



开放科学
(资源服务)
标识码
(OSID)

新疆汉语方言数字化平台建设研究

郭辉¹ 徐志宏² 洪一旌³ 于菲⁴ 刘飞⁵

1. 新疆医科大学国际教育学院 乌鲁木齐 830017;
2. 新疆医科大学图书馆 乌鲁木齐 830017;
3. 新疆医科大学人事处 乌鲁木齐 830017;
4. 新疆克拉玛依区第八小学 克拉玛依 834000;
5. 新疆医科大学资产处 乌鲁木齐 830017

摘要: [目的/意义] 新疆汉语方言数字化平台的建设对于保护和传承地方语言文化具有重要的理论和实践意义, 有助于新疆历史文化研究以及新疆语言文化的保护与传承。[方法/过程] 本文在介绍新疆汉语方言历史演变、比较研究概况, 在汉语方言数据库建设现状的基础上, 提出了建设新疆汉语方言数字化平台的总体框架以及功能设计。[结果/结论] 指出了在平台建设过程中可能遇到的挑战, 并提出了相应的解决方案, 以期对汉语方言的保护及数字化平台建设起到一定的借鉴作用。

关键词: 新疆汉语方言; 数字化平台; 语言保护

中图分类号: G35

Research on the Construction of Xinjiang Chinese Dialect Digital Platform

GUO Hui¹ XU Zhihong² HONG Yijing³ YU Fei⁴ LIU Fei⁵

1. International Education College, Xinjiang Medical University, Urumqi 830017, China;
2. Library of Xinjiang Medical University, Urumqi 830017, China;
3. Personnel Department, Xinjiang Medical University, Urumqi 830017, China;
4. Karamay No. 8 Primary School, Karamay 834000, China;
5. Assets Management Department, Xinjiang Medical University, Urumqi 830017, China

Abstract: [Objective/Significance] The construction of the digital platform for Xinjiang Chinese dialects holds significant theoretical and practical value for the protection and inheritance of local linguistic and cultural heritage. It provides robust support for

基金项目 教育部人文社科研究项目“新疆汉语方言语音数据库建设研究”(21XJJCZH001)。

作者简介 郭辉(1980-), 硕士, 讲师, 主要研究方向为教育教学管理; 徐志宏(1992-), 本科, 助理馆员, 主要研究方向为信息管理; 洪一旌(1984-), 硕士, 讲师, 主要研究方向为教育管理研究, 人力资源管理。于菲(1984-), 本科, 中小学一级教师, 主要研究方向为数学教育; 刘飞(1981-), 通讯作者, 本科, 研究馆员, 研究方向为教育管理研究、新疆汉语方言, E-mail: lf0118@163.com。

引用格式 郭辉, 徐志宏, 洪一旌, 等. 新疆汉语方言数字化平台建设研究[J]. 情报工程, 2026, 12(1): 71-85.

the research, preservation, and inheritance of Xinjiang's historical culture and linguistic resources. [Methods/Processes] Based on an overview of the historical evolution and comparative research of Xinjiang Chinese dialects, as well as the current development status of Chinese dialect database construction, this study proposes the overall framework and functional design for building the digital platform. [Results/Conclusions] Furthermore, it identifies potential challenges in the platform construction process and puts forward corresponding solutions, aiming to offer valuable references for the protection of Chinese dialects and the development of similar digital platforms.

Keywords: Xinjiang Chinese Dialect; Digital Platform; Language Protection

引言

新疆地处中国的西北部，是一个多民族、多语言、多文化交汇的地区，新疆汉语方言（指汉语在新疆地区形成的方言，又称新疆官话、新疆话）不仅仅是一种交流工具，更是承载着丰富历史文化内涵的活态遗产，它们在语言学研究、文化传承、社会交流等方面具有重要的研究价值。新中国成立以来，关于新疆汉语方言的调查研究成果斐然，但还存在明显的不足之处，主要表现为当前对新疆汉语方言的研究长期局限于书面和口头的狭窄范围，新疆汉语方言数字化平台建设尚属空白，这不利于保护和传承新疆文化。随着现代化和城镇化进程的推进以及国家通用语言文字的推广^[1]，方言使用人群逐渐减少，许多方言面临消失的危险。因此，为响应习近平总书记在第三次中央新疆工作座谈会中强调的“要做好文化润疆工作”的号召，为增强新疆各族群众对伟大祖国的认同、对中华民族的认同、对中华文化的认同，需建立一个系统的数字化平台来记录、保存和传播这些方言，相关工作具有十分紧迫的现实意义。

1 新疆汉语方言的历史演变

新疆汉语方言的历史演变是一个复杂且多

层次的过程，这一过程受到地理、历史、民族迁徙以及社会文化等多方面因素的影响。新疆地处亚洲内陆，历史上是多个民族、文化交汇的要地，这种多元化的背景使得新疆汉语方言在形成和演变过程中展现出独特的特征^[2]。通过对新疆汉语方言历史演变的研究，可以为新疆汉语方言数字化平台建设中的语料分类与层级架构设计、方言溯源标注、演变规律可视化呈现等工作提供依据，有助于更科学地对数字化平台进行定位和描述。

1.1 西汉时期

汉语方言传入新疆的时间可以追溯到西汉时期，随着丝绸之路的开通，来自中原的驻军、官吏、商人及随军人员纷纷涌入新疆地区，他们带来了中原的语言和文化。这一时期汉语的传播主要集中在南疆地区，尤其是在塔里木盆地一带，为新疆汉语方言的早期发展奠定了基础。

