专题 大模型训练数据

XBMU-bo-Lhasa31：藏语拉萨话语音识别数据集

马立克[1]，李冠宇[1,2*]，谢晨宇[1]，孙倩[1]，郭玉豪[1]

1. 西北民族大学语言与文化计算教育部重点实验室，兰州 730030
2. 青藏高原人文环境数据智能实验室，兰州 730000

摘要：藏语语音识别在藏语教育、新闻传播等领域具有重要应用价值。藏语拉萨话广泛使用于拉萨市及周边地区，由于地域等因素的影响，当前可用的藏语语音数据资源匮乏，高质量标注数据稀缺。为此，本研究构建了一个专业规范的藏语拉萨话语音识别数据集。数据集使用自制录音软件实地录制，采集自 51 位说话人，总时长 31.61 小时，包含 24,289 条语音样本，平均每条时长 4.68 秒。数据内容主要选自新闻领域文本，确保语言规范性和领域代表性。为保障数据质量，实施了严格的质量控制流程：首先，对原始文本进行分句处理和人工校验；其次，在录音完成后，采用语音端点检测（VAD）技术筛选优质录音样本；最后，对文本中的非发音符号进行规范化处理，以提高语音识别的准确性。本数据集的建立为藏语语音识别研究提供了重要基础资源，对推动藏语语音识别技术发展具有积极意义。

关键词：语音识别；藏语拉萨话；多说话人；语音语料库

数据库（集）基本信息简介

数据库（集）名称	XBMU-bo-Lhasa31：藏语拉萨话语音识别数据集
数据通信作者	李冠宇（xxlgy@xbmu.edu.cn）
数据作者	马立克、李冠宇、谢晨宇、孙倩、郭玉豪
数据时间范围	2020年
地理区域	拉萨及其周边地区
数据量	2.69 GB
数据格式	*.wav, *.txt
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01539
基金项目	国家自然科学基金（61633013）；2024年甘肃省科技重大专项计划（24ZDFA004）。
数据库（集）组成	数据集由音频文件、文本文件和说明文件组成。其中（1）wav是音频文件夹，其下按照说话人分成51个子文件夹，总时长31.61个小时，包含24,289条语音样本，平均每条时长4.68秒。（2）transcript文件中的文本与音频一一对应，其中所有文本数据都来源于新闻领域，并对文本中的非发音符号进行规范化处理。（3）readme.txt文件包含数据集的一些基础信息。（4）resource_lexicon.txt是发音词典文件。

www.sciengine.com/CSD
ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2025.0122.zh

文献 DOI：
10.11922/11-6035.csd.2025.0122.zh

数据 DOI：
10.57760/sciencedb.j00001.01539

文献分类：计算机科学技术

收稿日期：2025-06-23
录用日期：2025-11-14
发表日期：2026-03-09

* 论文通信作者
李冠宇：xxlgy@xbmu.edu.cn

引　言

藏语作为我国重要的少数民族语言之一，其语音识别技术的发展迅速[1]，对于促进民族地区教育信息化、新闻传播智能化以及跨语言交流具有重要意义。近年来，端到端语音识别技术（如Whisper[2]）的快速发展，显著降低了对精细标注数据的依赖，仅需音频和对应文本即可训练出高性能的语音识别系统。这一技术突破为资源稀缺语言的语音识别研究提供了新的机遇[3]。

拉萨话作为藏语三大方言中使用最广泛的变体，在西藏自治区及周边地区具有重要的社会文化影响力。然而，现有的藏语语音数据资源仍面临严重匮乏的问题。尽管端到端语音识别技术降低了对精细标注数据的依赖，但高质量、领域特定的数据集在数据量有限的情况下，仍然是提升系统性能的关键因素[4]。特别是在资源稀缺的语言场景中，精心构建的指定领域数据集可以弥补数据规模的不足，显著提高模型的识别效果[5]。新闻文本的规范性和标准性为语音识别提供了理想的语言素材，该领域的广泛应用场景确保了数据集有更广泛的使用空间。

为此，本研究构建了一个专业规范的藏语拉萨话语音识别数据集。本数据集采用自主研发的录音采集系统，由拉萨籍的同学在安静录音室完成录制，共包含 51 位说话人的 24,289 条语音样本，总时长达到 31.61 小时。文本使用经过分句处理和人工校验的文本，由说话人自行朗读并进行纠错。录音完成后采用语音端点检测（VAD）[6]技术过滤质量较低的音频。同时对文本进行规范化处理，删除不发音以及反写的字母，确保与音频的准确对应[7]。

