



文献 CSTR:

32001.14.11-6035.csd.2025.0131.zh



文献 DOI:

10.11922/11-6035.csd.2025.0131.zh

数据 DOI:

10.57760/sciencedb.27649

文献分类: 语言学

收稿日期: 2025-07-08

录用日期: 2026-01-28

发表日期: 2026-02-09

阿布都外力·阿力木¹, 李冠宇^{1,2*}

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: 在少数民族群体汉语普通话习得规律研究及相关语音技术开发中, 高质量、针对性强的发音语料库是重要基础。然而, 当前针对维吾尔族大学生这一群体的汉语普通话发音系统性语料资源较为匮乏, 难以满足对其发音特点的深入分析与应用需求。为此, 本研究以新疆南疆地区 48 名维吾尔族大学生为发音人, 同时选取 2 名汉族学生作为汉语普通话标准音参照, 基于《普通话水平测试实施大纲》精选实验材料, 系统涵盖汉语普通话 39 个韵母、21 个声母及单字调、双字调, 通过专业录音设备在标准化环境下采集语音数据, 经切分去噪后形成 14550 个音频文件 (总时长 146 分钟, 数据量 595.6 MB) 及对应文本文件。语料库严格控制选字原则与发音变量, 为分析维吾尔族大学生汉语发音规律、开发语音教学资源及优化相关技术提供了高质量数据支撑, 对推动民族地区语言教育与跨语言信息处理具有重要价值。

关键词: 维吾尔族大学生; 汉语普通话; 发音语料库; 第二语言习得

数据库 (集) 基本信息简介

数据库 (集) 名称	维吾尔族大学生汉语普通话发音语料库
数据通信作者	李冠宇 (xxlgy@xbmu.edu.cn)
数据作者	阿布都外力·阿力木、李冠宇
数据时间范围	2024年
地理区域	新疆维吾尔自治区南疆地区
数据量	595.6 MB
数据格式	*.wav, *.docx
数据服务系统网址	https://doi.org/10.57760/sciencedb.27649
基金项目	2024年甘肃省科技重大专项计划 (24ZDFA004); 中央高校基本科研业务费项目 (31920220094)
数据库 (集) 组成	数据集共包括音频文件和文本文件, 其中, (1) wav是发音人录音数据, 包含14550个*.wav音频文件, 总时长为146分钟, 数据量为595.6 MB; (2) docx是对应的词表及发音人信息, 数据量为27 KB。

引言

* 论文通信作者

李冠宇: xxlgy@xbmu.edu.cn

习近平总书记在全国教育大会上明确提出, 要加大国家通用语言文字推广力度, 促进铸牢中华民族共同体意识^[1]。党的二十大报告提出“加大国家通用语言

文字推广力度”^[2]，对加大国家通用语言文字推广力度作出了战略部署，进一步彰显了推广普及国家通用语言文字工作对全面建设社会主义现代化国家的极端重要性^[3]。因此，在新疆地区开展维吾尔族大学生普通话使用情况研究，既为推动普通话在全国范围内同步通用奠定基础，也为实现新疆长治久安核心战略目标提供有力支撑。

语音数据作为语言研究与自然语言处理领域的基础性资源，在语音识别、语言习得规律分析、跨语言对比研究等方面具有不可替代的价值。我国作为统一的多民族国家，汉语普通话作为通用语言，其在少数民族群体中的习得与发音特点研究，对于促进民族间交流、优化教育模式、提升多语种语音技术性能具有重要意义。其中，维吾尔族作为我国新疆地区的主要少数民族之一，其母语与汉语在语音系统上存在显著差异，维吾尔族大学生的汉语普通话发音规律研究既是第二语言习得研究的重要课题，也是面向少数民族的汉语语音技术开发的核心基础。中国的二语习得研究必须紧密结合本土实践，回应时代的迫切需求^[4]。现有语音数据集多侧重于语音识别^[5]、语音合成^[6]以及语音翻译^[7-8]等相关的研究。这些数据集围绕特定语言，在各自聚焦的语音技术研究中充分发挥作用，推动相关领域从理论探索走向实际应用，但对于维吾尔族大学生汉语普通话发音研究方面缺乏针对性设计。在发音人语言背景、普通话水平等变量的记录，以及声韵母声学特征、声调基频模式等语音学关键要素的系统性覆盖上存在明显不足，导致难以支撑对该群体汉语发音规律的深层次解析，也难以给针对性语音教学与技术开发提供充分精准的数据支撑，在一定程度上限制了相关研究的深入推进。

