专题 大模型训练数据

XBMU-AMDO31：藏语安多方言语音识别数据集

谢晨宇[1]，李冠宇[1,2*]，马立克[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 730030

摘要：近年来，尽管语音识别技术在高资源语种（如英语、汉语）中取得显著进展，但针对藏语等低资源复杂音系语种的研究进展仍然缓慢。安多藏语作为低资源复杂音系语言，其语音识别面临数据稀缺与可用数据集质量和多样性不足的双重挑战。由于缺乏公开的数据集，相关研究面临着诸多限制。为此，本文构建并介绍了一个开源的藏语安多方言语音识别数据集。语音样本最初采集于中国甘肃省夏河地区，共收录了 66 位以藏语为母语者共 31 小时录音以及相应的转录文本，后续经过人工质检与标准化处理，确保了方言纯正性的以及数据的质量和一致性。本语音数据集的所有资源均已开放，目前已在多篇藏语语音识别相关论文或研究中被使用，得到业内专家的一致好评，更证明了数据集的质量。本数据集为藏语安多方言的高质量语音数据提供了重要补充，其复杂音系特性为跨语种迁移学习、小样本语音技术研究提供独特样本支持。

关键词：语音识别；安多藏语数据集；多说话人；低资源

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

文献 CSTR：
32001.14.11-6035.csd.2025.0113.zh

文献 DOI：
10.11922/11-6035.csd.2025.0113.zh

数据 DOI：
10.57760/sciencedb.j00001.01518

文献分类：计算机科学技术

收稿日期：2025-06-05
录用日期：2025-11-03
发表日期：2026-03-02

数据库（集）基本信息简介

数据库（集）名称	XBMU-AMDO31：藏语安多方言语音识别数据集
数据通信作者	李冠宇（xxlgy@xbmu.edu.cn）
数据作者	谢晨宇、李冠宇、马立克、孙倩、郭玉豪
数据时间范围	2020年
地理区域	中国甘肃省夏河地区
数据量	3.37 GB
数据格式	*.wav, *.txt
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01518
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项目（24ZDFA004）。
数据库（集）组成	语音数据集总时长为31小时，采集自66位发音人（男女各33人）。数据按用途划分为三部分：训练集包含54位发音人共18,505条语句；开发集由6位发音人共2,045条语句构成；测试集则包含另外6位发音人共2,035条语句。所有发音人均未在不同子集中重复出现。

* 论文通信作者
李冠宇：xxlgy@xbmu.edu.cn

引　言

在过去几十年中，深度学习技术的出现使语音识别技术取得了显著进展。微软团队在 2017 年搭建完成的 Switchboard 语音识别框架就已经可以达到 5.1%的词错误率（WER），并首次与专业人类转录员水平相当（人类平均 WER 约 5%–6%）[1]。现如今广泛使用的语言，如英语、中文和法语的识别正确率甚至已经超越了人类表现，这很大程度上要归功于大规模语音数据库的可用性，推动了深度学习算法的进步。例如，普通话的开源语料库 AISHELL[2]和 WenetSpeech[3]，英语的开源语料库如 LibriSpeech[4]和 GigaSpeech[5]，在统计语音识别和端到端语音识别系统的发展中发挥了关键作用。然而，对于一些少数民族语言，它们的使用人口相对较少或地理位置偏僻，缺乏音频数据采集和文本转录的资源，这阻碍了语音识别技术的发展。尽管目前存在部分低资源语音数据库，但这些数据集普遍规模有限。以藏族广泛使用的藏语为例，该语言面临着严重的训练数据匮乏问题，至今仍未达到令人满意的语音识别性能。