1.2 唐代时期

到了唐代，新疆地区的汉语传播和使用进一步得到普及。唐朝在新疆设立了安西都护府等行政管理机构，吸引了更多汉人迁入新疆，唐代的汉语与当地语言相互交融，形成了早期

的新疆汉语方言。受突厥语等邻近民族语言的影响，新疆汉语方言的语音和词汇开始出现区域性差异。

1.3 清代时期

清代是新疆汉语方言发展的重要时期。清政府收复新疆后，大量内地东部、中部地区的汉族、回族、满族、锡伯族等移民来到新疆开垦田地，驻守边疆，特别是陕甘地区的移民，他们带来了西北方言。同时在南疆的维吾尔族也逐渐向北疆迁徙，迁徙人口众多，地域分布广泛。这一时期的移民潮使得新疆汉语方言的基础更加稳固，方言的内部差异也逐渐显现。南疆和北疆的方言在语音、词汇和语法上开始展现出不同的特点。

1.4 近现代时期

新疆汉语方言的演变受到屯垦戍边政策和人口流动的显著影响。20世纪中叶，新疆生产建设兵团的成立吸引了全国各地大量人口，带来了不同的方言和语言习惯。随着普通话推广政策的实施，普通话在新疆的普及对当地方言产生了深远的影响。当前，新疆汉语方言仍在不断演变中，受信息技术和新媒体的影响，年轻一代已开始在方言使用中体现出新的变化趋势。面对这样的变化，研究和保护新疆汉语方言的历史演变不仅有助于理解语言的多样性，也为文化传承提供了重要的参考。

2 新疆汉语方言比较研究概况

与普通话、其他方言、古代汉语、民族语言相比，新疆汉语方言在语音、词汇和语法等

方面展现出独特的魅力。这种比较研究有助于深入了解新疆的语言文化融合现象以及多元文化的交流互动，其研究方法还可以进一步借鉴现代语言学的先进技术和理论，如语音分析软件在语音研究中的应用，语料库语言学方法在词汇和语法研究中的运用等。

2.1 语音比较研究

梳理相关文献之后可以发现，目前关于新疆汉语方言语音比较研究已经非常丰富。其中共时比较研究相对于历时比较研究而言更为深入。共时比较研究中，第一个研究方向是方言片区内部不同本土语言的同异比较，主要表现为新疆汉语方言与以乌鲁木齐为代表的新疆其他城市方言的语音比较研究，其研究侧重于对语音、声调的同异比较。共时比较研究的第二个研究方向是不同方言片区语音的同异比较，其研究主要侧重于新疆中原官话南疆片、兰银官话北疆片、北京官话兵团片三个方言片在声母、韵母、声调等方面的同异比较，从而确定方言语音形成的历史渊源。共时比较研究的第三个研究方向是方言与普通话的比较研究，这也是新疆汉语方言语音共时比较研究的主体部分，这部分研究更多的是侧重新疆汉语方言与普通话在语音上声、韵、调等方面的差异，总结出新疆汉语方言语音的显著特点。共时比较研究的第四个研究方向是方言与周边民族外部语言的比较研究，目前这部分研究还比较薄弱，仅有新疆汉语方言与维吾尔语、湘语的语音比较研究^[3]。其中新疆汉语方言与维吾尔语的语音比较研究探寻了两种语言之间产生异同的原因，进一步明确两种语言各自的特点。新疆汉

语方言与湘语的比较研究,明确了大学生英语录音材料中的韵律差别,探寻方言对英语录音材料造成的具体阻碍^[4]。在历时比较方面,目前的研究主要侧重于中古音与方言比较,而且其研究多以新疆乌鲁木齐为研究区,因乌鲁木齐是新疆的自治区首府,人口众多且来源广泛,其研究更具代表性。

2.2 词汇比较研究

通过梳理文献可以发现,目前新疆汉语方言的词汇比较研究已经比较深入,其中词汇的内部比较研究成果更为丰富,主要集中在新疆汉语方言不同片区之间、新疆汉语方言与本土方言之间以及新疆本土方言之间的比较研究,其研究内容多为特征词、语义特点、词汇的结构特征等方面的比较研究^[5]。而新疆汉语方言词汇的外部比较研究成果相对较为薄弱,其研究内容只侧重于形容词、日常词汇与特征词的比较研究。

2.3 语法比较研究

新疆汉语方言语法的比较研究文献较少,视野较窄。要么是简单对比研究新疆汉语方言的典型语法现象,进一步探寻产生这种典型语法现象的原因;要么是通过对比研究新疆汉语方言语法差异,进一步研究这种语法差异可能导致的不同影响。新疆汉语方言语法的比较研究中,其与维吾尔语在句式方面的语法比较研究更为详细,最根本的原因是新疆汉语方言语法多与普通话相通,其显著的差异特点主要体现在带“子”后缀、单音节名词习惯重叠等方面^[6]。

2.4 历史比较研究

目前,关于新疆汉语方言的历史比较研究主要侧重于在历史层面对新疆汉语方言的形成进行研究,通过追溯新疆汉语方言形成的历史,梳理出语音随时间变化的过程,进一步突出新疆汉语方言的特点,探寻出新疆汉语方言发展的趋势^[7]。但是在揭示新疆汉语方言与其他方言的历史层次以及方言之间关系方面还比较薄弱。

2.5 地域文化比较研究

新疆汉语方言的地域文化比较研究成果较少,研究主要体现为新疆维吾尔语、奇台汉语方言与新疆汉语方言的比较研究,从方言的文化角度出发,比较研究出新疆汉语方言有别于其他方言的独特语言现象,通过语言现象反映出方言独具特色的深厚文化内涵。

3 汉语方言数据库建设现状

3.1 国内汉语方言数据库建设现状

改革开放以来,我国汉语语音语料库建设取得了长足发展。北京语言大学通过20世纪80年代采录的北京口语材料建成了“北京口语语料库”,现已对社会开放。中国社会科学院语言研究所在2005年启动了“现代汉语口语语料库”项目。中国科技大学讯飞语音公司开发的汉语语音库在近10年已完成市场化。上海方言数据库“依好学堂”是地方特色鲜明的方言数据库,以上海方言为主要收录对象,对上海方言的语音、词汇、语法等方面进行了全面而