相对而言，本数据集的建成不仅为维吾尔语母语者汉语普通话发音的声学特征分析、汉维语音对比研究提供了高质量实证材料，也为少数民族汉语语音教学资源开发、针对性发音矫正模型构建及面向维吾尔族群体的汉语语音识别系统优化奠定了数据基础。其系统性与针对性特点，将丰富少数民族大学生汉语语音精细化数据集的研究，助力推动民族地区语言教育现代化与跨语言信息处理技术的发展。

1 数据采集和处理方法

本数据集以《普通话水平测试实施大纲》为依据，选取常用词语作为语料来源，完整涵盖汉语普通话的所有韵母、声母与声调，包含音频文件及与之对应的文本文件。数据采集对象为 48 位新疆南疆地区的维吾尔族大学生，在西北民族大学专业录音室通过专业设备录制，经处理形成 14550 个音频文件。

1.1 发音词表来源

本实验单词均从国家语委普通话文字应用培训测试中心编写的《普通话水平测试实施大纲》（2021 年版）^[9]当中结合汉语的音节结构特点，选取比较常用的词语，包含所有的韵母、声母和声调（单字调和双字调）。

1.2 发音词表选字原则

（1）韵母表选字原则：按照韵母的三种类型列出了 39 个韵母，包括单元音韵母、复元音韵母和带鼻音韵母，没有考虑双音节词，单音节词选择声母时选择了塞音，如 b、d、p、t、g、k。尽量不选边音、鼻音、零声母或半元音，避免声母对实验的基频产生影响^[10]。如表 1 所示。

表 1 汉语普通话韵母实验词表的样本示例

Table 1 Sample examples of the experimental word list for mandarin finals

序号	韵母	例子
1	a	八, 搭, 塌
2	o	播, 泼, 摸
3	e	恶, 德, 特
4	i	逼, 低, 踢
5	u	逋, 嘟, 突
...	...	...

(2) 声母表选字原则: 本研究从汉语拼音系统中列出了 21 个声母作为实验材料, 在选择组合的韵母时, 优先考虑了低元音 a、前高元音 i 和后高元音 u, 在无法使用这些元音进行组合的情况下, 才转而选择了其他韵母。为了保持声调的一致性, 在声调选择上优先考虑了阴平, 在阴平不可用的情况下, 才选择了其他声调作为替代。如表 2 所示。

表 2 汉语普通话声母实验词表的样本示例

Table 2 Sample examples of the experimental word list for mandarin initials

序号	声母	例子
1	b	八, 逼, 逋
2	p	趴, 批, 普
3	m	妈, 咪, 木
4	f	发, 佛, 肤
5	d	搭, 低, 嘟
...	...	...

(3) 单音节词声调实验测试选字原则: 在前期选择单音节字的选字过程中, 本研究在选择声母时, 优先考虑了双唇塞音 b、舌尖塞音 d、双唇塞音 p 和舌尖塞音 t 等声母。根据朱晓农《语音学》(2010) 中提及的原则^[11], 实验首选例字的声母以塞音、擦音、塞擦音为宜。尽量不选择鼻音、边音, 因为鼻音和边音在发音时, 气流通过的路径与塞音、擦音和塞擦音不同。此外, 鼻音和边音在发音时声带都会振动, 产生浊音。浊音的声带振动会在语图上表现为连续的横杠, 这可能会与声调的周期性波动相混淆, 从而影响声调的分析。

在组合韵母时, 根据元音的音位特征, 优先选择了低元音[a]、前高元音[i]、后高元音[u], 在无法使用这些元音进行组合的情况下, 才转而选择了其他的韵母作为替代。如表 3 所示。

表 3 汉语普通话单音节词声调实验词表的样本示例

Table 3 Sample examples of the experimental word list for mandarin monosyllabic tones

序号	调类	例子
1	阴平	八, 逼, 逋, 搭, 低, 嘟, 趴, 批, 扑, 塌, 踢, 突
2	阳平	拔, 鼻, 醅, 达, 敌, 独, 爬, 皮, 葡, 桃, 题, 图

序号	调类	例子
...	...	...