对于藏语来说，受限于藏语语音资源的稀缺，深度学习方法应用在藏语语音识别上的局限性是显而易见的。深度学习技术普及之前，藏语语音识别研究主要在于改进声学建模单元或在 LFMMI（Lattice Free Maximum Mutual Information）[6]等做一些局部优化，除此之外藏语语音识别发展几乎陷入停滞。近十年来，端到端神经网络的出现并成功应用在语音识别任务中，直接将语音识别成对应转录文本，抛弃了声学词典，不再需要语言方面的专业知识，使用单一的网络构建声学模型和语言模型的信息，弥补了藏语缺少的标准发音词典的不足，这对藏语语音识别是一个机遇和挑战。然而由于藏语三大方言差异较大，尤其是安多方言发音没有声调。目前开源的藏语数据集，例如中央民族大学采集的 TIBMD@MUC[7]，多集中于卫藏方言，难以保证安多方言的纯正性，无法满足安多藏语方言语音识别的训练需求。同时，语音识别任务例如说话人识别需要大量特定的数据支撑，这也是藏语语音识别的痛点之一。

据作者所知，本文将要介绍的 XBMU-AMDO31 是目前最早公开同时也是规模最大的藏语安多方言语音识别数据集，由西北民族大学录制。它包括由 66 名说话人录制的 31 个小时的安多藏语语音数据，同时包括对应转录文本。本研究利用专业的录音设备采集和人工校对相结合的方式完成了本数据集的构建，可用于端到端语音识别模型或大模型语音识别的研究。本数据集内容涵盖新闻类、文学类，教育类、科技类等多个领域，覆盖面广，可用于多种场景。文本语料库有 100 个文本，每个文本大约有 500 句子，包含藏文几乎全部的常用字符，为确保数据质量，通过引入人工校对环节，通过双重验证机制提升了数据的准确性与可靠性。

1　数据采集和处理方法

藏语安多方言语音识别数据集包含 18505 条安多藏语语音，共 66 个说话人。数据集由藏语安多方言语音集和对应的藏文文本集两部分组成。

1.1　文本语料库

藏文作为拼音文字系统，其三大方言虽发音差异显著，但共享统一的书写体系。该文字系统以音节为基本单位，每个音节由 34 个藏文字母（含 4 个元音和 30 个辅音）按照特定构字规则组合而成。在传统音节结构中，字母根据空间位置可分为六类：基字（必选核心）、前加字、上加字、下

加字、后加字及再后加字。其中 30 个辅音均可担任基字，通过纵向叠加（最多 3 个元音字母位于基字上方，仅 1 个元音可置于下方）和横向扩展（最多附加 5 个修饰字母）的方式，构成 1–7 个字母组合的音节结构[8]。

文本语料库的转录文本对于语音录制是至关重要的，转录文本录制的音频将直接影响到语音识别系统的训练，因此文本内容应具有广泛领域代表性。数据采集采用分级策略：主要来源为权威网络资源（新闻门户网站占比 60%，政府官网占比 30%），次要来源为标准化教材（5%）和经典文学作品（5%），最终形成覆盖新闻时事、文学艺术、教育教学、科学技术等多元领域的平衡语料库。文本语料库有 100 个文本，每个文本大约有 500 个句子，包含藏文几乎全部的常用字符[9]。

1.2 语音数据库

藏语属于汉藏语系的藏缅语族藏语支，历史悠久，使用人口众多，其作为我国藏族人民交流和沟通的主要载体，具有丰富的表现力、优美雅致、语调抑扬顿挫等特点。藏语分为卫藏、安多、康巴三种方言，安多方言相对于另外两种方言稍有不同。卫藏方言和康巴方言有声调，而安多方言没有声调。就安多方言而言，其主要分布在中国甘肃、青海、四川等地，尽管分布范围广泛，但安多方言内部差别较小，安多方言以甘肃夏河话为代表。安多方言另外一个主要特点是包含较多的复辅音，安多方言共有 35 个声母和 8 个韵母。

语音数据库的录音文本从文本语料库内进行筛选，所选取的文本之间不存在重叠情况，录音文本会随机指派给不同的发音人开展独立的语音录制工作。参与此次语音数据录制的发音人均为母语是安多藏语的中国甘肃夏河地区藏族人士，其年龄处于 18 岁至 30 岁之间。

