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

OPEN ACCESS
CC BY

文献 CSTR：
32001.14.11-6035.csd.2025.0150.zh
文献 DOI：
10.11922/11-6035.csd.2025.0150.zh
数据 DOI：
10.57760/sciencedb.j00001.01606

文献分类：计算机科学技术

收稿日期：2025-08-09
修回日期：2025-12-01
录用日期：2026-01-28
发表日期：2026-06-03

* 论文通信作者
于洪志：yhz@xbmu.cn

藏语拉萨话元音发音时的超声舌成像数据集

张金溪[1]，于洪志[2,3*]，张晶[1]

1. 兰州财经大学，商务传媒学院，兰州 730020
2. 西北民族大学，语言与文化计算教育部重点实验室，兰州 730030
3. 西北民族大学，甘肃省民族语言文化智能信息处理重点实验室，兰州 730030

摘要： 本数据集是针对藏语拉萨话元音发音特征和语音工程研究而建立的超声舌成像数据集。研究采用超声舌位动态成像技术采集母语者的发音器官（舌体）运动数据，在标准录音室环境下采集了一男一女两名拉萨本地发音人的元音发音数据，发音人均为母语是拉萨话的大学生且无发音障碍。实验使用 B 型超声设备（探头频率 5.0 MHz）实时记录发音时的舌体运动动态图像，同时配备高质量音频采集设备（采样率 44.1 kHz）记录语音信号。数据集包含[a]、[i]、[u]、[e]、[ø]、[ɛ]、[y]、[o]八个单元音和[au]、[iu]两个复元音的发音数据，共计 200 组有效数据单元（含超声视频、发音音频及对应文本）。所有数据均按照统一标准进行采集，发音人需保持标准坐姿，头部和超声探头固定，发音强度控制在 65－75 dB 范围内。为确保数据质量，由两名语音学专家对采集数据进行双重校验，剔除不合格样本。本数据集弥补了藏语拉萨话发音生理研究的不足，为藏语语音教学、言语病理治疗、语音识别与合成、多模态深度学习研究提供了重要的基础数据支撑，也可用于跨语言发音生理特征的对比研究，也可为濒危语言保护提供技术支撑，具有较高的语言学研究与语音工程技术应用价值。

关键词： 藏语拉萨话元音；超声舌成像；人工智能；语音识别与合成；多模态深度学习

数据库（集）基本信息简介

数据库（集）名称	藏语拉萨话元音发音时的超声舌成像数据集
数据通信作者	于洪志（yhz@xbmu.cn）
数据作者	张金溪、于洪志、张晶
数据时间范围	2019年
数据量	8.15 GB（解压后60.9 GB）
数据格式	*.avi, *.xml, *.wav, *.pdf
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01606
基金项目	国家自然科学基金（62366045，62566057）；国家社会科学基金（19CYY039）；甘肃省中央引导地方科技发展资金项目（25ZYJA034）；兰州市科技计划项目（2025-2-42）；甘肃省教育教学成果培育项目（2023GSJXCGPY-60）；兰州财经大学科研专项；陇原青年英才项目“甘肃省内多民族语言共演化特征研究”。

数据库（集）组成	数据集共包括2个发音人数据文件： 发音人1（男生）的数据文件包括：（1）ultrasound1.zip是超声视频，包含100个avi格式文件和100个xml格式文件，数据量3.76 GB；（2）audio2.zip是发音音频，包含100个wav格式文件，数据量5.87 MB；（3）text3.zip是发音文本，包含1个pdf格式文件，数据量409 KB。 发音人2（女生）的数据文件包括：（1）ultrasound1.zip是超声视频，包含100个avi格式文件和100个xml格式文件，数据量4.37 GB；（2）audio2.zip是发音音频，包含100个wav格式文件，数据量6.25 MB；（3）text3.zip是发音文本，包含1个pdf格式文件，数据量409 KB。

引　言

超声成像技术凭借其临床安全性、非侵入性及实时动态监测等优势，在生物医学领域形成多维度应用体系。该技术不仅广泛应用于超声医学的三大分支——超声诊断学、治疗性超声应用及生物医学超声工程，实现人体生理组织的实时可视化与功能分析，同时已拓展至语言学研究领域[1]。在语音科学实践中，超声舌成像（ultrasonic tongue imaging，UTI）等非侵入式观测手段为语言发音机制解析、语音产生过程动态建模提供了关键技术支撑[2]，成为连接医学工程与人文科学的创新交叉研究工具，也已成为发音生理研究的首选工具。

超声舌成像技术作为跨学科的语言研究工具，在海外语言学界已形成成熟的应用体系。该技术深度应用于语言教学、二语习得及协同发音等发音生理机制研究[1]，同时为言语障碍康复、舌部术后功能重建等病理语音干预提供技术支持[2]，并在语音识别、语音合成及无声语音交互等前沿语音工程领域发挥重要作用[3]。相较之下，国内超声成像技术在语言语音研究中的应用尚处发展阶段，现有成果多聚焦于汉语普通话生理语音学[4]、普通话音频-超声同步的多模态数据库[5]及言语工程[6]领域，针对少数民族语言的系统性研究则更为稀缺，亟待拓展跨语言比较的研究维度。