(4) 双音节词声调实验测试选字原则：按照汉语普通话中阴平、阳平、上声、去声四个声调两两组合的形式，由此得到 16 组词语，其中每组设置 4 个词语，这样共计为 64 个词语，128 个字。词语搭配中声母选择清声母，包括了普通话中绝大部分的塞音、塞擦音和擦音^[12]，选择韵母时选择了低元音 a、前高元音 i、后高元音 u，为了保证 16 组搭配都有词语，在词语搭配不够的条件下，再放宽对韵母的限制。如表 4 所示。

表 4 汉语普通话双音节词声调实验词表的样本示例

Table 4 Sample examples of the experimental word list for mandarin disyllabic tones

调类	阴平	阳平	上声	去声
阴平	交织	剥离	公使	公式
	期间	凄凉	胶体	编制
	搭车	低头	包裹	波动
	编织	工时	驱使	交际
	...	...	...	...

1.3 发音人的选取

论文在研究中选择发音人时，从统计学的角度来考虑，如果拥有大规模的样本进行研究可能工作量稍大，难以完成。因此，在前期工作中选择实验发音人时，为了尽量兼顾所有学生，根据维吾尔族学生的普通话等级和年级，一共选择了 48 名学生（男女生各 24 名），其中普通话等级为二级甲等 10 名（5 男，5 女），二级乙等 16 名（8 男，8 女），三级甲等与三级乙等共有 22 名（11 男，11 女），都是新疆南疆地区的维吾尔族大学生。汉语普通话标准音选择就读于西北民族大学新闻传播学院广播电视学专业的 2 名学生（男女生各 1 名），普通话等级均为一级乙水平。本课题随机抽取的实验发音人均是来自西北民族大学的维吾尔族大学生，年龄在 18 岁至 23 岁左右，身体健康，发音器官正常，母语为维吾尔语，要求接受过双语教学，且平时用维吾尔语中心方言。虽然这些维吾尔族大学生基本上都掌握了汉语的使用，可以进行日常交流，但是从他们的发音中，仍然可以发现其母语口音的痕迹。

1.4 语音信号数据采集方法

语音信号采集工作在专业的录音室里进行，采集设备包括计算机、调音台、外置声卡以及指向性高灵敏度传声话筒，录音使用 Adobe Audition（2020 版本）软件，单通道，量化精度为 16Bit，采样频率为 44100 Hz。所选的实验测试字都进行随机排列，在正式录音之前，先让发音人充分熟悉实验材料，试读几遍，如果遇到不认识的字或词语，录音之前就向发音人解释清楚或者在词表上注音。为了保证录音质量，让发音人把每个词语读两遍。

录音完毕以后，请每位发音人详细填写他们的相关信息，将录音以 wav 格式存入电脑。如图 1 所示，用 Adobe Audition 软件按照实验方案选取最优的方式对录音进行切分，切分依据听辨，切除噪声段，如吞咽、咳嗽等，保留有效的语音片段，并以 wav 格式存入电脑。



图1 语音信号采集环境与软件界面

Figure 1 Recording environment and software interface for speech signal acquisition

2 数据样本描述

本数据集以维吾尔族大学生为研究对象，基于《普通话水平测试实施大纲》精选实验材料，系统涵盖汉语普通话所有韵母、声母及单字调、双字调等关键语音要素。在材料选取上，严格遵循语音学研究规范：韵母实验优先选择塞音声母以排除基频干扰，声母实验聚焦低元音 a 、前高元音 i 、后高元音 u 等核心韵母，声调实验则依据塞音、擦音等具有清晰声学边界的声母选字，确保数据对语音特征的精准反映。发音人选取兼顾性别均衡与普通话水平差异，并通过专业录音设备与标准化流程在专业录音室完成数据采集，最终形成包含 14550 条音频及对应文本的结构化数据集。