深入的记录和整理。国家语言文字工作委员会2008年启动建设的中国语言资源有声数据库是我国重要的语言资源保护项目成果之一。目前，汉语方言语音语料数据库的理论研究和实际研发都取得了不错的成绩，主要表现在方言语音、词汇、语法和俗语等方面。

3.1.1 语音数据库的研发概况

以方言语音为主要内容研发的数据库有侯精一开发的“现代汉语方言音库CD”，这个数据库是以录音的形式，准确记录和细致描述了现代汉语40个地点方言的基本面貌和主要特点，为学术界提供了一份极为珍贵的现代汉语方言有声资料。蒋平主持的“汉语方言声调资料库”旨在考查声调的普遍行为，检验前人所概括的关于声调的普遍规律。而中国科技大学研发的“粤语语音合成系统语料库”介绍语料库设计的原理和过程，结合粤语语音合成系统语料库的实现，提出“语境矢量”的独特设计和“语境总量”的概念以及计算方法。刘俐李主持的“现代汉语方言核心词·特征词集”则侧重于对汉语方言词汇的研究。刘村汉主持开发的“方言字音Excel处理系统”更多的是研究方言的字词和发音数字化。海柳文开发的“汉语方言民族语言语音材料处理软件”则通过设计计算机软件，将方言的语音材料一键式处理输出成txt、xls等格式文件。

3.1.2 词汇数据库的研发概况

词汇数据库简称词库，又可分为单语词库和多语词库，我国目前对方言词汇数据库的研发主要集中在单语词库上，产生了社科院主持的“地方普通话语音语料库”、潘悟云开发的“汉语方言计算机处理系统”、麦耘主持的“汉

语方言词汇数据库”等一系列大型词库。这些数据库有的以广泛收集方言词汇语料为主，如“地方普通话语音语料库”共涵盖10个城市，每个城市200名发音人，朗读约1900~2200句口语化语料；有的以词汇的不同音节建库，如“汉语方言词汇数据库”是按词语单音节、双音节和多音节分类建设的。词汇数据库的研制不仅方便了使用者对方言词语的查询，还提供了一些分析功能，如“地方普通话语音语料库”在存储汉语方言词汇语料的基础上，还对语音数据进行了正则汉字和拼音标注，并对典型数据作了精细标注，包括音节和声韵母的音段切分 and 实际发音的标注；“汉语方言词汇数据库”建立的语义类代码库和子数据库为方言数据库的词性标注和逻辑结构设计提供了例证。在“汉语方言计算机处理系统”中，使用者可直接实现网上方言词汇查询，是方言数据库和门户网站结合的先驱。

3.1.3 语法和俗语数据库的研发概况

以方言语法和俗语为内容的数据库，目前研制的还不多，比较规范的有刘丹青主持的“汉语方言语法特征语料库”、南京师范大学研开发的“汉语俗语语料库”和“现代汉语日常语域信息库”等。这类语料库语料收集难度大、标注比较复杂，但随着方言数据库技术的提高和标注原则的普及，以方言语法和俗语等为主要内容的数据库正在逐步出现。上海师范大学语言研究所潘悟云研制的“汉语方言计算机处理系统”有单机版和网络版两种版本，收录了若干汉语方言和民族语言词汇，可以实现词汇的查询。“汉藏语同源词研究·词汇语音数据库”由香港科技大学丁邦新教授和中国社会科学院

孙宏开教授主持，中国社会科学院江荻教授设计、研制，20多位专家学者参与整理、提供、核对数据，收录了汉藏语系122种语言和12种汉语方言的1500余条词汇，可以实现语音、语义单项和多项组合查询，是目前国内功能最强大的多语词汇数据库。

3.2 汉语方言数据库建设的缺憾

对国内外相关文献梳理之后发现，目前关于汉语方言语音语料数据库的研究仍然存在一些缺憾。对典型多民族地区的汉语方言语音语料数据库的研究还比较少，收词量偏少，或方言和语种偏少，数据结构比较简单，对每个词的音节特征、结构特征和语法语义特征缺乏全面标注，功能不够强大，词汇多数为音标标注，缺乏声音数据，大多为单机系统，没有面向所有读者的开放式网络版，或用户无法在网络上实现声音或文本数据的添加，系统的研制由不同单位或个人各自实施，缺乏统一数据标准，使用起来多有不便，有必要对词汇数据库的有关数据和功能制定一个统一的标准。

4 新疆汉语方言数字化平台建设方案

4.1 平台架构设计

新疆汉语方言数字化平台的架构设计综合考虑了平台的功能性、可扩展性和用户友好性^[8]。平台架构采用模块化结构，主要由数据层、应用层和展示层构成，各层协同工作，实现新疆汉语方言数据的高效管理、处理与展示，并充

分考虑新疆汉语方言在语音、词汇、语法等方面的地域特征，确保平台功能的针对性与实用性，如图1所示。

4.1.1 数据层

数据层负责存储和管理庞大的方言数据，采用分布式数据库技术以确保数据的安全性和一致性。数据库需支持多种数据格式，如音频、文本和视频，以适应不同类型的方言资源。

(1) 分布式数据库：采用分布式存储技术，确保数据的安全性与一致性。该数据库负责存储音频、文本、视频等多种格式的新疆汉语方言数据，以及词汇数据库、语法知识库中的结构化数据，实现各类数据的统一管理 with 高效调用。

(2) 音频存储方式（定制编码）：选用能精准保留新疆汉语方言语音细微特征的定制编码方式，如基于Opus的定制编码。该方式存储大量方言音频数据，为语音识别模块提供原始语音数据支持，确保方言语音的独特韵味得以完整保存。

