文献 DOI：
10.11922/csdata.120.2015.0024

文献分类：信息科学

收稿日期：2015-10-23
发表日期：2016-08-08

2015 年中国少数民族地区蒙藏维言语录音数据集

韦向峰[1*]，袁毅[1]，张全[1]，池哲洁[1,2]

摘　要：本文介绍了 2015 年采集的来自内蒙古、青海、西藏和新疆的蒙古语、藏语和维吾尔语言语数据集，使用客户端 / 服务器架构的言语数据远程采集系统，实现了蒙藏维少数民族言语数据的低成本高效采集，公开共享的言语数据集大小为 136 MB 近 800 句。该数据集对于少数民族言语分析研究与教学、语音识别与合成具有重要的研究价值和应用价值，本文的言语数据采集系统仅需少许改动也可以应用于其他语种或者方言的言语数据采集，具有操作简单、部署方便、成本低等特点。

关键词：言语数据；少数民族；蒙藏维；录音；远程采集

数据库（集）基本信息简介

数据库（集）中文名称	2015 年中国少数民族地区蒙藏维言语录音数据集		
数据库（集）英文名称	Mongolian, Tibetan, and Uyghur speech data from Chinese minority regions in 2015		
通讯作者	韦向峰（wxf@mail.ioa.ac.cn）		
数据作者	韦向峰、袁毅、张全、池哲洁		
数据时间范围	2015 年		
地理区域	中国内蒙古、青海、西藏、新疆		
数据格式	.mp3, .txt	数据量	136 MB
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/30		
基金项目	中国科学院信息化专项科技数据资源整合与共享工程重点数据库项目之“民族信息处理学科领域基础科学数据整合与集成应用”		
数据库（集）组成	数据集共包括 3 个数据文件，它们分别为：mp3-example-meng.zip，mp3-example-zang.zip，mp3-example-wei.zip。 其中： 1. mp3-example-meng.zip 是蒙古语言语数据，数据量 70 MB； 2. mp3-example-zang.zip 是藏语言语数据，数据量 26 MB； 3. mp3-example-wei.zip 是维吾尔语言语数据，数据量 40 MB		

引　言

中国是一个多民族国家，少数民族在言语方面有其独特性。收集和保护少数民族的言语数据对于研究少数民族的言语特征、弘扬少数民族文化、促进少数民族的言语信息化建设都有着重要意义。目前，少数民族言语数据的采集方式有以下几种常见的方式。第一种方式是田野调查语言学，如对口耳相传的故事、诗歌等进行现场言语采集或者视频记录；第二种方式是在录音室进行言语采集，在专业的录音环境中使用专业录音设备针对特定的文字内容进行特定人的言语录音；第三种方式是真实环境录音，即在真实自然的环境下记录人们的言语，如会议录音、电话信道录音等。在语音识别、语音合成、言语检测等领域，使用较多的是第二种方式。

飞龙[1]为了语音识别系统的需要，对 200 名持有较标准蒙古语口音的录音人

1. 中国科学院声学研究所，北京 100190；
2. 中国科学院大学，北京 100049
* 通讯作者（Email: wxf@mail.ioa.ac.cn）

员进行了录音，通过 16 KHz 采样和 16 位 PCM 量化后得到了语音识别所使用的语音语料库 69 136 句。山丹 [2] 在专用的语音录音室，以 44.1 KHz 采样速率、16 bit 采样精度、单声道对用于机器测评的蒙古语孤立音标和单词的发音进行了语音录制。陈小莹 [3] 根据藏语拉萨话语音合成系统语音合成语料库的需要，用外置声卡、麦克风、调音台和 Audition 软件在专业的录音室录制了 3000 句藏语言语数据，并对录音人的音质、语速、发音风格的一致性进行了控制。热依曼·吐尔逊等 [4] 在一般录音环境下采用 IBM 笔记本电脑、全双工卡、高灵敏度内置麦克风和高宝立式话筒（阻抗 250 Ω 、灵敏度 -56±3dB、频响范围 100 Hz ～ 16 KHz），以及 WavRecode 录音软件进行录音，录音内容为广播、电视、文艺作品和词典例句中具有代表性的维吾尔语语句（尽可能覆盖维吾尔语种所有的三音子），语句长度为 10 ～ 20 个音节。杨雅婷等 [5] 为建立维吾尔语口语语音语料库，分麦克风信道和电话信道进行采集，以 8 KHz 采样率直接进行采样，麦克风信道用 CoolEdit 录制，电话信道采用声音采集软硬件同时对 2 路电话录音。