3 数据质量控制和评估

为确保本数据集语音及文本数据的可靠性与有效性，针对维吾尔族学生的语言背景特点，在数据质量控制与评估环节实施了多维度措施。录音阶段，考虑到维吾尔族学生母语（维吾尔语）与汉语语音系统的差异，并考虑到部分学生可能对汉语生僻字词不熟悉，提前让发音人熟悉实验材料，对不认识的字或词语进行解释并标注拼音，避免因词汇障碍导致发音偏差，确保其发音能真实反映汉语普通话习得水平。所有语音均在专业录音室内录制，同时要求每位学生对每个词语朗读两遍，便于后续从重复样本中筛选更贴合汉语普通话发音特征的语音片段。录音完成后，使用 Adobe Audition 软件进行精细化处理，通过听辨逐段切分音频，剔除吞咽、咳嗽等噪声部分，仅保留有效语音片段，保证音频数据纯净。此外，详细记录每位学生的普通话等级、教育背景等信息，为后续结合其语言习得特点开展针对性评估提供依据，确保数据集在维吾尔族学生汉语发音研究及相关应用中的可靠性。

4 数据价值

汉语二语习得要进行多视角研究^[13]。本数据集作为维吾尔族大学生汉语普通话发音的语音资源，弥补了该群体系统性语音数据的匮乏，为汉语普通话语音研究及相关应用提供重要支撑。在学术研究层面，14550 个音频文件及对应文本数据，可用于深入分析维吾尔族大学生普通话发音的声学特征，比如声母的冲直条表现、声调的基频模式，还能针对性探究常见偏误规律，如阳平与上声的调值混淆、声母 c 与 s 的发音区分难题，为揭示母语对汉语语音习得的影响、完善少数民族汉语习得

理论提供实证依据。此外，为进一步发挥数据集的长效价值，后续可基于其现有基础，深挖维吾尔族人易读错、易发音错的语音要素特征与规律，并逐步将发音群体拓展至接受继续教育的成人，让数据支撑覆盖更广泛的教育场景。在应用层面，数据集可直接服务于少数民族汉语语音教学，助力开发针对性发音矫正工具与教学资源，同时为面向维吾尔族群体的汉语语音识别、语音合成技术优化提供高质量训练数据，推动多语种语音技术在民族地区落地，对促进民族间语言交流与文化融合具有重要实践价值。

5 数据使用方法和建议

本数据集的使用可结合语音学研究需求，先依据韵母、声母、单字调及双字调等类别，对 14550 个 wav 音频文件及对应 docx 文本数据进行系统性整理，确保数据分类清晰、检索便捷。如图 2 所示，建议采用 Praat 语音分析软件对切分后的音频文件进行声学标注，生成 TextGrid 文件并与原音频文件同步保存，标注时需明确音节、声韵母、声调的层级设置。其中音节层对应完整语音单位，声韵母层精准划分发音成分，声调层标注调类属性，以实现语音片段与文本内容的精准对应。在声学参数提取方面，可通过 Praat 手动操作或编写脚本程序完成，推荐提取基频与共振峰两个参数。同时可引用 Praat 官网公开脚本库或相关语音学研究附载的参数提取脚本，提取的参数将为探究维吾尔族大学生汉语普通话发音中声母的声学边界、声调的基频模式等提供可靠量化依据。