(3) 词汇数据库（含特色字段）：建立专门的词汇数据库表结构，除存储常规词汇信息外，增加词汇文化背景、来源等特色字段。该数据库方便对具有地域文化特色的词汇进行管理与研究，为词汇检索算法模块和语义分析模块提供数据基础。

(4) 语法知识库（语义网络存储）：以语义网络形式存储新疆汉语方言语法规则和例句，构建语法知识体系。该知识库方便语法分析模块调用，用于语法结构解析与语义分析，同时为用户提供深入了解方言语法的资源。

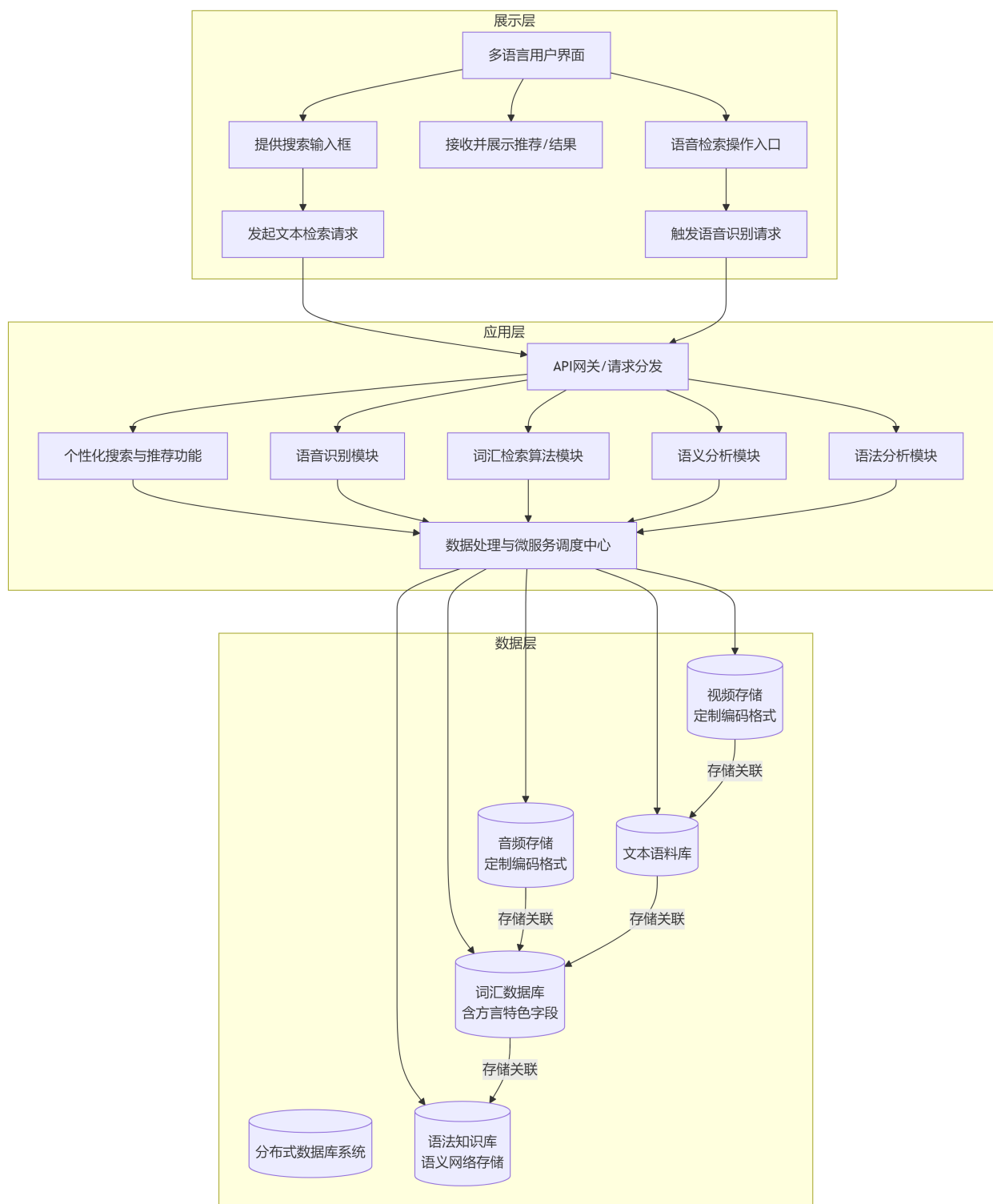


图1 新疆汉语方言数字化平台架构设计

4.1.2 应用层

应用层是平台的核心，提供数据处理和分

析的功能。该层通过先进的自然语言处理技术和机器学习算法，能够实现方言数据的自动标

注、语音识别和语义分析等功能。同时,为了提高数据处理的效率,应用层采用微服务架构,支持功能的独立部署和动态调整。

(1) 语音识别模块:采用针对新疆汉语方言语音地域特征训练的深度学习模型,如 TensorFlow 预训练完成的 SpeechCommands 声学模型,能将识别后的语音数据传输至数据处理模块进行后续处理。

(2) 语义分析模块:结合新疆汉语方言词汇语义特点及语法结构,分析用户输入内容或方言文本的语义。例如,对于具有多义性的方言词汇,能依据上下文准确判断语义,并与词汇数据库和语法知识库关联,深入挖掘语义内涵。

(3) 词汇检索模块:支持多维度词汇检索,用户可按词汇来源(如少数民族语言借词、历史移民带来的词汇等)、语义类别(如民俗文化词汇、日常生活词汇等)进行检索。该模块利用词汇数据库中特色字段,快速定位相关词汇,提高检索效率。

(4) 语法分析模块:利用依存句法分析等技术,针对新疆汉语方言特殊语法结构进行解析。如对独特句式结构和虚词用法进行分析,将分析结果传输至数据处理模块,同时与语法知识库交互,为用户提供语法结构可视化展示与详细分析结果。