整个录音流程划分为两个阶段。第一阶段，先邀请说话人录制单句音频用于试音，仅当音频不存在明显鼻音或嘶哑等问题时，才会邀请其参与完整语音数据集的录制。第二阶段，将西北民族大学研发的专用录音软件分发给说话人，要求其使用个人移动设备进行音频采集。录制需在安静且无回声的环境中进行，说话人需以清晰流畅的方式朗读转录文本，且全程会有专业人员提供指导。最终提交的音频将由专业人员对音质和文本匹配度进行二次质检，以确保形成高质量的语音数据库。所有音频的采样率设定为 16 kHz，采样精度为 16 比特，以无压缩的 WAV 格式保存并上传。

1.3 数据集整合

当语音数据集和文本语料库整合完毕后，对文本数据进行预处理、机器翻译校验、后处理、专家检验；对音频数据进行音频质量检测、重采样、专家审核、重命名，最终整合得到。数据集完整处理方案如图 1 所示。

数据处理具体处理步骤如下：1）文本预处理环节：通过人工核查方式，筛查是否存在音频与文本不对应的情况，即确认是否有缺少音频的文本或缺少文本的音频。2）机器翻译校验：借助机器翻译技术将藏文文本转换为可识别文本，以此检测原文文本的准确性。3）后处理操作：执行数据清洗工作，通过定义正则表达式匹配藏文字符，并对特殊字符（如书名号、双引号等）进行针对性处理。4）专家审核机制：从数据集中随机抽取部分样本，由专业领域专家进行审核与校对，确保数据质量。5）文本整合阶段：将处理完毕的音频与文本数据进行关联整合，最终形成完整的数据集。

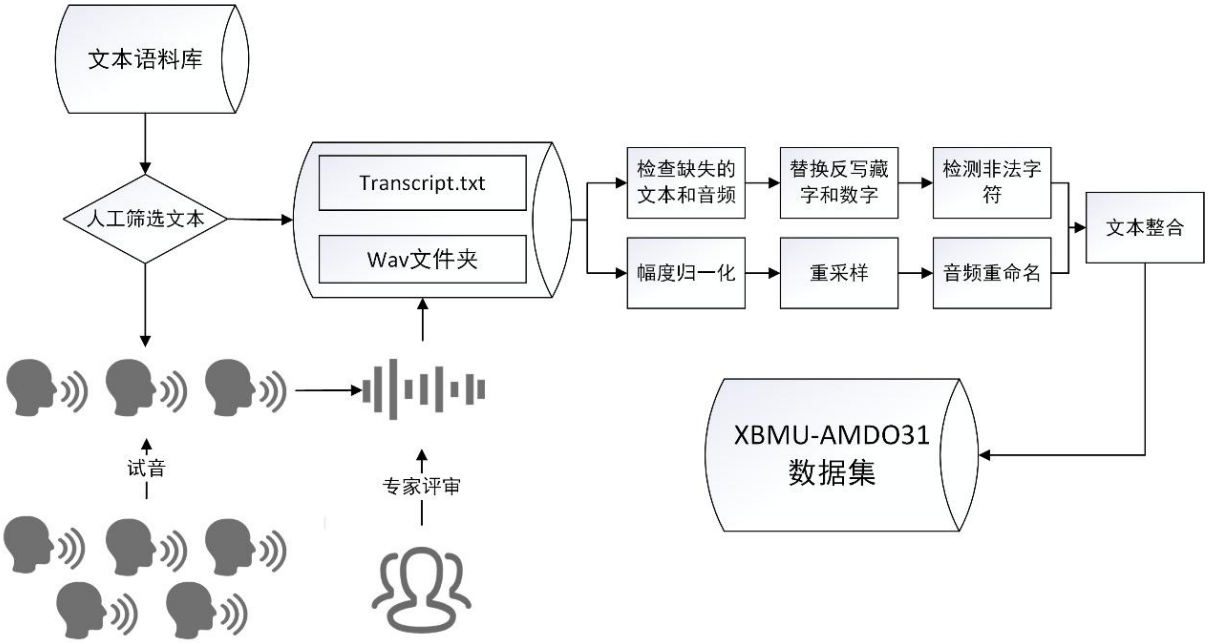


图 1　安多方言数据集采集和处理流程

Figure 1　Data collection and processing pipeline for the Amdo Tibetan dialect corpus