藏语作为汉藏语系的重要分支，其语音体系研究对理解语言演化、跨语言比较及语音工程应用具有显著价值。拉萨话作为藏语的标准方言和语音学研究的核心对象，其元音发音过程中舌位的动态变化不仅决定音质差异，更反映语言特有的发音协同效应[7]与器官运动规律[8]。然而传统研究依赖听觉转写或静态图像（如 X 光片），难以精准捕捉舌位连续变化，导致发音机制理解受限。而发音生理数据集通过动态舌位运动捕捉，为揭示语音产生机制提供关键证据[9]，尤其在声学-发音关联建模[10]、言语病理诊断[11]及濒危语言保护[12]等领域不可或缺。当前国际主流超声数据库（如 UltraSuite 数据库[13]、TAL 数据库[14]、SSR7000 数据库[15]）多聚焦英语语系，而少数民族语言特有的发音适应机制研究仍缺乏系统性数据支撑。在藏语发音研究领域，现有成果主要基于声学分析[16]及传统发音器官观察，尚未建立可量化的动态舌位影像数据库。

本数据集系统整合拉萨话元音的超声动态舌位影像、同步音频及文本标注，弥补了高海拔语言发音可视化数据库的不足。本数据集的潜在应用场景包括：（1）语音建模，为藏语发音合成/识别提供生理驱动参数；（2）教学优化，帮助藏语学习者纠正发音错误和提高学习效率；（3）言语障碍治疗，借助超声舌成像技术可以为言语障碍患者制定个性化治疗方案提供参考；（4）跨语言比较，支持汉藏语系发音类型学实证研究；（5）计算方法，推动多模态深度学习模型的舌位特征提取。（6）语言保护，有助于保护和发展藏语语言文化，能为藏语濒危方言保护提供技术支持。数据严格遵循国际发音数据库规范，为藏语语音研究提供可验证、可扩展的基础资源。

1 数据采集和处理方法

1.1 发音语料的筛选及发音人的选择

本研究构建的藏语拉萨话发音时的超声舌成像数据集，其语料基础源自两部权威方言研究著作：藏语卫藏方言语音特征研究文献[17]和藏语方言对比分析调查表[18]。鉴于研究的需要，我们采集的是具有实际意义的藏语拉萨话韵母（包含元音）发音数据（涵盖 10 个藏语拉萨话典型元音及复元音，这 10 个元音搭配了不同声母或韵尾，共有 100 个词汇，每位发音人对这 100 个词汇进行朗读，共计生成 2×100=200 组有效数据单元）。在语料选取策略上，研究团队聚焦发音人日常高频使用的词汇，这类语料具有三大优势：其一，发音人能够自然流畅地完成发音任务；其二，有效提升单次采集成功率，避免因表述卡顿导致的重复录制；其三，显著优化语音数据质量，为后续分析提供可靠基础。本研究对韵母（元音）的发音语料进行了有针对性的筛选，如表 1 藏语拉萨话韵母（元音）发音表所示。

表 1　藏语拉萨话韵母（元音）发音表

Table 1　Pronunciation of vowels in the Lhasa dialect of Tibetan

序号	韵母	藏文	拉丁转写	汉语
1	ཨ	རྟ	rta	马
2	ཨག	གཡག	gyag	公牦牛
3	ཨགས	ཕྱགས	phyags	扫
4	ཨང	ནང	nang	内部
5	ཨངས	གངས	gangs	雪
6	ཨད	སྐད	skad	声音
7	ཨན	སྨན	sman	药
8	ཨནད	འགྲན	vgrand	比赛
.....	.....	.....	.....	.....
93	ཨོར	འཁོར	vkhor	旋转
94	ཨོརད	སྐྱོརད	skyord	搀扶
95	ཨོལ	སོལ	sol	木炭
96	ཨོལད	དཀྲོལད	dkrold	弹
97	ཨོས	གོས	gos	衣服, 丝绸
98	ཨོའི	ཁོའི	khovi	生病
99	ཨོའུ	བོའུ	bovu	枪
100	ཨོའོ	བཞོའོ	bzhovo	挤奶呀

在发音人筛选方面，本研究制定了严格的入选条件。首要条件是发音人需以藏语拉萨话为第一语言，且未受普通话或其他语言/方言的语音干扰，能够流利使用纯正的拉萨话进行交流。此外，所有发音人均为在校大学生（年龄在 20–25 岁之间），需具备标准的拉萨话发音能力，要求吐字清晰、发音准确。在性别配比上，研究团队特别设置了男女均衡的样本结构。从生理条件考量，优先选择

体型偏瘦的发音人，因其下颌部位皮肤及肌肉组织较薄，更有利于超声舌成像设备的信号采集。基于上述标准，研究团队通过多轮招募与筛选，最终确定两位符合要求的发音合作者。所有入选者均为拉萨市中心城区长大的母语者（一男一女），在保证语音样本代表性的同时，也兼顾了声学数据采集的可行性。

1.2 数据采集系统及数据采集过程

超声舌成像视频采集采用加拿大 Ultrasonix Medical 公司研制的 SonixOP 超声成像系统（图 1），其系统组件涵盖 LCD 显示模块、带触控界面的操作控制台、移动式设备支架及换能器探头。其中核心成像组件选用 EC9-5 型凸阵换能器（Transducer）。



图 1　超声设备（左）、稳定头盔（中）、显示屏（右）

Figure 1　Ultrasound device (left), stabilization headset (middle), and display screen (right)