(5) 数据处理模块:负责整合各模块处理后的数据,进行统一存储、管理与进一步分析。该模块依托微服务架构,实现各功能模块的独立部署与动态调整,提高数据处理效率与平台扩展性。

(6) 微服务架构:为应用层各功能模块提

供架构支持,实现模块间解耦,方便功能的独立升级、扩展与维护。例如,当需要优化语音识别功能时,可单独对语音识别模块进行更新,不影响其他模块运行。

4.1.3 展示层

展示层以用户体验为导向,提供友好的用户界面和交互功能。界面设计简洁直观,支持多语言显示,能够在多平台(移动端、PC端等)操作使用,以便不同背景的用户能够轻松访问和使用平台。展示层提供个性化的搜索和推荐功能^[9],用户可以快速找到所需的方言资料。应用边缘计算(Edge Computing),通过文字自动识别系统,输入关键字符及相关内容即可获取信息,也可通过语音检索,来实现搜索需求,在架构设计中,特别关注平台的可扩展性和灵活性。随着方言数据的不断增加和用户需求的变化,平台应具备快速响应和适应的能力。通过采用云计算和虚拟化技术,平台能够在需要时动态扩展资源,确保系统的高效运行。

(1) 多语言界面:平台界面支持多种语言显示,包括简体中文、维吾尔语、哈萨克语和英语等四种语言切换。还考虑到了不同语言的排版差异和字体需求,方便不同地区、不同背景用户使用平台,尤其考虑到新疆地区多民族、多语言的特点,平台可以根据用户的语言偏好提供最佳体验,同时保持核心功能的统一性和稳定性。

(2) 个性化搜索:用户可根据自身需求输入关键词、方言特征等进行搜索,如输入新疆汉语方言中特色词汇、特定语音描述等,精准获取相关方言资料。

(3) 推荐功能:依据用户搜索历史、浏览

记录等，为用户推荐可能感兴趣的新疆汉语方言内容，如具有相似语音特征的方言片段、相关词汇文化解读等。

(4) 语音检索入口：用户通过语音输入方式，实现方言内容检索。平台针对新疆汉语方言语音特点优化语音识别引擎，提高检索准确性。例如，用户说出新疆汉语方言中特色的词汇或短语，即可快速获取相关资料。

4.2 特殊架构说明

4.2.1 语音地域特征的特殊架构

新疆汉语方言语音具有独特的声母、韵母及声调特征，部分地区平翘舌不分，声调调值与普通话不同。在数据层，音频存储格式选用能精准捕捉语音细微特征的编码方式，如基于 Opus 的定制编码，以保留方言独特的发音韵味。例如，在采集乌鲁木齐地区方言语音时，能清晰记录其将“zh、ch、sh”读为“z、c、s”的发音特点。应用层引入深度学习模型，针对新疆汉语方言语音特点进行预训练，优化语音识别模块。例如，利用 TensorFlow 构建声学模型，通过大量新疆汉语方言语音样本训练，提升对独特发音的识别准确率，识别出伊犁地区方言中一些特殊的声调发音。

4.2.2 词汇地域特征的特殊架构

词汇方面，新疆汉语方言包含大量具有地域文化特色的词汇，如源自少数民族语言的借词。数据层建立专门的词汇数据库表结构，增加词汇文化背景、来源等字段，方便对特色词汇进行管理与研究。例如，对于“巴扎”这个源自维吾尔语的词汇，可以在数据库中记录其文化背景为维吾尔族传统的集市，来源为维吾

尔语。应用层开发智能词汇检索算法，支持词汇来源、语义类别等多维度检索，如可通过“少数民族语言借词”标签快速筛选相关词汇，方便用户查找如“馕”“热瓦普”等词汇。

4.2.3 语法地域特征的特殊架构

语法上，新疆汉语方言存在一些独特的句式结构和虚词用法。数据层构建语法知识库，以语义网络形式存储语法规则和例句，方便调用与分析。例如，将新疆汉语方言中“把”字句的特殊用法及相关例句存储在语法知识库中。应用层设计语法分析模块，利用依存句法分析等技术，对方言特殊语法结构进行解析，为用户提供语法结构可视化展示与分析结果。比如，当用户输入一句具有新疆汉语方言特色语法的句子，如“我把那本书看了两遍呢”，语法分析模块能展示出该句中“把”字结构的特点及与普通话“把”字句的差异。

4.3 数据采集与处理

4.3.1 数据的采集

数据采集与处理是新疆汉语方言数字化平台建设中的核心环节，直接关系到平台的准确性和实用性。为实现全面、精确的方言数据收集，以支持后续的分析与应用，需制定科学合理的方案。

数据采集首先需要确定方言样本的范围与代表性。新疆地域辽阔，方言类型多样化，因此在采集过程中，需要涵盖各主要方言区，以确保数据的全面性。具体操作中，可以通过随机抽样与分层抽样相结合的方式选择调查对象^[10]。例如，在乌鲁木齐、哈密、昌吉、伊犁等地区，分别选取若干个具有代表性的方

言点进行深入调研。采集方法上,可以采用面对面访谈、录音收集、问卷调查等多种手段相结合的方式。面对面访谈能够获取丰富的语音语料,而录音收集则保证了语音数据的真实性和完整性。问卷调查有助于收集广泛的语言使用情况及其社会背景信息。对于语音数据的采集,可以使用高质量的录音设备,以保证录音的清晰度和准确性。在数据处理阶段,需要对采集到的语音和文本数据进行系统化整理和分析。语音数据的处理包括转录、标注和编码等步骤。转录工作需要专业人员根据录音内容逐字记录,确保文本与录音高度一致。标注则涉及对语音数据的音节、音素等进行详细标记,以便后续的语音分析和机器学习应用。通过人工智能驱动的设计(AI-Driven Design),不断使用AI技术来优化方言数据的标注、识别和分析过程,提高系统的准确性和效率。编码环节中,可以采用国际音标(International Phonetic Alphabet, IPA)等标准化工具,确保数据的通用性和可比性。可以使用Prometheus、Grafana等工具进行实时监控,使用ELK Stack(Elasticsearch、Logstash、Kibana)进行日志收集和分析。