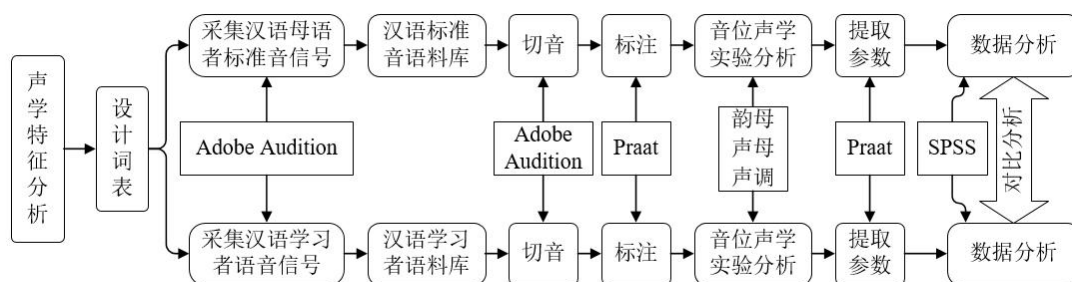


图 2 实验操作流程

Figure 2 Experimental workflow

致 谢

本数据集的构建获得西北民族大学维吾尔族同学的支持与帮助，在此表示由衷感谢！

作者分工职责

阿布都外力·阿力木（1997—），男，新疆阿克苏人，博士研究生，研究方向为实验语音学。主要承担工作：数据采集、数据标注与质量控制。

李冠宇（1973—），男，甘肃秦安人，博士，副教授，研究方向为语音识别、语音合成。主要承担工作：数据质量控制与综合管理。

参考文献

- [1] 薛冬雪. 在高校“大思政”体系建设中积极主动开展铸牢中华民族共同体意识教育[J]. 中国共青团, 2024(18): 24–25. [XUE D X. Actively carry out the consciousness education of building a strong Chinese national community in the construction of “great ideological and political” system in colleges and universities[J]. The Communist Youth League of China, 2024(18): 24–25.]
- [2] 习近平. 高举中国特色社会主义伟大旗帜为全面建设社会主义现代化国家而团结奋斗——在中国共产党第二十次全国代表大会上的报告[M]. 北京: 人民出版社, 2022. [XI J P. Hold High the Great Banner of Socialism with Chinese Characteristics and Strive in Unity to Build a Modern Socialist Country in All Respects: Report to the 20th National Congress of the Communist Party of China [M]. Beijing: People's Publishing House, 2022.]
- [3] 王启涛. 民族地区推广和普及国家通用语言文字工作的回顾和展望[J]. 西南民族大学学报(人文社会科学版), 2023, 44(7): 1–18. DOI:10.3969/j.issn.1004-3926.2023.07.001. [WANG Q T. Review and prospect of the work of promoting and popularizing the national common language and characters in ethnic areas[J]. Journal of Southwest University for Nationalities, 2023, 44(7): 1–18. DOI:10.3969/j.issn.1004-3926.2023.07.001.]
- [4] 杨连瑞, 张雨. 中国二语习得: 回顾过去, 展望未来[J/OL]. 现代外语, 2025: 1–10. (2025-04-29). DOI: 10.20071/j.cnki.xdwy.20250427.011. [YANG L R, ZHANG Y. China's second language acquisition: looking back and looking forward to the future[J/OL]. Modern Foreign Languages, 2025: 1–10. (2025-04-29). DOI: 10.20071/j.cnki.xdwy.20250427.011.]
- [5] 刘志强, 马志强, 张晓旭, 等. 蒙古语语音语料库 IMUT-MC[DS/OL]. V1. Science Data Bank, 2022 [2025-07-22]. DOI: 10.11922/sciencedb.j00001.00349. [LIU Z Q, MA Z Q, ZHANG X X, et al. Mongolian Speech Corpus IMUT-MC[DS/OL]. V1. Science Data Bank, 2022[2025-07-22]. DOI: 10.11922/sciencedb.j00001.00349.]
- [6] 韦向峰, 袁毅, 张全, 等. 基于端点检测的蒙藏维语音片段数据集[DS/OL]. V1. Science Data Bank, 2019[2025-07-22]. DOI: 10.11922/sciencedb.788. [WEI X F, YUAN Y, ZHANG Q, et al. A dataset of Mongolian, Tibetan and Uyghur speech fragments based on voice activity detection [DS/OL]. V1. Science Data Bank, 2019[2025-07-22]. DOI: 10.11922/sciencedb. 788.]
- [7] 朱丽平, 李宁. 中-蒙和维-汉语音翻译数据集[DS/OL]. V2. Science Data Bank, 2022[2025-07-22]. DOI: 10.11922/sciencedb.j00001.00356. [ZHU L P, LI N. Chinese-Mongolian and Uyghur-Chinese Speech Translation Dataset [DS/OL]. V2. Science Data Bank, 2022[2025-07-22]. DOI: 10.11922/sciencedb.j00001.00356.]
- [8] 戚肖克, 特尼格尔, 孙媛, 等. 蒙汉语音翻译数据集[DS/OL]. V2. Science Data Bank, 2022. (2025-07-22). DOI: 10.11922/sciencedb.j00001.00345. [QI X K, TE N G E, SUN Y, et al. A dataset of Mongolian-Chinese speech translation[DS/OL]. V2. Science Data Bank, 2022. (2025-07-22). DOI: 10.11922/sciencedb.j00001.00345.]
- [9] 国家语委普通话文字应用培训测试中心. 普通话水平测试实施大纲[M]. 北京: 北京语文出版社, 2021. [National Language Commission's Mandarin Language Application Training and Testing Center. Outline for Putonghua Proficiency Test[M]. Beijing: Beijing Language Press, 2021.]