为确保获得最优舌体运动成像质量，研究团队依据表 2 所示的设备参数配置进行系统性调试，最终通过显示界面优化使舌体轮廓线与硬腭轮廓线达到最佳可视化效果。

表 2　超声设备参数设置

Table 2　Ultrasound device parameter settings

序号	参数名称	参数值
1	Freq 探头频率	5.0 M
2	Depth 深度	8.0 cm
3	Sector 扇区	100%
4	Gain 增益	50%
5	FrRate 帧频	中 25 Hz
6	Dyn 动态范围	80 dB
7	Persist 余辉	6
8	B-Mode Map B 型图	8
9	Chroma Map 色度图	1
10	SoS	1540m/s
11	Power 声能输出	0
12	MI 机械指数	<0.67

序号	参数名称	参数值
13	TIS 热指数-软组织	<0.53
14	柔化	关
15	放大	100%
16	Clarity 清晰度	High

语音音频采集，此部分使用了 Cool Edit Pro 软件（如图 2 所示）对语音音频进行了采集。为了确保音质清晰，同时使用了 Creative/创新 Blaster X-Fi Surround 5.1 Pro 声卡（如图 3 所示）、BEHRINGER/百灵达 Q502USB 声卡调音台（如图 4 所示）。



图 2　Cool Edit Pro 软件

Figure 2　Cool Edit Pro software



图 3　声卡

Figure 3　Sound card



图 4　调音台

Figure 4　Audio mixing desk

超声舌成像采集过程通常需要发音人端坐在设备前，保持头部稳定。技术人员会调整超声探头位置，将其轻轻置于发音人下颌下方，通过耦合剂确保声波有效传导。在采集过程中，发音人需按要求发出特定音节或保持静息状态，此时高频超声波会实时捕捉舌体运动轨迹，形成动态二维图像，为后续发音分析提供精确的舌位参数。在超声舌成像采集过程中，同时使用另一台计算机的音频软件完成发音人语音信号的采集工作。

2 数据样本描述

通过超声成像系统可以实现发音人舌体运动过程的动态记录，并以视频文件形式保存。每个视频序列包含 99 至 120 帧超声舌成像数据，图 5 呈现了以藏文韵母དཀྲིས发音（它的国际音标为[tʂiˀ]）语料为例的舌体运动动态采集结果。具体而言，图 6 展示了该语料第 45、58、68、80 帧的超声舌成像（对应上左、上右、下左、下右图），成像结果中舌体轮廓线呈现高清晰度与分层特征，为后续轮廓线提取工作奠定了坚实基础。