与 TIBMD@MUC[8]和 Nict-tib1[9]等藏语数据集相比，本数据集在录音环境控制和质量把控方面具有显著优势。所有语音数据均在专业录音室环境下采集，确保音频纯净度高、背景杂音小。在后期处理中，对所有语音进行端点检测，切除过长的静音片段。同时，文本数据经过多轮人工校对和标准化处理，确保语音-文本对齐的准确性。这些方法使得本数据集在数据质量上超过了原有的开源数据集。

本数据集的建立在藏语教育、新闻传播等领域具有重要应用价值，本数据集可为基于端到端深度学习的藏语语音识别算法研究提供高质量的训练数据，可以有效改善当前藏语语音识别研究中面临的数据匮乏问题[10]。其次，数据集中的标准发音和规范文本能够为藏语语音合成系统的开发提供重要支持，助力藏语智能语音交互应用的实现[11]。此外，本数据集还可作为基础研究资源，促进藏语多方言比较研究以及藏汉/藏英等多语种语音处理技术的发展[12]。

1　数据采集和处理方法

XBMU-bo-Lhasa31 数据集是一个单通道藏语拉萨话语音识别数据集，包含音频文件、对应文本、发音词典及元数据说明。基于新闻文本的规范性和应用广泛性，本研究选用新闻领域作为核心语料来源。数据集采集自 51 位西北民族大学拉萨籍学生（年龄 20–25 岁），在专业录音室环境下通过移动端应用完成录制，每位说话人平均贡献 480 条语音样本。

1.1　数据采集

语音识别数据集包括音频和文本两部分，如图 1 所示整个数据采集过程分成如下 6 部分：选择领域文本、网络爬虫、初步处理及专家评审、导入移动端设备、朗读纠错和上传云服务器，具体步骤如下：



图 1　藏语拉萨话语音识别数据集采集和处理图

Figure 1　Collection and processing diagram of the dataset of speech recognition for the Lhasa Dialect of Tibetan

（1）选择领域文本：本研究在语料选取上主要基于两个核心标准—文本的规范性与内容覆盖范围。为确保数据质量，最终选定新闻领域作为主要语料来源，并以中国藏语广播网作为核心数据采集渠道。

（2）网络爬虫：本研究基于 Scrapy 框架开发了定向网络爬虫[13]，对国内现有的藏语新闻网站的公开新闻数据进行采集。通过分析网站结构和页面元素，抓取新闻标题、正文内容、发布时间等结构化数据，并通过设置 2–3 秒的请求间隔和自动代理轮换机制，确保采集过程符合网站规范。所有获取的文本数据均经过去重、编码转换和格式标准化处理，最终构建成规范藏语新闻文本语料库。

（3）初步处理及专家评审：获取的藏语文本首先依据标点符号进行分句处理，随后通过去重筛选获得唯一性语句。最终由语言学专家逐句审核，确保文本规范性和准确性，保证转录文本符合以下规则：语法规范并且语句通顺，包含大多数的调音节、音联和协同发音，覆盖尽量多的发音现象；涉及尽可能多的领域，使转录文本更具普遍性，审核通过的文本构成 transcript 文本。

（4）导入移动端设备：将分句处理后的文本数据导入自主研发的移动端录音应用程序，为说话人提供便捷的录音操作界面。

（5）朗读纠错：说话人被邀请录制一句音频进行试音，在不存在严重的鼻音和嘶哑的情况下，说话人在移动端设备完成注册获取唯一说话人 ID 后，系统随机分配藏语文本进行朗读录音。通过实时试听与重录机制，说话人可自主修正发音问题，直至完成符合质量要求的有效录音。

（6）上传云服务器：录音完成后，音频文件以“说话人 ID+文本 ID”的命名规则上传至云服务器存储，所有的音频，采样率被设置为 16 kHz，采样精度设置为 16 bit，保存格式为*.wav，所有的音频通过无压缩上传。同时建立对应的文本索引关系，确保音频与文本数据的准确关联。

1.2　数据预处理

对于上传云服务器的数据，数据的形式为每位说话人一个单独的文件夹。然后分别对音频和文

本进行处理构成所需数据集，分别对音频和文本进行如下操作：

（1）文本：对原始文本执行以下标准化操作：移除包含非藏语内容（如数字、英文词汇）的音频样本；清除不发音符号（包括装饰符号“”及标点符号）；将低频反写字符统一映射为标准藏文（包括藏文中用于音译梵语词汇如“”是“”的反写）；基于字符统计结果剔除不在图 2 中的非法字符（如无法显示的乱码等），最终生成标准化的 transcript 文件。