为降低在专业的录音室环境中采集少数民族言语数据的成本，本文主要研究蒙古语、藏语、维吾尔语言语数据的远程采集，所依托的项目是中国科学院信息化专项科技数据资源整合与共享工程重点数据库项目之“民族信息处理学科领域基础科学数据整合与集成应用”。该项目由中国科学院合肥物质科学研究院牵头负责，联合在民族语言信息处理和自然语言处理方面具有丰富积累的中国科学院软件研究所和中国科学院声学研究所承担。项目目标是整合中国科学院院内民族信息处理学科领域内相关研究单位的民族语言数据资源（汉 / 蒙 / 藏 / 维），为中国科学院各类信息化工作及科普提供多民族语言信息处理支撑，为我国少数民族地区的信息化建设提供公共数据和服务支持。中国科学院合肥物质科学研究院在蒙古语的汉蒙双语词典、汉蒙句对齐语料以及维吾尔语的汉维双语词典、汉维双语句对齐语料上都积累了丰富的文本语料资源 [6]，中国科学院软件研究所在藏语的汉藏词典、句子级对齐的汉藏双语平行语料上积累了丰富的文本语料资源 [7]。

因此，本文在中国科学院合肥物质科学研究院和中国科学院软件研究所提供的汉蒙、汉藏、汉维平行语料的文本语句的基础上，对来自内蒙古、青海、西藏和新疆地区的少数民族发音人进行远程言语录音，采用客户端 / 服务器的（C/S）架构实现多客户端多用户言语数据的采集。在客户端，用户使用统一的专业麦克风和目前主流的笔记本电脑，运行专门的客户端软件，对照少数民族文字文本语料进行朗读录音。录音完成后，通过客户端软件和 Internet 直接上传到服务器。在服务器端，需要接收各客户端传过来的言语数据，并在 MySQL 数据库中记录言语数据对应的用户 ID、任务 ID、对应文字文本等信息。

1 数据采集和处理方法

整个言语数据远程采集系统主要分为客户端和服务器两大部分，其中服务器的数据管理采用的是 MySQL 数据库，客户端的数据因为数据量较小所以采用 Microsoft SQL Server Compact 数据库进行管理。客户端和服务器之间需要进行数据传输和信息交换，以实现客户端的登录退出功能、文本语料下载功能和上传录音数据功能（图 1）。

本文的研究背景是对蒙、藏、维三种少数民族的言语数据进行采集，言语数据对应的文本由中国科学院合肥智能机械研究所和中国科学院软件研究所提供，

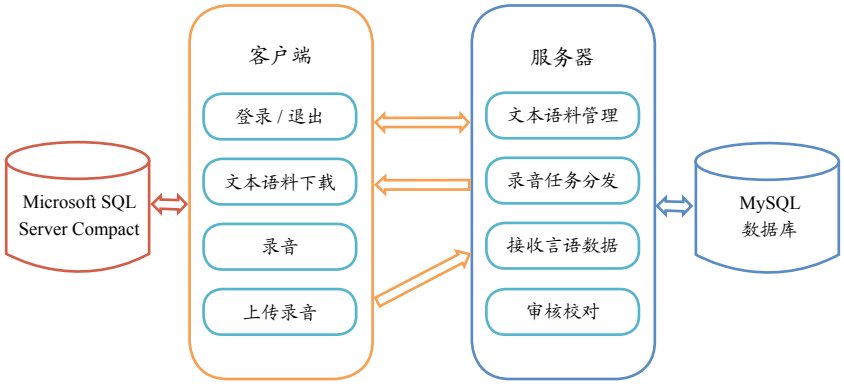

图 1 言语数据远程采集系统

即汉蒙、汉藏、汉维双语平行语料库[6～7]。言语数据采集按照一次任务的制定、分配、录音、上传、审核等环节完成一次成功的言语数据采集，每次任务设计为 100 ～ 1000 句不等，从汉蒙、汉藏、汉维双语平行语料库的文本中随机抽取。一个录音用户可以录制多次任务，一次任务可以分配给多个不同用户进行录制。