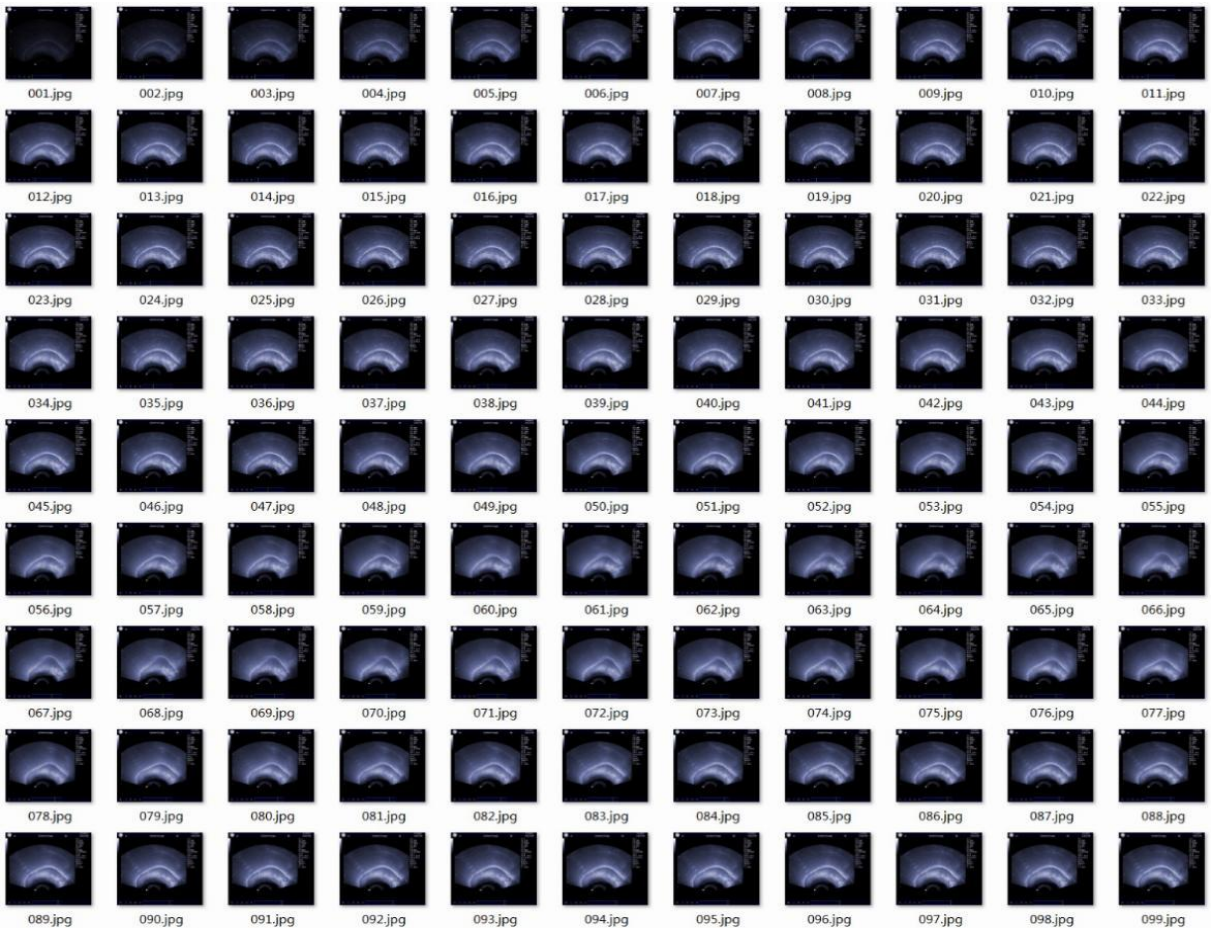

图 5　藏文དཀྲིས发音时的超声舌成像（第 1-99 帧）

Figure 5　Ultrasound tongue images during Tibetanདཀྲིས articulation (frames 1-99)

3 数据质量控制和评估

基于超声舌成像进行的所有科学研究，除了采用直接观察法外，最基础和最核心的工作是对超

声舌成像序列图像中的舌体轮廓线进行跟踪与提取。有了这些数据才能进行进一步的量化分析。本部分使用 Edgetrack 工具[19]对超声舌成像序列实施舌体轮廓线的自动化追踪与提取，相关处理结果如图 7 所示。图 7 清晰标注了舌体成像标志：上面黄色点状虚线代表上腭轮廓线，下面红色点状虚线代表舌体轮廓线。其中，舌体轮廓线的左侧显示的是舌根区域，舌体轮廓线的上方对应舌面轮廓，舌体轮廓线的右侧呈现的是舌尖位置。基于提取的轮廓线数据，可系统分析舌体的空间定位与形态特征，包括垂直高度、水平位移及整体姿态等关键参数，为量化研究提供了结构化数据支撑。

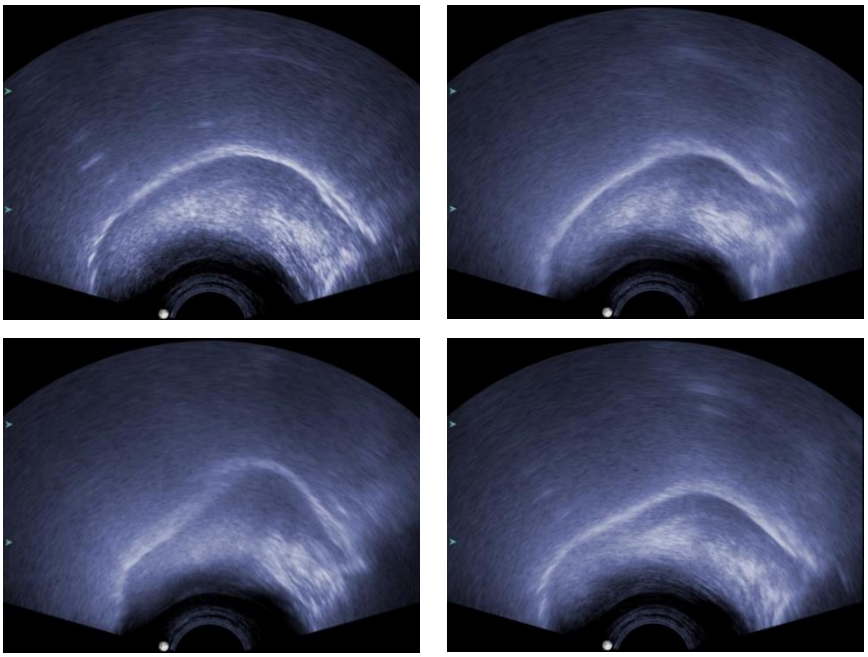

图 6　第 45、58、68、80 帧的超声舌成像（对应上左、上右、下左、下右图）

Figure 6　Ultrasound tongue images at frames 45, 58, 68, and 80 (corresponding to the upper-left, upper-right, lower-left, and lower-right panels, respectively)