图 2　藏文 Unicode 编码区块

Figure 2　Tibetan Unicode code block

（2）音频：检查音频和文本是否一一对应，然后通过 VAD 识别并去除音频中过长的静音片段，保留有效语音内容，并计算有效音频时长与对应文本长度的比值，对超出合理范围的异常样本进行人工复核，由专家判定是否保留或重新录制。针对不同说话人录音音量差异较大的问题，采用音频归一化处理技术，将所有语音信号的幅度统一映射至标准范围[-1,1]内，消除音量波动对语音识别模型训练的影响。该流程确保了最终数据集中的每个样本均满足音频清晰、文本准确、音量适中的质量要求。

最后数据集的格式如图 3 所示，每个说话人对应一个文件夹，并统计了数据的基本信息。

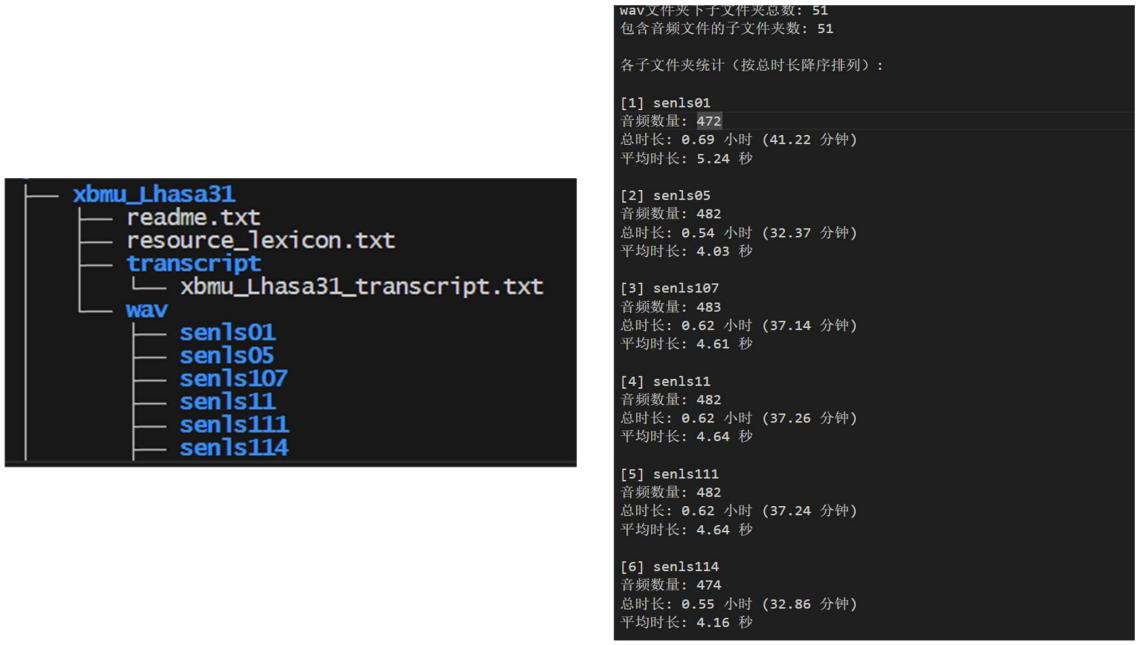

图 3　XBMU-bo-Lhasa31 数据集及相关信息

Figure 3　XBMU-bo-Lhasa31 dataset and its detailed information

2　数据样本描述

XBMU-bo-Lhasa31 数据集是一个专业构建的藏语拉萨话语音识别数据集，共采集自 51 位西北民族大学拉萨籍在校学生（平均年龄 20–25 岁，7 名男性和 44 名女性）。本数据集包含 24,289 条高质量语音样本，总时长达 31.61 小时，平均每条语音时长为 4.68 秒，采样率被设置为 16 kHz，采样精度设置为 16 bit，保存格式为*.wav。数据集采用结构化存储方案，包含以下核心组成部分：1）wav 音频文件夹，下设 51 个以说话人 ID 命名的子文件夹，每个子文件夹存储对应说话人的全部录音文件；2）完整文本转录文件；3）标准藏语发音词典；4）详细的数据集说明文档。