数据采集的第一步是在服务器端由研究人员或工作人员制定任务（语种、文本句子数、范围）；第二步，把制定好的任务分配给远程的某个民族语言发音人，并通知该发音人下载任务的文本语料；第三步，该发音人使用本文的言语数据远程采集系统的客户端软件下载任务的文本语料；第四步，任务的文本语料下载成功后，该发音人对照文本语料中的蒙藏维少数民族语言文字朗读语句并进行录音；第五步，任务中所有语句的录音都完成后，该发音人通过客户端软件上传言语录音数据到服务器；第六步，研究人员或语言专家通过开发的 Web 版本的言语数据审核系统对上传到服务器的言语数据进行审核，判定任务的言语数据是否合格；第七步，也是最后一步，把合格的言语数据存入蒙藏维言语数据库中。

使用言语数据远程采集系统的用户从客户端下载新任务的文本语料后，可以对照语料文本进行朗读录音。客户端的录音功能模块是相对独立的，它可以提供：（1）少数民族的文本语料显示；（2）开始录音、停止录音、取消录音、播放录音和重新录音等功能；（3）浏览查看上一句或下一句的录音情况；（4）增大或减小显示的字体，改变字体库等等。图 2 所示的是对某个藏语语句进行录音时的客户端录音功能模块界面。

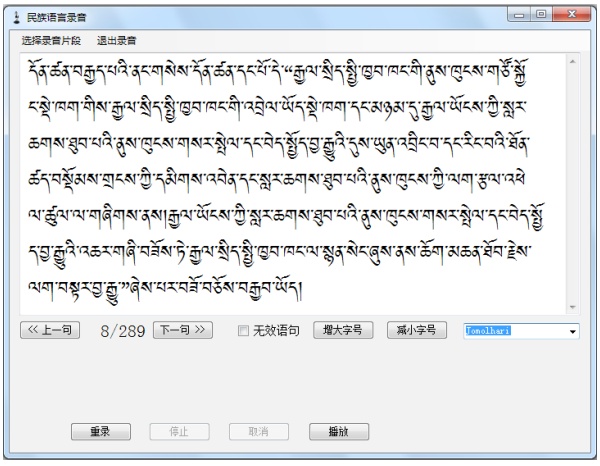

图 2 言语数据采集系统客户端的录音界面

2 数据样本描述

本数据集的一个样本包含三个文件：第一个是 txt 格式的汉语语句的文字文本；第二个是 txt 格式的少数民族语言（蒙古语、藏语或维吾尔语）与汉语语句对应的文字文本；第三个是 mp3 格式的按照少数民族文字文本朗读的言语数据文件。