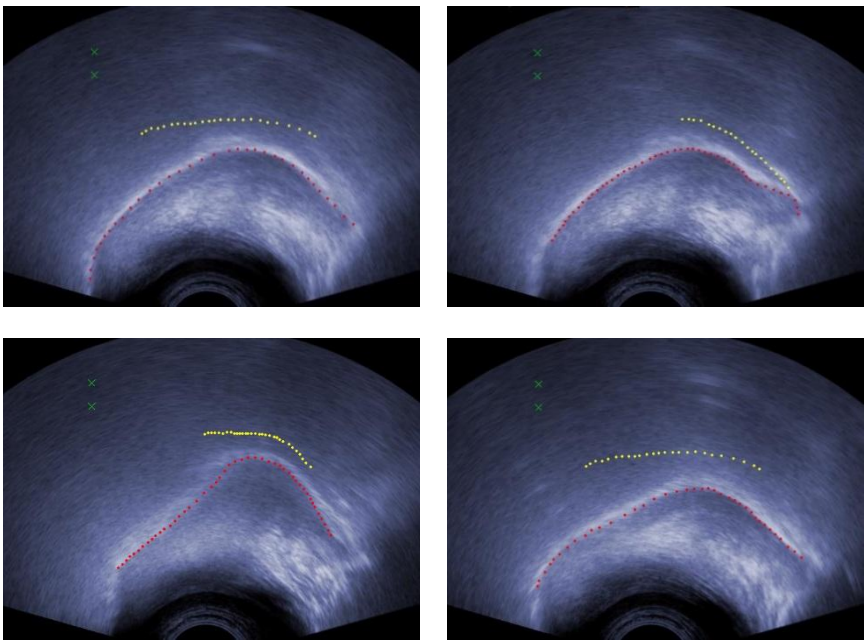

图 7　舌体轮廓线的跟踪与提取

Figure 7　Tracking and extraction of the tongue contour

基于前面跟踪与提取的数据可以做更深入的科学研究，比如，可以绘制出藏语拉萨话元音[i]发音稳定阶段的舌位极值点和上腭极值点，具体包括拉萨话元音发音稳定的舌位最高点、舌位最低点、舌位最远点、舌位最近点和舌位狭窄点，以及上腭最高点和上腭狭窄点，如图 8 所示；还可以对藏语拉萨话元音[i]发音稳定阶段的舌位曲线与其他几个阶段的舌位曲线相似度进行比较，如图 9 所示。

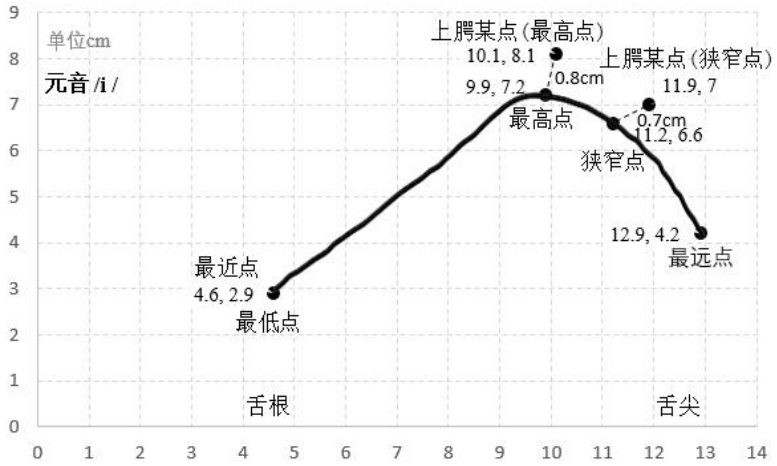

图 8　舌位极值点

Figure 8　Tongue position extrema points

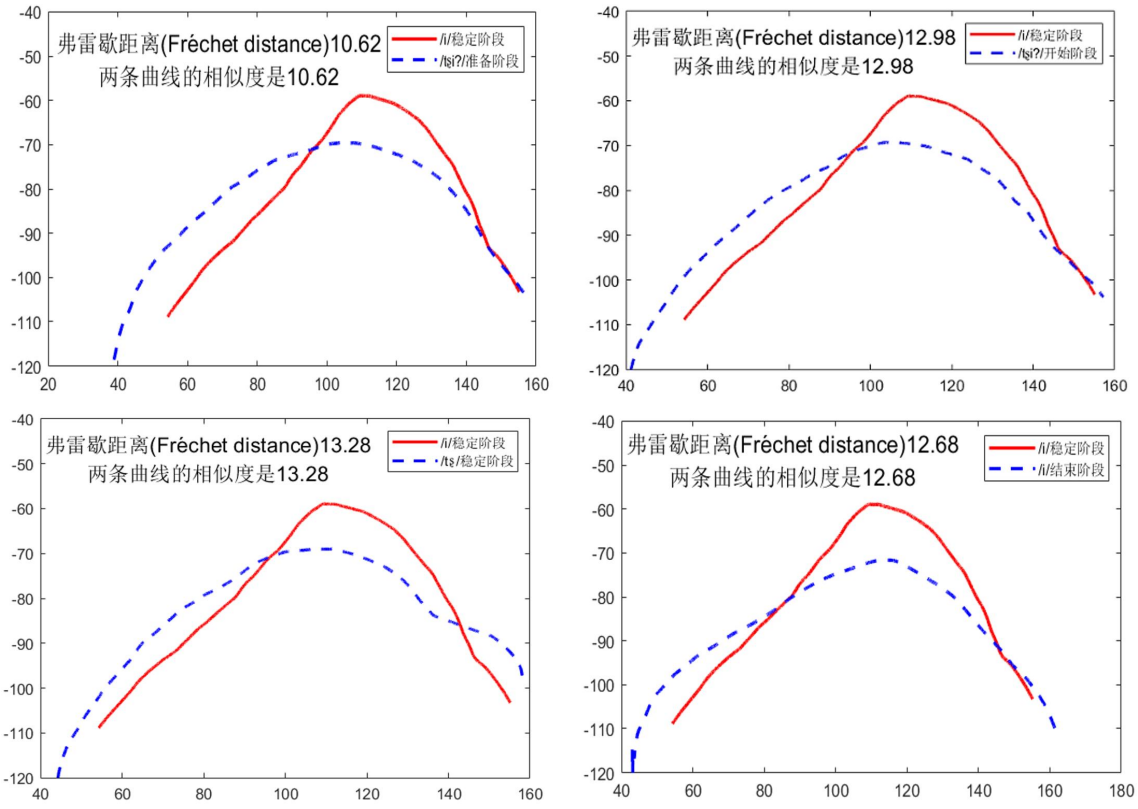

图 9　舌位曲线相似度

Figure 9　Tongue contour similarity analysis

4 数据价值

首先，从数据来源的创新性与独特性方面来看，国内外现有藏语语音数据集多以声学参数（如音频、共振峰）或文本语料为主，缺乏发音器官的动态生理数据；研究维度静态化，忽视协同发音过程中舌位连续变化的动态轨迹。本数据集采用医用级超声设备系统性采集藏语拉萨话元音的动态舌位运动数据，通过超声波检测技术捕获元音稳定段及协同发音时的舌形变化，弥补了藏语拉萨话发音生理数据的不足。