所有音频文件均采用“说话人标识_文本序号.wav”的统一命名格式。以“senls01-zang_24_2.wav”为例，其中“senls01”为说话人唯一注册 ID，“zang_24_2”为对应文本编码。通过多位说话人录制相同文本的设计，确保数据发音的多样性。对本数据集说话人的音频文件数目和音频总有效时长（min）进行统计，结果如表 1 所示。

表 1　音频数据统计表

Table 1　Audio statistics

说话人标识	音频文件数目	时长（分钟）	说话人标识	音频文件数目	时长（分钟）
senls86	484	41.59	senls16	475	37.81
senls116	478	41.56	senls149	476	37.45
senls72	481	41.54	senls11	482	37.26
senls36	478	41.52	senls111	482	37.24
senls256	479	41.37	senls107	483	37.14
senls01	472	41.22	senls150	472	36.72
senls12	477	40.92	senls117	477	36.02

说话人标识	音频文件数目	时长（分钟）	说话人标识	音频文件数目	时长（分钟）
senls22	476	40.85	senls25	476	35.36
senls37	476	40.76	senls99	472	35.17
senls55	479	40.59	senls92	472	35.06
senls226	479	40.39	senls45	478	35.01
senls39	484	40.27	senls169	476	34.46
senls15	479	40.24	senls229	476	34.36
senls21	479	40.13	senls157	477	34.02
senls274	484	39.95	senls190	477	33.98
senls237	480	39.63	senls191	469	33.80
senls34	477	39.18	senls295	468	33.76
senls41	480	38.99	senls234	455	33.56
senls184	484	38.86	senls114	474	32.86
senls197	483	38.64	senls05	482	32.37
senls40	487	38.61	senls205	443	31.92
senls48	477	38.53	senls130	473	31.90
senls27	477	38.16	senls46	476	31.26
senls255	477	38.06	senls322	475	30.61
senls28	476	38.06	senls35	478	29.83
senls193	462	37.83			