图 3 ～ 5 所示的是蒙古语言语数据文件及其对应的汉语、蒙古语文字文本。

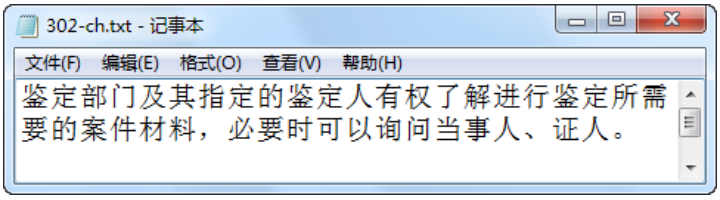

图 3 蒙古语言语数据对应的汉语文字文本

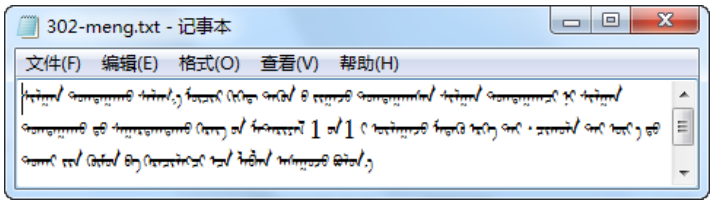

图 4 蒙古语言语数据对应的蒙古语文字文本

图 5 蒙古语言语数据对应的 mp3 文件

图 6 ～ 8 所示的是藏语言语数据文件及其对应的汉语、藏语文字文本。

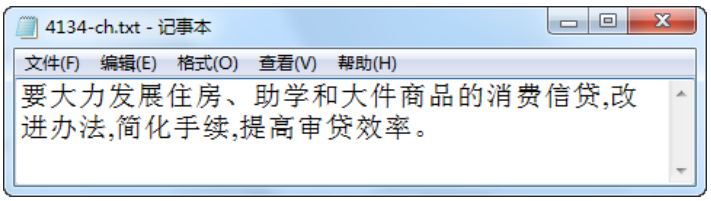

图 6 藏语言语数据对应的汉语文字文本

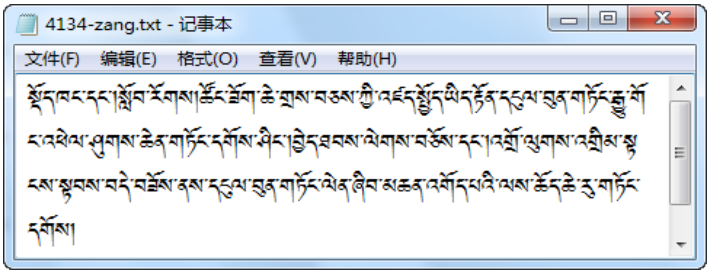

图 7 藏语言语数据对应的藏语文字文本

图 8 藏语言语数据对应的 mp3 文件

图9～11所示的是维吾尔语言语数据文件及其对应的汉语、维吾尔语文字文本。

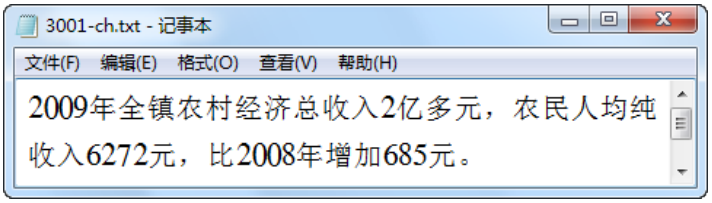

图 9 维吾尔语言语数据对应的汉语文字文本

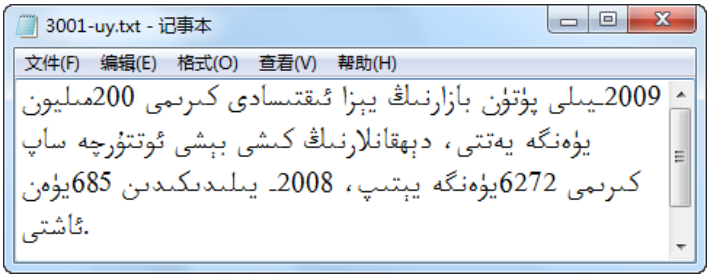

图 10 维吾尔语言语数据对应的维吾尔语文字文本

图 11 维吾尔语言语数据对应的 mp3 文件

3 数据质量控制和评估

为保证客户端上传到服务器的言语数据的质量，在客户端用户上传所录制的言语数据后，审核人员在服务器端还需要对这些数据进行审核。为此专门开发了 Web 版本的民族语言言语审核系统，审核用户可以通过 Internet 实现远程登录并对言语数据进行审核，审核结果分为通过和不通过。对于审核结果为不通过的言语数据，将不能进入最终的言语数据库。

审核人员包括精通民族语言的专家和精通言语分析的专家，通过多人多遍的审核方式来确保言语数据的质量。根据言语数据的自身特点，审核时出现以下情况的言语数据将被判为不合格：（1）有明显的背景噪音；（2）声音不清晰、音量过小或者无声音；（3）有过多的话筒震动音；（4）有长时间的停顿或空白；（5）有大量念错的字词或者与句子文字不对应的录音；（6）以非自然方式说话的录音；（7）其他审核者认为不合格的录音。对于一次任务中部分语句合格而部分语句不合格的任务，在剔除不合格的语句后，将任务中合格的言语数据存入最终的言语数据库。正式采集到的言语数据需要经过至少 2 名专家至少各 1 遍的审核，由于在正式采集言语数据之前对发音人进行了培训、测试和筛选，正式采集到的言语数据按语句统计的审核结果通过率达到了 97.1%。

4 数据价值

本数据集公开共享了近800句蒙古语、藏语和维吾尔语的语句级言语数据资料，在国内没有这样的先例，填补了中国少数民族公开共享的蒙藏维言语数据语料的空白。本数据集可以用于少数民族的语音参数研究、语言教学研究，也可以用于少数民族语音识别和语音合成系统的开发与应用，具有广泛的学术研究价值和较