文本数据的处理则包括文本清洗、分类和存储。清洗工作主要是去除冗余信息、纠正错别字,确保数据的准确性^[11]。分类过程中,需要依据方言的语音、词汇、语法等特征进行细致划分,以便后续分析。经过处理的数据需要存储在安全的数据库中,同时保证数据的可访问性和备份安全。通过科学的采集与处理方法,可以为新疆汉语方言的研究提供可靠的数据基础,为数字化平台的建设奠定坚实的基础^[12]。这不仅有助于方言的保护和传承,也为语言学研究提供了新的视角和数据支持。

4.3.2 采集数据的处理

对音频文件进行降噪处理,同时增强方言特征频段。对视频文件进行剪辑,同时提取音频文件,同样进行降噪及放大处理。将结构化数据存储在MySQL中,将非结构化数据存储在MongoDB中。

4.3.3 音/视频文件处理

经过完整的文件数据处理流程,平台不仅完成基础的数据清洗,更深度挖掘新疆汉语方言在语言交融、声学特征、文化传播等方面的独特价值。

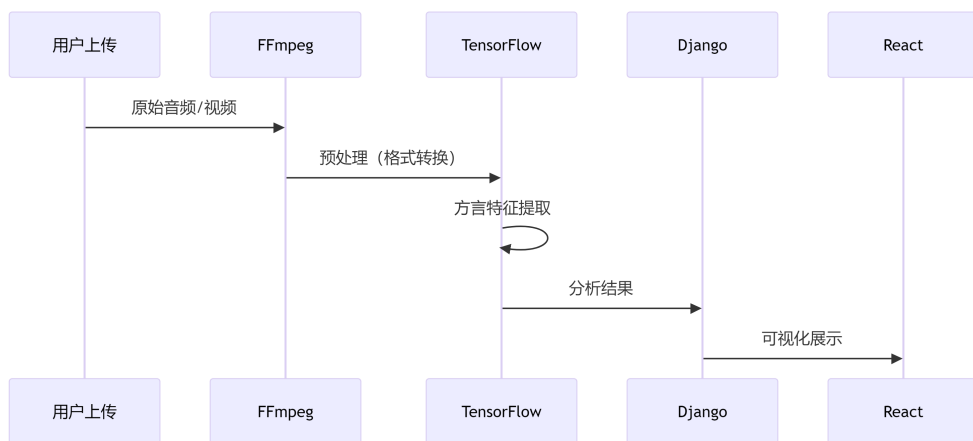


图2 音频文件处理流程

4.3.4 数据脱敏处理

通过剔除敏感数据，实现数据集的深度共享，方便各类用户进行方言学习、学术研究以及实际应用。

综合以上技术实现与工具选择，旨在打造一个高效、稳定、可扩展的新疆汉语方言数字化平台，为用户提供便利的方言数据查询和研究分析功能，推动方言资源的保护和传承。

4.4 技术实现与工具选择

在新疆汉语方言数字化平台的建设过程中，技术实现与工具选择是关键步骤，直接关系到平台的功能性、扩展性和用户体验。

在技术实现方面，首先需要确定开发语言和框架。考虑到平台需要支持多用户访问和大规模数据处理，可以选择 Python 作为主要开发语言，结合 Django 框架进行平台的后端开发，数据存储是平台的核心部分之一，新疆汉语方言数据丰富多样，可以使用 MySQL 数据库进行结构化数据存储，同时，为了支持非结构化数据，如音频和视频文件，可以采用 MongoDB 作为补充的数据库管理系统，提升数据管理的灵活性。在数据采集和处理方面，需要利用现代的语音识别和自然语言处理技术。Google 的 TensorFlow 和 Facebook 的 PyTorch 都是深度学习领域的领先工具，具备强大的语音识别能力，能够有效提取方言特征。相较于基于 C++ 编写的开源语音识别工具包 Kaldi，TensorFlow 和 PyTorch 克服了 Kaldi 在语音识别中不仅需要定制 MPI 调度，而且在训练声学模型、语言模型时需要分批分阶段训练导致误差逐级传递的缺点，如 TensorFlow 可以直接调用已经预训练

完成的声学模型 SpeechCommands，在此基础上对方言特征进行微调，在方言数据的识别中达到轻量化、高效率的效果。结合语音识别技术，还可以使用 FFmpeg 进行音频数据的格式转换和压缩，确保数据在存储和传输过程中的效率和质量。平台的部署与维护同样重要，可以采用 Docker 进行容器化部署，以提高应用的可移植性和环境一致性。Kubernetes 可以进一步实现应用的自动化部署、扩展和管理，确保平台在高并发访问时的稳定性和响应速度。在工具选择上，还需配备完善的开发和测试工具，Gitee 作为版本控制工具，有助于代码的版本管理和协作开发，Jenkins 可以用于持续集成和持续部署，确保代码质量和快速迭代。在测试方面，使用 pytest 进行 Python 代码的单元测试、集成测试等。对于前端代码，使用 Jest 或 Mocha 等测试框架进行单元测试。使用 Selenium 或 Cypress 等工具进行端到端测试。