音频与转录文本一一对应，文本文件仿照 aishell 数据集格式[14]，采用如图 4 所示格式，文本中删除了藏语的音节分隔符和截止符，如有需要可自行添加。

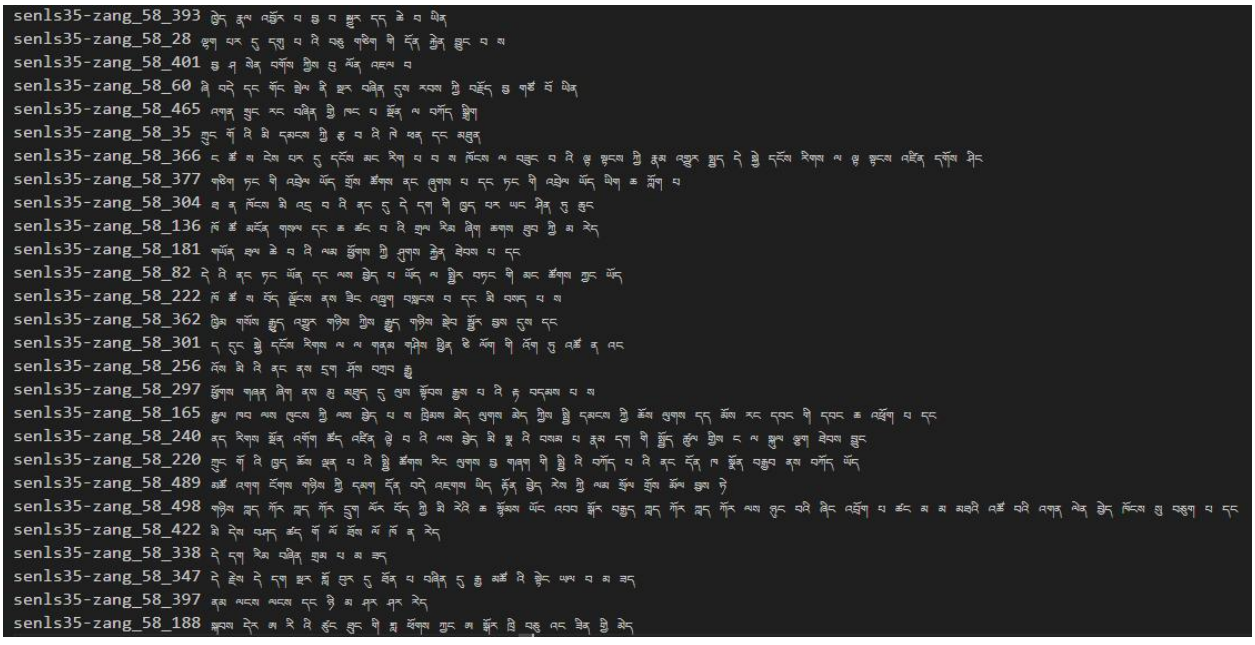

图 4　XBMU-bo-Lhasa31 数据集及相关信息

Figure 4　XBMU-bo-Lhasa31 dataset and its detailed information

本研究还构建了一套包含 5,933 个藏文音节的发音字典，如图 5 所示，采用国际音标（IPA）转写方案（如“ཀོཀ”转写为/k o k/），为传统语音识别模型提供音素级标注支持。该字典基于藏语传统音韵学理论设计，严格遵循藏语发音规则，覆盖常用词汇及新闻术语，确保对数据集文本的完整映射。

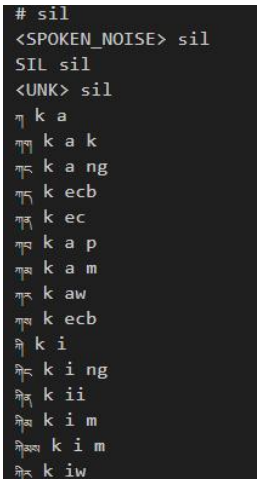

图 5　藏语音节发音词典

Figure 5　Pronunciation dictionary of Tibetan syllables

3　数据质量控制和评估

本研究构建了全流程、多层次的数据质量保障机制，该机制贯穿数据采集、处理与验证的各个环节。在数据采集阶段，组建了由藏语语言学专家和语音技术专家组成的文本评审小组，对所有候选文本进行严格筛选，确保语言规范性和内容适用性。录音过程中采用“实时监听-即时反馈”的交互式采集模式，要求每位发音人在完成录音后立即回放检核，对存在发音模糊、语调异常或背景噪声的样本进行现场重录，直至达到质量要求方可上传。

在技术处理层面，建立了自动化与人工相结合的双重质检体系：

（1）自动化检测：通过语音活动检测（VAD）算法精确识别并剔除无效静音片段；采用基于信号处理的音量归一化技术，将所有音频样本的幅度标准化至[-1,1]区间，消除因采集设备或发音习惯导致的音量差异。

（2）人工校验：组织专业标注团队对音频文本对齐情况进行逐条核查，重点修正转写错误和异常发音样本；开发专用的文本净化工具，结合正则表达式匹配与人工复查，彻底清除非藏语字符及非功能性符号。

为量化评估数据质量，实施了分层抽样检测机制：从全部数据中随机抽取 5%的样本（约 1,214 条），由 3 名认证语音评估专家按照以下标准独立评分，部分打分如表 2 所示：1）0 分：存在明显质量问题（如严重噪声、文本不符）。2）1 分：基本合格但存在轻微瑕疵（如微弱底噪）。3）2 分：发音清晰、文本准确、音质优良。

表 2　藏语拉萨话语音识别审核表

Table 2　Review form of speech recognition for the Lhasa Dialect of Tibetan

序号	音频	文本	专家 1	专家 2	专家 3
1	senls35-zang_58_297	ཕྱོགས གཞན ཞིག ནས མུ མཐུད དུ ལུས སྟོབས རྒྱས པ འི རྟ བདམས པ ས	2	2	2
2	senls11-zang_34_367	སྐྱེ འཕེལ དང སྤྲགས པ འི སྐྱེ འཚར རེད	2	2	2
3	senls184-zang_74_33	རྩོམ རིག སྐྱུ རྩལ ནི མིའི རིགས སྤྱི ཚོགས ཀྱི ལག ལེན བྱ སྤྱོད ཀྱི མཚོན ཚུལ གསོན ཉམས ལྡན པ ཞིག ཡིན ལ	2	2	1
4	senls184-zang_74_479	མེས རྒྱལ སྲུང སྐྱོབ བྱེད པ དང བཙན འཛུལ ལ འགོག རྒོལ བྱེད རྒྱུ ནི ཀྱུང ཧྭ མི དམངས སྤྱི མཐུན རྒྱལ ཁབ ཀྱི སྤྱི དམངས ཡོངས ཀྱི ལས འགན མཆོག ཏུ གྱུར པ ཞིག ཡིན	2	2	2
5	senls39-zang_62_218	དད ལྡན མང ཚོགས ལ མཐུན སྒྲིལ བྱས ཏེ དཔལ འབྱོར དང སྤྱི ཚོགས གོང དུ སྤེལ བ ར ལེགས སྐྱེས འབུལ དགོས	2	2	2
6	senls39-zang_62_294	མཚམས ལ ལ ར མོ འི དབང བོ ཕལ ལམ མི སྨིན པ ར འགྱུར ངེས ཡིན	1	2	2