2　数据样本描述

语音数据集共 31 个小时，由 66 个说话人进行录制，其中包含 33 名男性和 33 名女性。语音数据集被分为训练集、开发集和测试集以及对应的文本文件。

图 2 是藏语安多方言语音识别数据集的文本示例，第一列为音频 wav 的名称，第二列为音频对应的藏文文本，音频的相关信息以“_”作为分隔符。其中“a”代表音频文件是安多方言（AMDO 的首字母）；“0”或“1”代表说话人的性别，其中“0”代表女性，“1”代表男性；紧随其后的是以分割符“-”分开的说话人信息，左右两边分别代表说话人姓名简写和说话人所属组别；最后的数字则是音频信息的唯一编号，共 22585 条。以 a_1_zx-A51_22562 为例，即为男性（1）说话人（zx-A51）录制的藏语安多方言（a）语音，这条数据为第 22562 条。

a_0_crc-A27_11700 ང་དེ་བོད་གི་ས་ཆ་སོང་ངས་རྟ་ཚོན་རྒྱུད་དགའ་གི
a_0_lmc2-A20_8689 དེའི་ཚ་ཤས་གང་དང་གང་ལ་དཉེ་བ་འཇེབ་རྒྱབ་ཡིའི་ཚོ
a_0_snyz-A21_9073 སྐོབ་གསོའི་སྐོབ་ཁྲིད་ཀྱི་དགོས་མཁོ་དང་ཡང་མཐུན་དགོས་པ་དང
a_1_gbj-A61_27277 རྒྱལ་ཁྱིད་ཐྱི་ཁྲབ་ངང་ཚོས་ལུགས་ལས་དོན་སྒྲུབ་ལ་ཚགས་ཚོར་འགོད་དུ་འཛུག་ཚོག
a_1_ymcr-A54_24298 ཡ་སྐོ་མཐུན་ཁྲིད་ཤེ་ཞིང་ལས་ཚན་རིག་ངང་ཚོ་ཡོད་ཐིས
a_1_yqj-A13_5401 ཐྲ་ན་མེད་པའི་སྒོང་སྐོལས་དེས་ཆུབས་ཅེན་ཀྱི་བྱ་གཞག་ལ་འདེགས་སྐོར་དང་སྤུས་འབེད་ཁེད་བཞིན་ཡོད་
a_1_zx-A51_22553 བདག་གིས་ད་དུང་དཔེར་བཙོད་ཅིག་བཞག་རུས་ཐུབ་ཚུལ་གཅིག་ཀི་འཚུང་ཁུངས་ནི
a_1_zmj-A30_13065 ང་ཚོའི་ཤེ་ལས་ཀྱིལ་ཅིན་ལེ་ཡིས་འགན་ཁུར་བའི་ལས་ལུགས་ལག་བསྟར་ཐུ་དགོས

图 2　藏语安多方言数据集文本示例

Figure 2　Textual examples from the Amdo Tibetan dialect corpus

训练集由 54 位发音人（26 位女性和 28 位男性）组成，共采集 18,505 条语句；开发集包含 6 位发音人（4 位女性和 2 位男性）录制的 2,045 条语句；测试集则由 6 位发音人（3 位男性和 3 位女性）提供的 2,035 条语句构成。每位发音人平均录制约 450 条语句，个别发音人录制的语句数量不足 200 条。数据集中不同说话人时长统计表如表 1–3 所示。表 4 提供了语音数据集所有子集的简要描述。

