性。

**可扩展与持续完善的需求：**综合考虑藏语应用场景未来扩展（如情感语音、多说话人会话、跨语种混合语音等）以及新技术发展趋势，对需要预留扩展空间或定义扩展字段的内容予以原则性规定，使标准具备后续修订和拓展的基础。

## 三、试验、验证情况分析

### （一）试验情况

（1）标注规则适用性验证

试验重点验证本规范针对藏语三大方言发音差异、口语/书面语表达差异的规则适配性。试验显示：本规范针对藏语方言发音变体、口语吞音变音、新词古词转写、语音片段对齐的规则清晰明确，可指导标注人员快速完成不同类型藏语语音的标注，三大方言样本均实现完整标注，未出现规则空白或歧义问题，适配性满足行业应用预期要求。

（2）标注结果技术价值验证

试验采用标注完成的数据集分别训练端到端藏语语音识别模型、基于大模型微调的藏语方言识别模型，对比相同规模非规范标注数据训练的模型性能：试验结果显示，使用本规范标注数据训练的模型，可降低藏语三大方言平均词错误率，

有效支撑藏语语音识别技术落地，验证了规范的技术价值。

**（二）预期的经济效益**

本标准将为藏语语音技术的研究和产业化应用奠定基础，促进智能语音设备、语言学习软件、语音翻译工具等相关产业的发展。这不仅能够催生新的经济增长点，还能为藏族地区的中小企业提供技术支持，增强其市场竞争力。

**（三）预期的社会效益**

本标准的制定和推广将提升藏语语音技术的可用性和普及度，使藏族群众更便捷地使用现代信息技术，缩小数字鸿沟。同时，通过规范化的语音标注，藏语的教学和传承将更加高效，有助于提升藏族群众的文化自信和民族认同感，推动社会和谐发展。

## 四、与国际、国外同类标准技术内容的对比情况

本标准技术内容为我国独创，无同类国际标准、国外标准。

当前国内外尚未形成针对藏语语音标注的专项标准体系，因此本标准未采用国际标准。国际上，英语、汉语等大语种已建立相对成熟的语音标注规范，如英语领域广泛应用的 ISO 246 系列语言资源管理标准（ISO 24617-1 语义标注框架），以及汉语领域的 GB/T 21023-2007《中文语音识别系统通用技术规范》。然而，针对藏语语音标注方面，仍存在三大方言标注规则、语音数据采集、时间边界定位、噪声标注等关键技术环节的统一规范仍属空白等问题。现有研究多集中于学术机构零散实践，缺乏可推广的标准化标注框架，制约藏语智能语音技术研发与产业化应用。本标准的制定将填补该领域标准空白，为全球藏语语言资源建设提供中国方案。

## 五、以国际标准为基础的起草情况

无同类的国际、国外标准，本标准未采用国际、国外标准。

## 六、与有关法律、行政法规及相关标准的关系

本标准未替代或废止现行相关标准，与现行相关法律、法规、规章和强制性国家标准协调一致。

本标准规范性引用了GB/T 13000—2025《信息技术 通用编码字符集（UCS）》，规定在藏语语音标注应采用GB/T 13000的字符编码体系，以保障藏文字符在跨系统、跨国界交换传输中的一致性；规范性引用了GB/T 45674—2025《网络安全技术 生成式人工智能数据标注安全规范》，以保障待标注数据的安全性。

本标准在起草过程中还参考了以下我国国家标准：

- GB/T 42755—2023 人工智能 面向机器学习的数据标注规程

**七、重大分歧意见的处理经过和依据**

无。

**八、涉及专利的有关说明**

本标准未涉及专利、版权和著作权等知识产权问题。

**九、实施国家标准的要求以及措施建议**

建议本标准作为推荐性国家标准发布，自发布之日起6个月实施。在标准实施中，建议及时在行业内及相关单位开展标准的宣贯，使相关单位能够学习并应用本标准，并结合实际应用要求贯彻实施标准。

**十、其他应当说明的事项**

无。

国家标准《信息技术 藏语语音识别数据标注规范》编制工作组

2026-11-18