4　数据价值

本数据集在低资源语言语音数据建设方面具有显著的学术价值和应用潜力，其创新性主要体现在三个方面：首先，在数据采集方面采用创新的采集策略和质量控制体系，在说话人多样性、语音纯净度和文本规范性等方面为后续构建语音识别数据集提供方法；其次，将自动化检测与人工校验有机结合，为其他少数民族语言数据建设提供了可借鉴的技术框架；最后，其多维度设计突破了传统单一任务数据集的局限性，能同时支持语音识别、语音合成和说话人识别等多个研究方向。与国内外同类数据集相比，本数据集在标准化程度和易用性方面具有明显优势，精心设计的存储结构和命名规则显著降低了数据预处理难度。本研究构建了一个支持语音识别任务研究的标准化低资源语言语音数据集，其创新的采集标注方案和模块化架构显著提升了数据质量和易用性，为语言资源建设和语音技术发展提供了重要基础支撑。

5　数据使用方法和建议

本节旨在简要说明 XBMU-bo-Lhasa31 数据集在藏语拉萨话语音识别任务中的可用性与基础使用思路。本数据集围绕真实拉萨话语境完成采集、清洗与标注，覆盖常见词汇与高频音节组合，文本与音节边界对齐可靠，可直接用于端到端识别模型的训练与评测。为保证可比性与复现性，选用 Whisper 作为统一实验基座。Whisper 采用基于 Transformer 的端到端编码—解码结构，依托大规模多语言多领域预训练，具有较强的跨语种迁移与小样本适配能力，能够在不增加复杂工程负担的前

提下快速吸收新语种信息，因此适合作为本数据集的实验底座与对照框架。

在具体使用上，建议按照不同的数据规模选择不同的训练策略，以下提供了三种不同的训练策略。其一，直接在原始词表与结构上进行全参微调，用于建立可复现可对比的实验基线；其二，在原有词表基础上增补少量藏语 token，先进行短周期预训练以感知新符号分布，再通过 LoRA 实现参数高效微调，在效果与成本之间取得平衡，作为本数据集的标准使用范式；其三，以音素为建模单元，使用发音词典统计音素建模单元，在原词表中增加音素标签训练 ASR 输出音素序列，并由语言模型将识别的音素转写回藏文音节，用于验证发音词典与音系约束的有效性。如表 3 所示，所有实验均以 Whisper 为基座模型训练并在同一评测规程下计算音节错误率（CER）与音素错误率（PER）。

表 3　不同微调策略下的藏语语音识别结果

Table 3　Tibetan speech recognition results under different fine-tuning strategies

方法	建模单元	训练流程概述	CER(%)	PER(%)
全参微调	字节码	直接以本数据集对 Whisper 全量参数端到端微调	45.80	-
预训练+全参微调	藏语 token	在原词表上增添藏语 token，先短周期预训练适配新符号，再以 LoRA 微调注意力与 FFN 等关键层	10.16	-
音素	音素输出	以音素为 ASR 输出，结合发音词典与语言模型回写为藏文音节	13.35	5.70

上述实验表明，XBMU-bo-Lhasa31 数据集能够有效支撑拉萨话语音识别的训练与评测。基于 Whisper 的对比实践进一步说明：对于原有词表未覆盖的新语种，直接进行全参微调并不稳妥；在词表中引入适配新语言的 token，先完成短周期适配再采用 LoRA 微调，能够在效果与成本之间取得更为稳健的平衡并带来可靠的识别表现。以音素为建模单元的策略在发音层面具有明确优势，但当前受制于语言模型与同域文本规模，音素到音节的回写仍存在改进空间；随着更强语言模型与更充足文本资源的加入，该路线的可读性与终端输出质量有望进一步提升。值得强调的是，本研究随数据集提供的发音词典在相关实验中展现出稳定、可解释的发音层级建模能力，验证了其在拉萨话语音识别中的有效性与实用价值。综上，本数据集不仅为不同训练范式提供了清晰且可复现的实践路径，也为后续在更强语言模型与多源文本驱动下的系统优化奠定了坚实基础。