表 1　训练集统计表

Table 1　Training set statistics

说话人唯一标识	音频数目	时长（小时）	说话人唯一标识	音频数目	时长（小时）
a_0_chj	213	0.2052	a_1_cg2	74	0.1004
a_0_djc2	152	0.2550	a_1_crzx	447	0.6066
a_0_djzm	420	0.6124	a_1_crzx2	449	0.5196
a_0_dmc	450	0.7335	a_1_dhlcr2	450	0.6571
a_0_dzc	449	0.5081	a_1_lmj	64	0.0802
a_0_dzj	449	0.5774	a_1_lmj2	446	0.6619
a_0_dzzm	125	0.1601	a_1_njcr	449	0.5986
a_0_gdmc	450	0.9867	a_1_rqzx	450	0.6745
a_0_gzj2	448	0.5738	a_1_sbz	446	0.6675
a_0_gzzm	449	0.4594	a_1_sjk	450	0.6468
a_0_hmc	226	0.3511	a_1_wdcr	19	0.0317
a_0_yz	450	0.5417	a_1_djcd	35	0.0604
a_0_zmj	450	0.9640	a_1_djcd2	448	0.5575
a_0_ztj	449	0.3916	a_1_djrq	447	0.6682
a_0_zxj2	449	0.6188	a_1_dz	152	0.1236
a_0_jhmj	446	0.5091	a_1_dzcr1	450	0.7215
a_0_lmc	450	0.6795	a_1_dzj2	205	0.4020
a_0_lmc3	430	0.6137	a_1_gbj	448	0.4941
a_0_lmcr	450	0.5468	a_1_gzdj	448	0.5029
a_0_njc	165	0.1818	a_1_gzj	450	0.5186
a_0_njm	144	0.2794	a_1_lmdz	22	0.0346
a_0_snyz	443	0.5513	a_1_yqj	449	0.8396
a_0_snzm	449	0.7079	a_1_zgcr	93	0.1597
a_0_snzm2	450	0.4269	a_1_zmj	62	0.0737
a_0_wmc	450	0.6163	a_1_zx	61	0.0797
a_0_wmc2	448	0.5791	a_1_wmdz2	449	0.5230
a_1_wmdz	139	0.2285	a_1_ymcr	449	0.5110

表 2　开发集统计表

Table 2　Development set statistics

说话人唯一标识	音频数目	时长（小时）	说话人唯一标识	音频数目	时长（小时）
a_0_crc	447	0.6683	a_0_rzlm	450	0.5275
a_0_ctj	145	0.1799	a_1_rqzx2	449	0.4269
a_0_lmc2	450	0.7961	a_1_zmzx	104	0.2096

表 3　测试集统计表

Table 3　Test set statistics

说话人唯一标识	音频数目	时长（小时）	说话人唯一标识	音频数目	时长（小时）
a_0_cacm	448	0.4615	a_1_crzx3	450	0.7131
a_0_jg	100	0.1370	a_1_lhj	142	0.2064
a_0_zkj	445	0.7013	a_1_zmj2	450	0.6289

表 4　数据结构

Table 4　Data structure

子集	音频时长（小时）	男性（人数）	女性（人数）
训练集	25.37	28	26
开发集	2.81	2	4
测试集	2.85	3	3

3　数据质量控制和评估

本数据集通过机器录制和人工质检来保证数据集的质量。为了避免藏语音频和文本出现偏差，随机抽取了部分样本，并对其进行检测。邀请藏语语言专家和西北民族大学藏语专家对录制的音频进行打分评价，如图 3 所示。数据质量高的数据最终整理成为藏语语音识别数据。