- [10] 贺晨吟. 西班牙语国家留学生习得汉语声调偏误实验分析[D]. 广东: 暨南大学, 2009. [HE C Y. Experiments and Analysis of Spanish-speaking Learners' Acquisition of Mandarin Tones[D]. Guangdong: Jinan University, 2009.]
- [11] 朱晓农. 语音学[M]. 北京: 商务印书馆, 2010. [ZHU X N. Phonetics[M]. Beijing: The Commercial Press, 2010.]
- [12] 廖艳莎. 拉萨藏族大学生汉语普通话声调偏误的实验研究[D]. 甘肃: 西北民族大学, 2011. [LIAO Y S. Experiments and Researches of Mandarin Tonal Errors among Lhasa Tibetan College Student [D]. Gansu: Northwest Minzu University, 2011.]
- [13] 赵杨, 姜帅. 汉语二语习得研究的现状与展望[J]. 现代外语, 2025, 48(4): 577 - 585. DOI:10.20071/j.cnki.xdwy.20250427.002. [ZHAO Y, JIANG S. Status-quo and prospect of research on Chinese as a second language[J]. Modern Foreign Languages, 2025, 48(4): 577 - 585. DOI:10.20071/j.cnki.xdwy.20250427.002.]

论文引用格式

阿布都外力·阿力木, 李冠宇. 维吾尔族大学生汉语普通话发音语料库[J/OL]. 中国科学数据, 2026, 11(1). (2026-02-09). DOI: 10.11922/11-6035.csd.2025.0131.zh.

数据引用格式

阿布都外力·阿力木, 李冠宇. 维吾尔族大学生汉语普通话发音语料库[DS/OL]. V2. Science Data Bank, 2025. (2025-10-24). DOI: 10.57760/sciencedb.27649.

A corpus of mandarin pronunciation by Uyghur college students

Abuduwaili ALIMU¹, LI Guanyu^{1,2*}

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

Abstract: In research on the acquisition patterns of mandarin by ethnic minority groups and the development of related speech technologies, high-quality and highly targeted pronunciation corpora serve as an important foundation. However, systematic corpora for mandarin pronunciation specifically designed for Uyghur college students remain scarce, making it difficult to meet the needs of in-depth analysis and practical applications of their pronunciation characteristics. To address this gap, this study constructs a dedicated corpus of mandarin pronunciation by Uyghur college students. The corpus includes recordings from 48 Uyghur college students from southern Xinjiang as primary speakers, with 2 Han students selected as reference speakers for standard mandarin pronunciation. Based on the “Outline for Putonghua Proficiency Test”, we carefully selected experimental materials to systematically cover 39 finals, 21 initials, as well as monosyllabic and disyllabic tone patterns in mandarin. Speech data were recorded using

professional equipment in a standardized acoustic environment. After segmentation and denoising, the corpus comprises 14,550 audio files (with a total duration of 146 minutes and a data volume of 595.6MB), along with corresponding transcription files. The corpus strictly controls character selection principles and pronunciation variables, providing high-quality data support for analyzing the mandarin pronunciation patterns of Uyghur college students, developing mandarin pronunciation teaching resources, and optimizing related speech technologies. Overall, the corpus holds great value for promoting language education in ethnic minority regions and cross-lingual speech and information processing research.

Keywords: Uyghur college students; mandarin; pronunciation corpus; the second language acquisition

Dataset Profile

Title	A corpus of mandarin pronunciation by Uyghur college students
Data corresponding author	LI Guanyu (xxlgy@xbmu.edu.cn)
Data authors	Abuduwaili ALIMU, LI Guanyu
Time range	2024
Geographical scope	Southern Xinjiang Uygur Autonomous Region, China
Data volume	595.6 MB
Data format	*.wav, *.docx
Data service system	https://doi.org/10.57760/sciencedb.27649
Sources of funding	Gansu Province Science and Technology Major Special Program in 2024 (24ZDFA004); the Fundamental Research Funds for the Central Universities (31920220094).
Dataset composition	The dataset includes audio files and text files. Among them, (1) the wav files consist of recording data of speakers, containing 14,550 *.wav audio files with a total duration of 146 minutes and a data volume of 595.6MB; (2) the docx files consist of the corresponding word lists and speaker information, with a data size of 27 KB.