其次，在数据采集与质量控制体系构建中，本研究采用多模态同时采集技术，系统获取发音过程中的动态生理数据与声学信号。通过高精度超声舌成像设备与专业语音采集系统的同步联动，实现了舌部运动轨迹（超声模态）、语音声波（音频模态）与发音文本标注（语料模态）的三模态精准对齐，数据覆盖拉萨话元音音位的典型发音场景。在数据采集阶段，严格筛选符合方言母语者标准的发音人，通过声学验证与感知评估双重筛选，确保发音样本的典型性；在环境控制层面，采用专业录音棚标准构建声学环境，将背景噪声控制在最低水平；在数据处理阶段，实施双盲标注与交叉验证，对异常数据进行自动检测与人工复核，最终通过语音识别系统与超声运动轨迹的双重校验，确保数据质量。

再次，从语言科学与技术应用领域来看：在学术研究层面，该技术为语音学研究提供了动态观测手段，通过可视化舌位运动轨迹精准解析发音机制，既可构建汉-藏、藏-英等跨语言元音空间对比模型以深化语言类型学研究，又能为发音音系学建立实证基础。在语言教学领域，其生成的动态发音模型可优化藏语二语教学，通过实时反馈舌部运动特征辅助设计发音矫正工具；同时为临床语言学提供重要参照，既可建立拉萨话发音的生理基线数据，又能为言语障碍患者定制个性化康复方案。在技术赋能层面，该技术通过补充低资源藏语语音识别的生理特征维度，有效提升人工智能模型在噪声环境下的端到端识别鲁棒性，其跨模态数据还能增强语音-语义关联，推动藏语知识图谱的立体化构建。特别值得关注的是，在语言资源保护领域，该技术可建立包含动态发音生理数据的数字化语音库，为濒危方言的发音体系复原提供科学依据，助力构建“声音-动作-语义”的语言活态的保存体系。

最后，从数据集具体应用示例来看，本数据集具有广泛的应用场景[1-3]，这些应用场景或示例包括但不限于以下几方面。一是数据集可以作为语音合成模型的训练数据。本数据集提供的精确舌位运动信息，可用于训练更自然、更准确的藏语（拉萨话）发音合成器，特别是对元音音色的模拟。二是数据集可以作为语言学理论研究的实证基础。研究者可以利用本数据集，对拉萨话乃至藏语方言中元音的发音生理机制进行建模，验证或完善现有的发音语言学理论。三是数据集可以作为第二语言教学的辅助工具。本数据可以可视化地展示标准拉萨话元音的发音舌位，为藏语学习者提供直观的发音指导。四是数据集可以作为临床语音学的参考标准。在言语障碍治疗中，本数据集可作为一个健康的、标准的拉萨话发音生理参考，用于对比和诊断患者的发音异常。

总之，本数据集通过高精度多模态采集、跨学科数据处理方法（声学和生理）及严格质控，构建了聚焦藏语拉萨话元音发音生理的超声舌成像资源。本数据集兼具完备的内容体系（涵盖 CV、CVN 音节结构与协同发音特征）和多元应用场景（基础科研、教学实训、智能语音研发、言语临床诊疗、民族语言保护），能够助力藏语语音学研究从单一声学层面拓展至生理研究维度，是支撑相关领域跨越式发展的重要基础资源。

5 AIGC 使用声明

在撰写本手稿期间，作者未使用 AIGC 工具。所有手稿和数据均由作者独立完成，作者对论文的整体内容、数据真实性及学术诚信负全部责任。

作者分工职责

张金溪（1985—），男，山东省沂水县人，博士，副教授，研究方向为实验语音学、语音工程、深度学习。主要承担工作：数据采集、校对及评估，论文撰写，数据质量管理。

于洪志（1947—），女，山东省龙口市人，教授，博士生导师，研究方向为自然语言处理。主要承担工作：数据审核、数据综合控制。

张晶（1989—），女，甘肃省庆阳市人，博士，副教授，研究方向为实验语音学、藏语语音学。主要承担工作：发音语料筛选，发音人选择，数据筛选。

参考文献

[1] 张金溪, 李永宏, 寇赟. 基于超声成像的语言研究综述[J]. 西北民族大学学报(哲学社会科学版), 2019(6): 143–153. DOI:10.14084/j.cnki.cn62-1185/c.2019.06.014. [ZHANG J X, LI Y H, KOU Y. Overview of linguistic study based on ultrasound imaging technology[J]. Journal of Northwest Minzu University (Philosophy and Social Science), 2019(6): 143–153. DOI:10.14084/j.cnki.cn62-1185/c.2019.06.014.]