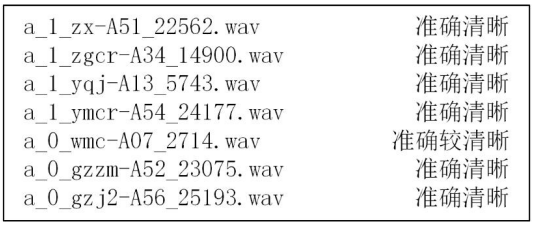


图 3　安多方言音频审核结果图

Figure 3　Quality verification results of the Amdo Tibetan dialect corpus

同时对文本清洗后的内容进行了人工再校验，反复对比文本和原始文稿，防止有非藏文字符出现，将质量极低的文本删除，重新补录文本。依据藏语三大方言的语音识别方法研究[10]将歧义数据进行修正。在藏语语音识别的文本标注和转写中必须严格遵循语言学规范和实际发音规则，不能简单地进行字符映射或数字替换。为了保证数据集整体的质量和准确性，基于音位优先以及借词处理的原则，转写必须反映实际发音，而非字形相似性。梵语借词中的特殊字母（如ཌ、ཥ、ཫ）需保留原字符，不可替换为藏语原生字母。在数字替换上不能将阿拉伯数字“1000”直接替换为藏文数字“྅”或字符“ཁྲ་”（意为“千”），而是应该按语音转写：需根据发音标注，例如“1000”在口语中可能读作“ཆིག་ཁྲ་”（chig tong，即“一千”），而非逐字读“྅”；“2024”可能读作“ཉི་ཁྲ་ཉི་ཤུ་རྩ་བཞི”（nyi tong nyi shu tsa zhi）。部分藏语歧义字符与数字发音转写对照如图 4 所示。

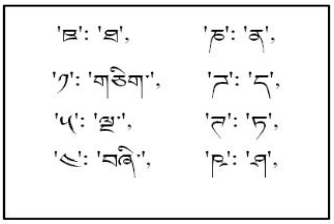


图 4　歧义字符与数字发音转写对照

Figure 4　Comparison of ambiguous characters and numerical pronunciation transcription

为了表明 XBMU-AMDO31 数据集具有较高的转录质量，分别搭建了传统的统计模型和端到端语音识别模型，通过使用音节为建模单元的方法比较不同模型在数据集上的表现能力。基于 Kaldi 工具包[11]训练了混合系统（Time Delay Neural Network-Hidden Markov Model，TDNN-HMM）[12]的语音识别基线模型，同时基于 ESPnet 工具包[13]和 Icefall 工具包训练了端到端语音识别基线模型来进行比较，实验结果如表 5 所示。

表 5　基线系统测试结果

Figure 5　Baseline system test results

模型	GMM-HMM	TDNN-HMM	Conformer,CTC	Conformer, pruned transducer
开发集音节错误率	19.71%	17.71%	14.82%	13.57%
测试集音节错误率	25.40%	15.29%	13.84%	11.72%

为确保实验结果的可靠性，本研究额外使用中央民族大学公开的多方言藏语数据集中的数据作为基准测试集，用于对比评估不同模型的识别性能，实验结果如表 6 所示。

表 6　中央民族大学语料测试

Figure 6　Corpus test of Minzu University of China

模型	GMM-HMM	TDNN-HMM	Conformer,CTC	Conformer, pruned transducer
音节错误率	49.13%	44.73%	26.1%	21.72%

实验结果可以很好反映出 XBMU-AMDO31 数据集的整体质量，并从侧面反映出语音识别模型架构的发展趋势。尽管用音节做建模单元已经成为搭建语音识别模型所使用的方法之一，然而以音节作为建模单元时，其性能表现往往受限于特定领域，难以实现跨领域的稳定识别效果。这一局限性主要源于低资源语料库难以全面覆盖所有文字变体。在此情况下，采用更细粒度的建模单元可能更有利于提升低资源语音识别系统的泛化能力。

4　数据价值

现阶段语音识别的发展速度极快，国内外的研究普遍基于大规模语音语料，数据集的稀缺是导致低资源语料语音识别发展速度缓慢的重要原因之一。在发布这篇论文之前，XBMU-AMDO31 就已在多篇高质量论文中作为数据集被使用。由 66 个不同性别的说话人所组成的数据集，可供科研工作