大的社会应用价值。

本数据集的采集系统采用客户端 / 服务器（C/S）架构，可供多用户远程同时使用，这大大降低了言语数据的采集成本和采集难度，提高了言语数据采集的效率。本文实现的言语数据远程采集系统，不仅可以用于蒙藏维三种少数民族的言语数据采集，也可以不经修改或者仅修改文字编码显示后用于其他民族或者方言的言语数据采集。这对于中国少数民族语言和方言的言语数据采集研究、言语数据分析研究都具有十分重要的意义，有助于推动中国少数民族言语数据的收集与保护，也有助于少数民族言语应用系统方面的开发与应用。

致 谢

中国科学院合肥智能机械研究所的陈雷，中国科学院软件研究所的马龙龙、刘汇丹对本文提供了不少有建设性的建议，在此表示感谢。

作者分工职责

韦向峰（1976—），男，广西宜州人，博士，副研究员，研究方向为自然语言处理、语音识别和语音合成等。主要承担工作：总体技术方案设计与组织实施。

袁毅（1967—），男，北京市人，学士，高级工程师，研究方向为计算机网络、语音识别和语音合成等。主要承担工作：服务器端言语数据处理系统的实现。

张全（1968—），男，陕西西安人，博士，研究员，研究方向为自然语言处理、语音识别和语音合成等。主要承担工作：言语数据审核系统的实现。

池哲洁（1988—），男，福建尤溪人，博士研究生，研究方向为自然语言处理、语音识别和语音合成等。主要承担工作：客户端言语数据采集软件的实现。

参考文献

[1] 飞龙 . 蒙古语语音识别系统的研究与优化 [D]. 呼和浩特 : 内蒙古大学 , 2009.

[2] 山丹 . 面向蒙古语标准音机器测评的语音数据库设计与建设 [J]. 西部内蒙古论坛 , 2010, (3): 58 ～ 62.

[3] 陈小莹 . 藏语拉萨话语音合成语料库的研究与建立 [J]. 科技信息 , 2013, (9): 13 ～ 14.

[4] 热依曼 · 吐尔逊 , 依皮提哈尔 · 买买提 . 维吾尔语语音语料库管理软件的研究与实现 [J]. 新疆大学学报（自然科学版）, 2011, 28(2): 242 ～ 247.

[5] 杨雅婷 , 马博 , 王磊 , 等 . 维吾尔语口语语音语料库的设计与研究 [C]// 第五届全国青年计算语言学研讨会 (YWCL2010): 208 ～ 214.

[6] ZeDe Zhu, Miao Li, Lei Chen, et al. Building Comparable Corpus Based on Bilingual LDA Model[C]// In Processings of the 51st Annual Meeting of the Association for Computational Linguistics(ACL), Sofia, Bulgaria, August 4 ～ 9, 2013: 278 ～ 282.

[7] 刘汇丹 , 诺明花 , 马龙龙 , 等 . Web 藏文文本资源挖掘与利用研究 [J]. 中文信息学报 , 2015, 29(1): 170 ～ 177.

引用数据

(1) 韦向峰 , 袁毅 , 张全 , 池哲洁 . 2015 年中国少数民族地区蒙藏维言语录音数据集 [DB/OL]. Science Data Bank. DOI: 10.11922/sciencedb.120.30.

Mongolian, Tibetan, and Uyghur speech data from Chinese minority regions in 2015

Wei Xiangfeng, Yuan Yi , Zhang Quan, Chi Zhejie

ABSTRACT This paper introduces a Mongolian, Tibetan and Uyghur speech data set in 2015, which was collected using a remote speech acquisition software system based on Client/Server architecture. The system reduced the cost and improved the efficiency of collecting Mongolian, Tibetan and Uyghur speech data. The data set contains nearly 800 sentences, with a total size of 136 MB. The speech data is of great theoretical and practical value for speech analysis and teaching, speech recognition and synthesis concerning the minority languages in China. The system can be applied into acquiring other language/dialect speeches with slight modification, and it is easy to operate and economic to install.

KEYWORDS speech data; Chinese minorities; Mongolian Tibetan and Uyghur; recording; remote collection

引文格式： 韦向峰, 袁毅, 张全, 池哲洁. 2015 年中国少数民族地区蒙藏维言语录音数据集[J/OL]. 中国科学数据 , 2016, 1(2). DOI: 10.11922/csdata.120.2015.0024.