[2] 张金溪, 李永宏, 寇赟. 基于超声成像技术的发音生理舌体研究方法综述[J]. 计算机应用研究, 2021, 38(1): 15–22. DOI:10.19734/j.issn.1001-3695.2019.11.0651. [ZHANG J X, LI Y H, KOU Y. Research method of pronunciation physiological tongue based on ultrasonic imaging technique[J]. Application Research of Computers, 2021, 38(1): 15–22. DOI:10.19734/j.issn.1001-3695.2019.11.0651.]

[3] 张金溪, 张克宏, 于洪志. 深度学习在超声舌成像处理中的研究进展[J]. 中国图象图形学报, 2026, 31(2): 391–432. [ZHANG J X, ZHANG K H, YU H Z. Advances in deep learning techniques for ultrasound tongue imaging processing[J]. Journal of Image and Graphics, 2026, 31(2): 391–432.]

[4] 陈彧. 基于超声波检测的汉语普通话基础元音发音的舌体运动研究[D]. 天津: 南开大学, 2011. [CHEN Y. An investigation of togue movement in the production of Mandarin’s basic vowels using ultrasound[D]. Tianjin: Nankai University, 2011.]

[5] YANG Y D, SU R F, ZHAO S F, et al. An audio-ultrasound synchronized database of tongue movement for mandarin speech[J]. Scientific Data, 2025, 12: 607. DOI:10.1038/s41597-025-04917-w.

[6] 路文焕, 曲悦欣, 杨亚龙, 等. 无声语音接口中超声图像的混合特征提取[J]. 清华大学学报(自然科学版), 2017, 57(11): 1159–1162, 1169. DOI:10.16511/j.cnki.qhdxxb.2017.26.060. [LU W H, QU Y X, YANG Y L, et al. Hybrid feature extraction from ultrasound images for a silent speech interface[J]. Journal of Tsinghua University (Science and Technology), 2017, 57(11): 1159–1162, 1169. DOI:10.16511/j.cnki.qhdxxb.2017.26.060.]

[7] 骆明琼, 朱磊. 国外协同发音理论发展百年: 概览、问题与展望[J]. 外国语(上海外国语大学学报),

2025, 48(1): 58–71. [LUO M Q, ZHU L. Development of theories and models on coarticulation in western countries: overview, questions and future directions[J]. Journal of Foreign Languages, 2025, 48(1): 58–71.]

[8] 唐郅, 侯进. 基于深度神经网络的语音驱动发音器官的运动合成[J]. 自动化学报, 2016, 42(6): 923–930. DOI:10.16383/j.aas.2016.c150726. [TANG Z, HOU J. Speech-driven articulator motion synthesis with deep neural networks[J]. Acta Automatica Sinica, 2016, 42(6): 923–930. DOI:10.16383/j.aas.2016.c150726.]

[9] 谭亚鑫, 沈向荣. 基于超声动态舌形的普通话元音发音特点研究[J]. 听力学及言语疾病杂志, 2024, 32(4): 317–321. DOI:10.3969/j.issn.1006-7299.2024.04.006. [TAN Y X, SHEN X R. A study on pronunciation characteristics of mandarin vowels based on ultrasound tongue image[J]. Journal of Audiology and Speech Pathology, 2024, 32(4): 317–321. DOI:10.3969/j.issn.1006-7299.2024.04.006.]

[10] LI F J, SHEN F, MA D, et al. End-to-end mandarin speech reconstruction based on ultrasound tongue images using deep learning[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2025, 33: 140–149. DOI:10.1109/TNSRE.2024.3520498.

[11] MARCOUX I, MÉNARD L, LAPORTE C. Usability of two ultrasound tongue imaging devices in speech-language pathology[J]. American Journal of Speech-Language Pathology, 2025, 34(3): 1407–1424. DOI:10.1044/2025_ajslp-24-00293.

[12] 孙毅. 濒危隐喻抢救挖掘与语言资源整理保护[J]. 中国外语, 2025, 22(3): 37–44. DOI:10.13564/j.cnki.issn.1672-9382.2025.03.003. [SUN Y. Rescuing mining of endangered metaphors and protective collation of language resources[J]. Foreign Languages in China, 2025, 22(3): 37–44. DOI:10.13564/j.cnki.issn.1672-9382.2025.03.003.]

[13] ESHKY A, RIBEIRO M S, CLELAND J, et al. UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions[C]. Interspeech 2018. ISCA, 2018: 1888–1892. DOI:10.21437/interspeech.2018-1736.

[14] RIBEIRO M S, SANGER J, ZHANG J X, et al. Tal: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos[C]. 2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen, China. IEEE, 2021: 1109–1116. DOI:10.1109/SLT48900.2021.9383619.

[15] KIMURA N, SU Z X, SAEKI T, et al. SSR7000: a synchronized corpus of ultrasound tongue imaging for end-to-end silent speech recognition[C]. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France. 2022.

[16] 陈小莹, 艾金勇, 于洪志. 藏语拉萨话单音节嗓音声学参数分析[J]. 中文信息学报, 2015, 29(3): 184–189. [CHEN X Y, AI J Y, YU H Z. Acoustic analysis of monosyllabic voice of Lhasa Tibetan[J]. Journal of Chinese Information Processing, 2015, 29(3): 184–189.]