人员从事各个方面的藏语语音识别研究，例如说话人识别，说话人语音分离，说话人语音克隆，年龄与性别识别以及语音生物特征认证等。命名规则的完善方便使用者能够快速修改预处理流程，开展相关实验，并用于低资源语料语音识别模型的搭建。特定的多说话人藏语安多方言数据集，使这些任务成为可能。后续计划开源更多的低资源语音数据集，和全世界研究人员共享学术成果，推动低资源语音识别领域的发展。

5 数据使用方法和建议

本数据集提供了丰富的藏语安多方言和对应的转录文本，可供研究人员用于深度学习模型的训练和验证，从而为低资源语音识别的研究提供有价值的支持。

近期，低资源语音识别领域的研究主要集中在数据高效建模、跨语言迁移学习和多模态融合等方向，尤其是在大规模多语言语音（Massively Multilingual Speech）项目和声学事件检测与定位（Sound Event Localization and Detection）相关技术的结合成为热点之后，通过自监督预训练和跨语言迁移实现语种的扩展，极大程度上保留了低资源语言能力。

针对低资源语种数据稀缺的难题，联合训练 ASR 与文本到语音（TTS），可以显著降低低资源语种的数据需求[14]，同时也可以考虑将 SELD 与 ASR 结合，利用环境声音和语音的共享特征来提升低资源场景的鲁棒性。通过多模态数据增强的方法（如合成混响、噪声注入）来模拟真实环境[15]，在非洲和南美原住民语言数据集中验证，WER 降低 10%–15%，在 XBMU-AMDO31 数据集上也同样适用。

随着大模型以及迁移学习技术的出现，结合语言学先验知识与数据驱动方法，本数据集很可能可以将低资源语音识别的技术带向新的高度，同时为低资源语言的智能化处理开辟更优路径。

致　谢

本数据集的整合以及论文的撰写得到了西北民族大学语音工程组李运鹏、李森严、张钦雅以及宁洁雯等研究人员的支持，在此深表感谢。

数据作者分工职责

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。

马立克（2000—），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] XIONG W, WU L, ALLEVA F, et al. The microsoft 2017 conversational speech recognition system[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 5934–5938. DOI:10.1109/ICASSP.2018.8461870.

[2] 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.

[3] ZHANG B B, LV H, GUO P C, et al. WENETSPEECH: a 10000+ hours multi-domain mandarin corpus for speech recognition[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 6182–6186. DOI:10.1109/ICASSP43922.2022.9746682.

[4] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. IEEE, 2015: 5206–5210. DOI:10.1109/ICASSP.2015.7178964.

[5] CHEN G G, CHAI S Z, WANG G B, et al. GigaSpeech: an evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio[EB/OL]. 2021: arXiv: 2106.06909. https://arxiv.org/abs/2106.06909.

[6] POVEY D, PEDDINTI V, GALVEZ D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]//Interspeech 2016. ISCA, 2016: 2751–2755. DOI:10.21437/interspeech.2016-595.

[7] 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.

[8] 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. [SONG Z H. Research on Speech Synthesis Technology for Lhasa Tibetan Based on Fully End-to-End Methods[D]. Lanzhou: Northwest Minzu University, 2023.]

[9] 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. [LI S Y. Research on Low-Resource Speech Recognition Based on Transfer Learning and Language Model Fusion[D]. Lanzhou: Northwest Minzu University, 2023.]

[10] [王嘉文. 藏语三大方言的语音识别方法研究[D].西藏: 西藏大学, 2024. [WANG J W. Research on Speech Recognition Methods for Three Major Dialects of Tibetan[D]. Xizang: Xizang University, 2024.]

[11] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]//IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011.

[12] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural network architecture for efficient

modeling of long temporal contexts[C]//Interspeech 2015. ISCA, 2015: 3214 – 3218. DOI:10.21437/interspeech.2015-647.