致　谢

感谢西北民族大学中国民族信息技术研究院以及拉萨地区藏族同学的大力支持。

作者分工职责

马立克（2001—），男，甘肃兰州人，硕士研究生，研究方向为语音识别。主要承担工作：数据集预处理和整合、论文撰写。

李冠宇（1973—），男，甘肃秦安人，博士，教授，研究方向为语音识别、语音合成。主要承担工作：数据集采集和论文指导与修改。

谢晨宇（2002—），男，湖北十堰人，硕士研究生，研究方向为语音识别。主要承担工作：数据标注与质量控制。

孙倩（2001—），女，安徽阜阳人，硕士研究生，研究方向为语音识别。主要承担工作：数据处理和审查。

郭玉豪（2000—），男，山西运城人，硕士研究生，研究方向为语音合成。主要承担工作：数据集整合。

参考文献

[1] 边巴旺堆, 王希, 王君堡. 藏语语音识别研究进展综述[J]. 高原科学研究, 2022, 6(4): 76–84. DOI:10.16249/j.cnki.2096-4617.2022.04.009. [BIAN B, WANG X, WANG J B. An overview of the research progress of Tibetan speech recognition[J]. Plateau Science Research, 2022, 6(4): 76–84. DOI:10.16249/j.cnki.2096-4617.2022.04.009.]

[2] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]. Proceedings of the 40th International Conference on Machine Learning. ACM, 2023: 28492–28518. DOI:10.5555/3618408.3619590.

[3] MA L K, LI G Y, ZHE R Y. A study on speech recognition of Tibetan Amdo based on whisper[M]. Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2024). DordrechtAtlantis Press International BV2024: 367–373. DOI:10.2991/978-94-6463-490-7_40.

[4] 李森严. 基于迁移学习和语言模型融合的低资源语音识别研究[D]. 兰州: 西北民族大学, 2023. DOI:10.27408/d.cnki.gxmzc.2023.000156. [LI S Y. Low-resource speech recognition based on transfer learning and language model fusion[D]. Lanzhou: Northwest Minzu University, 2023. DOI:10.27408/d.cnki.gxmzc.2023.000156.]

[5] 梁鑫月. 儿童蒙古语语音识别研究[D].呼和浩特: 内蒙古大学, 2021. DOI:10.27224/d.cnki.gnmdu.2021.000421. [LIANG X Y. Research on Mongolian speech recognition for children[D]. Hohhot: Inner Mongolia University, 2021. DOI:10.27224/d.cnki.gnmdu.2021.000421.]

[6] DING S J, RIKHYE R, LIANG Q, et al. Personal VAD 2.0: optimizing personal voice activity detection for on-device speech recognition[C]//Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022). Incheon, Korea: ISCA, 2022: 3744-3748.DOI:10.21437/Interspeech.2022-856.

[7] 王嘉文. 藏语三大方言的语音识别方法研究[D]. 拉萨: 西藏大学, 2024. DOI:10.27735/d.cnki.gxzdx.2024.000063. [WANG J W. Research on speech recognition methods for three major Tibetan dialects[D]. Lhasa: Tibet University, 2024. DOI:10.27735/d.cnki.gxzdx.2024.000063.]

[8] ZHAO Y, XU X N, YUE J J, et al. An open speech resource for Tibetan multi-dialect and multitask recognition[J]. International Journal of Computational Science and Engineering, 2020, 22(2/3): 297. DOI:10.1504/ijcse.2020.107351.

[9] SOKY K, GONG Z, LI S. Nict-tib1: A public speech corpus of Lhasa dialect for benchmarking Tibetan language speech recognition systems[C]//2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA). IEEE, 2022: 1-5.

[10] 李运鹏, 张钦雅, 者润玉, 等. 跨语种迁移的藏语语音识别[J]. 中央民族大学学报(自然科学版), 2024, 33(4): 30－38. DOI:10.3969/j.issn.1005-8036.2024.04.005. [LI Y P, ZHANG Q Y, ZHE R Y, et al. Tibetan speech recognition for cross-language transfer[J]. Journal of MinZu University of China (Natural Sciences Edition), 2024, 33(4): 30－38. DOI:10.3969/j.issn.1005-8036.2024.04.005.]

[11] 宋志浩. 基于完全端到端方法的藏语拉萨话语音合成技术研究[D]. 兰州: 西北民族大学, 2023. DOI:10.27408/d.cnki.gxmzc.2023.000290. [SONG Z H. Research on fully end-to-end speech synthesis technology for Lhasa Tibetan[D]. Lanzhou: Northwest Minzu University, 2023. DOI:10.27408/d.cnki.gxmzc.2023.000290.]