另外，新疆汉语方言的特殊性主要体现在存在多民族语言交融的现象，如新疆汉语方言当中，受少数民族语言的影响，称洋葱为“皮牙子”，类似的例子还有“坎土曼”（一种铁制农具）等。还有一种比较典型的情况是新疆汉语方言口音中，很多名词词尾都以“子”结束，比如丫头子（女孩）、儿娃子（男孩）等，它们使得新疆汉语方言在语音识别领域面临一项独特挑战，那就是高频借词（如“皮牙子”“儿娃子”）与汉语语法、声调体系的深度混合，构成了典型的语码转换现象。传统单一的语音识别模型在面对此类混合语料时，由于其训练目标与数据分布的局限性，常出现两类系统性错误，一是词汇误判，即把借词误识别为发音

近似的汉语词汇；二是声学特征冲突，即借词的非汉语音系特征干扰句子整体声学模式，导致全局识别性能下降。针对此种现象，平台创新设计了一种双 ASR 协同识别架构对混合语码进行优化处理，具体来说就是在部署 TensorFlow 时，基于 Wav2Vec2.0 的主识别模型负责对输入的音频进行通用端到端识别，通过精准建模兰银官话的连续声调特征、主流词汇及基础语法结构实现方言主干内容的高精度、流畅转写，同时运行的轻量级维语借词检测架构作为并行协处理器，专注于实时检测音频流中是否存在预定义的维语借词发音片段，实现对混合语码中关键借词成分的高灵敏度、低延迟转写，最终通过两个架构的联合解码，系统给出一条融合了标准的汉语方言转写与正确的维语借词原文的识别文本，避免传统 ASR 架构在语码转换时出现识别错误。

4.5 平台创新之处

4.5.1 架构设计创新

对采集的数据进行 MySQL 存储与云端备份相结合的方式保存，并使用离线优先策略，确保核心功能在网络不稳定的环境下可以正常运行。另外支持用户自主安装 Praat 脚本等专业工具，用于声学分析和半自动注释等任务。

4.5.2 技术应用创新

使用 Matplotlib 绘制声调曲线，将标准普通话声调模式与新疆汉语方言声调模式进行专业声学可视化分析。使用 Vosk 实现离线方言识别，即可在无网络的环境下分析方言特征。

4.5.3 功能设置创新

完成基于内容的视频检索，首先提取需查询视频的特征，然后计算在已提取特定声调模式的视频数据库中的相似度，最后按方言特征检索返回最相似的结果。

5 平台应用与效果评估

5.1 用户需求分析

用户需求分析是数字化平台建设中至关重要的一环。首先，了解用户群体的构成是分析的基础，然后通过问卷调查、访谈等方式收集需求，结合行为数据进行用户画像，最终确定新疆汉语方言数字化平台的主要潜在用户群包括语言学研究者、教育工作者、本地居民以及对新疆文化感兴趣的外地人士。每个群体对平台的需求都有其独特的侧重点，如表 1 所示。平台在使用过程中也是动态优化的，通过对使用数据的持续监控和分析，可以了解用户在使用过程中遇到的问题 and 不满，为平台的优化提供重要依据。

表 1 用户需求分析

用户类型	核心需求	关键功能要求
语言学研究者	详尽精准的方言数据，支持语言特征研究和比较分析	1. 多维度数据检索功能 2. 复杂查询条件组合 3. 语音、词汇、语法特征可视化工具
教育工作者	辅助汉语方言课程教学，提高学生参与度	1. 方言音频 / 视频资源库
本地居民	了解保护语言文化遗产，交流方言使用经验	1. 方言学习简易工具 2. 方言社区社交功能
外地人士	了解新疆语言多样性及文化内涵	1. 多语言界面支持 2. 文化背景深度解读 3. 方言文化交融展示

用户需求分析为平台的开发和持续优化提供了清晰的方向，确保平台能够满足不同用户群体的需求，实现其预期的功能和社会影响。

5.2 平台UI设计

UI 不仅是交互界面，更是平衡学术严谨性与大众参与性、融合技术创新与文化传承的战略枢纽，直接决定了平台的生存能力与发展空间。UI 设计不仅是界面美观问题，更直接影响核心功能的使用效率，它必须满足不同群体的差异化需求，比如学者需要专业工具，居民需要易用社交功能等，这些需求都要通过 UI 来实现，UI 设计的好坏决定了数据呈现的清晰度和交互时的流畅度，其设计原则如表 2 所示。

表 2 UI 设计原则

具体内容	说明与应用场景
文化特色	多元视觉语言 1. 新疆特色美景作为界面背景 2. 维吾尔语及哈萨克语界面排版从右向左
场景交互	核心功能展示 1. 数据检索框 2. 声调曲线对比工具（音频、视频同步分析） 3. 语言交融情况（借词提示） 4. 音频、视频资源库
特性优化	离线优先 弱网增强 1. 加载时本地缓存资源 2. 网络恢复后自动同步数据 1. 音频降级传输（保真模式） 2. 视频降级传输（低质量模式）

5.3 平台功能测试与优化

在数字化平台的构建过程中，功能测试与

优化是确保平台可靠性和用户体验的关键环节。通过系统化的功能测试，可以识别和修复潜在问题，从而提升平台的稳定性与综合性能。

5.3.1 基本功能测试

基本功能测试旨在验证平台各核心模块的正确性与运行效率。测试内容主要包括：基础输入输出的准确性验证、数据处理流程的可靠性检验，以及用户界面的响应性能评估。通过引入自动化测试工具，能够高效识别程序中的逻辑缺陷与潜在性能瓶颈。此外，结合用户关于界面友好性与功能易用性的反馈，对交互设计进行了多轮迭代优化，显著改善了用户的整体操作体验。为确保平台在不同环境下的稳定运行，还进行了跨平台兼容性测试。通过在多种操作系统与设备类型上执行测试，发现并解决了可能影响平台广泛适配性的问题，保障了用户在不同终端上获得一致、流畅的使用体验。