[17] 瞿霭堂, 劲松. 藏语卫藏方言研究[M]. 北京: 中国藏学出版社, 2017. [QU A T, JIN S. A study of the Ü-Tsang dialect of Tibetan[M]. Beijing: China Tibetology Publishing House, 2017.]

[18] 孔江平, 于洪志, 李永宏, 等. 藏语方言调查表[M]. 北京: 商务印书馆, 2011. [KONG J P, YU H Z, LI Y H, et al. Questionnaire of Tibetan dialects [M]. Beijing: The Commercial Press, 2011.]

[19] LI M, KAMBHAMETTU C, STONE M. Automatic contour tracking in ultrasound images[J]. Clinical Linguistics & Phonetics, 2005, 19(6/7):545–554. DOI:10.1080/02699200500113616.

论文引用格式

张金溪，于洪志，张晶. 藏语拉萨话元音发音时的超声舌成像数据集[J/OL]. 中国科学数据, 2026, 11(2). (2026-06-03). DOI: 10.11922/11-6035.csd.2025.0150.zh.

数据引用格式

张金溪，于洪志，张晶. 藏语拉萨话元音发音时的超声舌成像数据集[DS/OL]. V1. Science Data Bank, 2025. (2025-09-22). DOI: 10.57760/sciencedb.j00001.01606.

A dataset of ultrasound tongue imaging for vowel articulation in the Lhasa dialect of Tibetan

ZHANG Jinxi[1], YU Hongzhi[2,3*], ZHANG Jing[1]

1. School of Business and Media, Lanzhou University of Finance and Economics, Lanzhou 730020, P.R.China

2. Key Laboratory of Linguistic and Cultural Computing Ministry of Education (Northwest Minzu University), Lanzhou 730030, P.R.China

3. Key Laboratory of Minzu Languages and Cultures Intelligent Information Processing, Gansu Province (Northwest Minzu University), Lanzhou 730030, P.R.China

*Email: yhz@xbmu.cn (YU Hongzhi)

Abstract: This dataset is an ultrasound tongue imaging collection established to study vowel articulation features in the Lhasa dialect of Tibetan and support speech engineering research. The study employed dynamic ultrasound tongue imaging technology to capture articulatory organ (tongue) movement data from native speakers. Data collection was conducted in a standard recording studio environment with two Lhasa-native participants (one male and one female), all university students with Lhasa dialect as their mother tongue and no speech disorders. The experiment utilized B-mode ultrasound equipment (probe frequency: 5.0 MHz) to record real-time tongue movement images during articulation, paired with high-fidelity audio recording devices (sampling rate: 44.1 kHz) to capture speech signals. The dataset includes eight monophthongs ([a], [i], [u], [ɐ], [ø], [ɛ], [y], [o]) and two diphthongs ([au], [iu]), yielding 200 valid data units (including ultrasound videos, audio recordings, and corresponding text annotations). All data were collected under standardized protocols: participants maintained an upright sitting posture with head stabilization and ultrasound probe fixation, while speech intensity was controlled within 65-75 dB SPL. To ensure data quality, the collected data were double-checked by two phonetics experts to eliminate unqualified samples. This dataset fills a gap in articulatory research on Lhasa Tibetan speech production, providing essential foundational data support for Tibetan language phonetics teaching, speech pathology treatment, speech recognition and synthesis, and multimodal deep learning research. It can also be used for cross-language comparative studies of articulatory features, offers technical support for endangered language preservation, and holds significant value for linguistic research and speech engineering applications.

Keywords: Tibetan Lhasa dialect vowels; ultrasound tongue imaging; artificial intelligence; speech recognition and synthesis; multimodal deep learning

Dataset Profile

Title	A dataset of ultrasound tongue imaging for vowel articulation in the Lhasa dialect of Tibetan
Data corresponding author	YU Hongzhi (yhz@xbmu.cn)
Data authors	ZHANG Jinxi, YU Hongzhi, ZHANG Jing
Time range	2019
Data volume	8.15 GB (60.9 GB after decompression)
Data format	*.avi, *.xml, *.wav, *.pdf
Data service system	https://doi.org/10.57760/sciencedb.j00001.01606
Sources of funding	The National Natural Science Foundation of China (62366045, 62566057); the National Social Science Fund of China (19CYY039); the Central Government Guides Local Science and Technology Development Fund Project of Gansu Province (25ZYJA034); the Science and Technology Plan Project Lanzhou of City (2025-2-42); the Education and Teaching Achievement Cultivation Project of Gansu Province (2023GSJXCGPY-60); the Scientific Research Special Fund of Lanzhou University of Finance and Economics; the Longyuan Young Talents Project "Research on the Co-evolutionary Characteristics of Multi-ethnic Languages in Gansu Province".
Dataset composition	The dataset comprises data files from two participants. The data file for Speaker 1 (male) includes: (1) ultrasound1.zip – ultrasound video data containing 100 AVI files and 100 XML files (3.76 GB); (2) audio2.zip—audio recordings containing 100 WAV files (5.87 MB); (3) text3.zip—transcript files containing one PDF file (409 KB). The data files for Speaker 2 (female) includes: (1) ultrasound1.zip—ultrasound video data containing 100 AVI files and 100 XML files (4.37 GB); (2) audio2.zip—audio recordings containing 100 WAV files (6.25 MB); (3) text3.zip–transcript files containing one PDF file (409 KB).