[12] MENG Y, LI J, LIN G, et al. Dolphin: a large-scale automatic speech recognition model for eastern languages[EB/OL]. (2025-03-26) [2026-03-06]. https://arxiv.org/abs/2503.20212.

[13] FAN Y H. Design and implementation of distributed crawler system based on scrapy[J]. IOP Conference Series: Earth and Environmental Science, 2018, 108: 042086. DOI:10.1088/1755-1315/108/4/042086.

[14] BU H, DU J Y, NA X Y, et al. AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline[C]. 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I /O Systems and Assessment (O-COCOSDA). Seoul, Korea. IEEE, 2017: 1–5. DOI:10.1109/ICSDA.2017.8384449..

论文引用格式

马立克, 李冠宇, 谢晨宇, 等. XBMU-bo-Lhasa31：藏语拉萨话语音识别数据集[J/OL]. 中国科学数据, 2026, 11(1). (2026-03-09). DOI: 10.11922/11-6035.csd.2025.0122.zh.

数据引用格式

马立克, 李冠宇, 谢晨宇, 等. XBMU-bo-Lhasa31：藏语拉萨话语音识别数据集[DS/OL]. V1. Science Data Bank, 2025. (2025-07-01). DOI: 10.57760/sciencedb.j00001.01539.

XBMU-bo-Lhasa31: A dataset of speech recognition for the Lhasa Dialect of Tibetan

MA Like[1], LI Guanyu[1,2*], XIE Chenyu[1], SUN Qian[1], GUO Yuhao[1]

1. Key Laboratory of Language and Cultural Computing, Ministry of Education, Northwest Minzu University, Lanzhou 730030, P. R. China

2. Qinghai Tibet Plateau Humanistic Environment Data Intelligence Laboratory, Lanzhou 730030, P. R. China

*Email: xxlgy@xbmu.edu.cn (LI Guanyu)

Abstract: Tibetan speech recognition has important application value in fields such as Tibetan language education, news dissemination and other fields. The Lhasa dialect of Tibetan is widely used in Lhasa City

and its surrounding regions. However, due to geographical and other constrains, currently available Tibetan speech data resources remained limited and high-quality annotated data are particularly scarce. For this reason, this study constructs a professionally designed and standardized speech recognition dataset for the Lhasa dialect of Tibetan. The dataset was recorded in real-world environments using self-developed recording software, and was collected from 51 speakers, with a total duration of 31.61 hours, containing 24,289 speech samples, with an average duration of 4.68 seconds per sample. The data content was primarily selected from news-related texts to ensure linguistic standardization and domain representativeness. In order to guarantee data quality, we implemented a strict quality control process: firstly, the original texts were segmented into sentences and manually verified; after the recordings were completed, the Voice Activity Detection (VAD) technique was used to filter and regain high-quality speech samples; in addition, non-pronounced symbols in the text were normalized to improve the accuracy of speech recognition. The establishment of this dataset provides an important foundational resource for Tibetan speech recognition and is expected to facilitate the development of Tibetan speech recognition technology.

Keywords: Speech Recognition; Tibetan Lhasa dialect; Multi-Speaker; phonetic corpus

Dataset Profile

Title	XBMU-bo-Lhasa31: A dataset of speech recognition for the Lhasa Dialect of Tibetan
Data corresponding author	LI Guanyu (xxlgy@xbmu.edu.cn)
Data authors	MA Like, LI Guanyu1, XIE Chenyu, SUN Qian, GUO Yuhao
Time range	2020
Geographical scope	Lhasa and its surrounding regions
Data volume	2.69 GB
Data format	*.wav, *.txt
Data service system	https://doi.org/10.57760/sciencedb.j00001.01539
Sources of funding	National Natural Science Foundation of China (61633013); Gansu Province Science and Technology Major Special Program in 2024 (24ZDFA004).
Dataset composition	The dataset consists of audio files, text files and description files. Specifically, (1) the wav folder contains the audio folder and is divided into 51 subfolders according to the speakers, with a total duration of 31.61 hours, containing 24,289 speech samples, with an average duration of 4.68 seconds per sample. (2) The text in the transcript file corresponds to the audio on a one-to-one basis, where all the textual data are derived from the news domain, and the non-pronounced symbols in the text have been normalized. (3) The readme.txt file contains basic information about the dataset. (4) resource_lexicon.txt is a pronunciation lexicon file.