[13] WATANABE S, HORI T, KARITA S, et al. ESPnet: end-to-end speech processing toolkit[EB/OL]. 2018: arXiv: 1804.00015. https://arxiv.org/abs/1804.00015.

[14] PRATAP V, TJANDRA A, SHI B, et al. Scaling speech technology to 1,000+ languages[J]. Journal of Machine Learning Research, 2024, 25(97): 1-52.

[15] VILLALBA J, BORGSTROM J, JAHAN M, et al. Advances in language recognition in low resource African languages: the JHU-MIT submission for NIST LRE22[C]//INTERSPEECH 2023. ISCA, 2023: 521 – 525. DOI:10.21437/interspeech.2023-1094.

论文引用格式

谢晨宇, 李冠宇, 马立克, 等. XBMU-AMDO31：藏语安多方言语音识别数据集[J/OL]. 中国科学数据, 2026, 11(1). (2026-03-02). DOI: 10.11922/11-6035.csd.2025.0113.zh.

数据引用格式

谢晨宇, 李冠宇, 马立克, 等. XBMU-AMDO31：藏语安多方言语音识别数据集[DS/OL]. V1. Science Data Bank, 2025. (2025-06-27). DOI: 10.57760/sciencedb.j00001.01518.

XBMU-AMDO31: A dataset of speech recognition dataset for the Amdo dialect of Tibetan

XIE Chenyu[1], LI Guanyu[1,2*], MA Like[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 Human Environment Data Intelligent Laboratory, Lanzhou 730030, P.R. China

*Email: xxlgy@xbmu.edu.cn

Abstract: In recent years, while significant progress has been made in speech recognition technology for high-resource languages (such as English and Mandarin), research on low-resource languages with complex phonologies, like Tibetan, has progressed relatively slow. As a low-resource and phonologically complex language, Amdo Tibetan faces dual challenges in speech recognition: data scarcity and insufficient quality and diversity of available datasets. The lack of publicly accessible datasets has imposed numerous constraints on related research. To address these challenges, this paper introduces and presents an open-source speech recognition dataset for the Amdo Tibetan dialect. The speech samples were initially collected in Xiahe County, Gansu Province, China, comprising 31 hours of recordings from 66 native speakers along with corresponding transcriptions. Subsequent manual quality control and standardization were applied to ensure the authenticity of the dialect as well as the consistency and quality of the data. All resources in this dataset have been made publicly available and have already been utilized in multiple

research papers and studies on Tibetan speech recognition, receiving widespread acclaim from experts in the field—further validating the dataset's quality. This dataset serves as an important supplement to high-quality speech data for Amdo Tibetan, and provides unique support for cross-lingual transfer learning and few-shot speech technology research due to its complex phonological characteristics.

Keywords: speech recognition; Amdo Tibetan dataset; multi-speaker; low-resource

Dataset Profile

Title	XBMU-AMDO31: A dataset of speech recognition dataset for the Amdo dialect of Tibetan
Data corresponding author	LI Guanyu (xxlgy@xbmu.edu.cn)
Data authors	XIE Chenyu, LI Guanyu, MA Like, SUN Qian, GUO Yuhao
Time range	2020
Geographical scope	Xiahe region, Gansu Province, China
Data volume	3.37 GB
Data format	*.wav, *.txt
Data service system	https://doi.org/10.57760/sciencedb.j00001.01518
Source of funding	National Natural Science Foundation of China (61633013); Gansu Provincial Science and Technology Major Project (24ZDFA004)
Dataset composition	The speech dataset contains a total of 31 hours of recordings collected from 66 speakers, including 33 males and 33 females. The dataset is divided into three parts: the training set consists of 18,505 utterances from 54 speakers; the development set includes 2,045 utterances from 6 speakers; and the test set comprises 2,035 utterances from another 6 speakers. None of the speakers are repeated across the different subsets.