5.3.2 平台性能测试

新疆汉语方言数字化平台需支持长期稳定运行，且面临长时间、大规模方言数据采集的任务。从技术层面分析，平台的性能挑战主要集中在三方面：一是对音频、视频等多模态数据进行分析所产生的计算压力；二是在弱网络环境下保障数据传输的稳定性；三是多语言界面动态渲染的效率。系统的性能测试正是为了提前揭示和应对这些挑战。为此，采用了自动化性能测试工具来模拟高并发与复杂业务场景，以全面评估系统承载能力。自动化测试不仅提升了测试效率，也保证了测试过程与结果的可重复性及一致性。性能测试方案如表 3 所示。

表3 性能测试方案

测试维度	核心指标	目标值	测量工具
响应效率	页面加载时间(首屏)	≤ 1.5s (4G 网络)	WebPageTest
并发能力	最大并发用户数	≥ 1000	LoadRunner
	事务处理能力(TPS)	≥ 50(复杂查询)	Gatling
资源利用	CPU 峰值使用率	≤ 75%	Prometheus
弱网适应性	弱网传输成功率(100kbps)	≥ 85%	Charles Proxy
	离线功能完整度	≥ 95%	Lighthouse

测试过程中的用户反馈为性能优化提供了重要方向。例如,针对用户反映的特定操作响应延迟问题,进行了专项的性能剖析与诊断,并据此优化了数据处理算法与服务器响应机制。在功能测试与优化的过程中,通过持续监控与分析关键性能指标数据,定量评估了各项优化措施的实际效果。通过对比优化前后的性能指标,能够量化优化带来的性能提升,并据此调整后续的优化策略。功能测试与优化是一项持续性的系统工程,通过不断地迭代和完善,新疆汉语方言数字化平台能够在不断变化的技术环境中保持其先进性和实用性,从而提供更高效、可靠的服务。

5.4 社会影响

新疆汉语方言数字化平台的建设在社会层面具有重要影响。

首先,该平台的推出有助于保护和传承新疆丰富多样的汉语方言资源。这些方言不仅是语言文化的宝贵遗产,还承载着当地丰富的历史和文化信息。在现代化进程中,由于普通话的普及和城市化的加剧,许多方言面临着消失的风险。数字化平台通过系统化地记录和存储方言数据,提供了一个有效的保护机制,为后人保存了宝贵的语言资源。

其次,平台的建设能够促进学术研究的深入发展。学者们可以通过平台获取全面的方言数据,进行语言学、社会学和人类学等多学科研究。这将有助于揭示方言的演变规律、语言接触现象以及文化交流的历史脉络。此外,平台上丰富的数据资源也可以用于语言教育和培训,帮助语言学习者更好地掌握方言,理解其文化背景。

从社会经济角度来看,数字化平台的建立可以推动文化旅游业的发展。通过对方言的数字化展示,游客可以更直观地了解新疆的文化多样性,体验当地独特的语言风俗。这不仅有助于提升新疆的文化吸引力,也能带动相关产业的经济增长^[14]。

6 结语

新疆汉语方言数字化平台的建设对于保护和传承新疆语言文化具有重要的理论和实践意义,能够为汉语方言研究者提供一个丰富的语料库,可以作为文化教育的资源,帮助年轻一代了解和学习家乡的方言文化,增强文化认同感。同时还可以促进多学科的交叉研究,推动语言学、信息技术、社会学等领域的协同发展,服务于学术界和教育界,增强公众对语言文化保护的意识和参与度。

参考文献

- [1] 长沙晚报掌上长沙. 你口中土不拉几的“方言”，竟是一种巨大的资源！[EB/OL]. (2018-09-18) [2025-01-18]. <https://baijiahao.baidu.com/s?Id=1611954005898045472&wfr=spider&for=pc>.
- [2] 董印其, 郭玮. 新疆汉语方言形成的历史概述[J]. 乌鲁木齐职业大学学报(人文社会科学版), 2006(4): 76-82.
- [3] 塔伊尔江·穆罕默德. 一部填补空白的力作——评《新疆汉语方言与维吾尔语比较研究》[J]. 新疆社科论坛, 2011(6): 92-93.
- [4] 张振华. 湘语与新疆汉语方言对大学生英语朗读中的韵律影响研究[J]. 邵阳学院学报(社会科学版), 2011, 10(1): 69-72.
- [5] 丽娜·海奴拉. 新疆汉语方言词汇研究文献综述[J]. 文教资料, 2019(27): 30-32.
- [6] 董印其, 陈岳. 新疆汉语方言研究30年文献述评[J]. 新疆师范大学学报(哲学社会科学版), 2012, 33(4): 68-73.
- [7] 张洋, 娣丽达·买买提明. 新疆汉语方言的历史、形成、确认及其特点[J]. 新疆社科论坛, 2009(6): 74-78.
- [8] 李先明, 谷国栋. 智慧城市基础数据库平台总体架构设计[J]. 通讯世界, 2018(10): 112-113.
- [9] 宋虎. 数字化转型对商业银行经营绩效的影响研究[D]. 曲阜: 曲阜师范大学, 2024.
- [10] 孙伟伟. 长沙市城市居民健康信息行为调查与对策研究[D]. 长沙: 中南大学, 2013.
- [11] 丁菊薇. 汉语方言资源数据库管理系统[D]. 兰州: 西北民族大学, 2019.
- [12] 成萌. 我国自然博物馆生物标本数字化的方法研究[J]. 山东工业技术, 2015(18): 138.
- [13] 高原, 顾明亮, 孙平, 等. 多用途汉语方言语音数据库的设计[J]. 计算机工程与应用, 2012, 48(5): 118-120.
- [14] 师帅. 新媒体视阈下关中方言文旅产品开发与传播策略研究[J]. 旅游与摄影, 2022(16): 99-101.

(责任编辑: 何彦青)