# 云计算技术在中介语口语语料库建设中的应用[①]

林君峰[②]

[摘　要]　不断扩大的语料库建设规模，与仍主要依靠人工的建设方式是一对矛盾。提升语料库建设的计算机技术含量有助于减轻建设人员的负担，提高建设效率和建设质量。口语音频语料的人工转写工作费时费力，应通过应用云计算技术改进相关工作。

[关键词]　中介语；口语语料库；云计算

## The Application of Cloud Computing in the Construction of Inter-language Spoken Corpus

Lin Junfeng

[**Abstract**]　Today's lagre-scale construction of corpus is still mainly rely on hand labor. More application of computer technology in the construction will help to reduce the burden of construction personnel, and improve the efficiency and quality. Transferring audio file to text by artificial is a time-consuming work, and should be improved by the application of cloud computing technology.

[**Key words**]　inter-language; spoken corpus; cloud computing

# 1　引　言

## 1.1　当前建设方式的局限

当前，“汉语中介语语料库建设正在跨入一个繁荣发展的重要时期”，口语语料库的重要作用也受到更多关注。“除了可以考察学习者口语中词汇、语法、语义、语用等方面的实际表现之外，还可以了解学习者实际的汉语语音面貌，可以对其进行声、韵、调等方面的考察与分析。”(张宝林、崔希亮，2015)南京大学、暨南大学等高校都建有汉语中介语口语语料库，预计建成规模最大的“全球汉语中介语语料库”(目标 5 000 万字)和收录数百甚至上千小时时长

①　本文受教育部哲学社会科学研究重大课题攻关项目“全球汉语中介语语料库建设和研究”资助，批准号 12JZD018。

②　作者简介：林君峰，福建师范大学海外教育学院讲师，研究方向为语料库建设与应用、计算机辅助教学。

的口语语料。

然而,在口语语料库建设规模不断扩大的同时,其建设方式仍主要依靠人工。口语音频语料的人工转写工作相比书面语语料费时费力,转写过程中经常需要反复播放,一份音频文件的转写时间可能要数倍于其时长,转写效率不高。这一方面影响语料库建设的进度,另一方面也会挤占用于标注的工作时间,不利于保持标注质量。此外,音频语料通过网络开放共享时,需要占用的服务器运算资源、带宽等也比较大,费用成本会高一些,这也可能会影响建设单位对外开放语料库的意愿。

要改进口语语料库建设工作,需要应用相关计算机技术,以减轻建设人员负担,降低运营费用,提高建设效率和建设质量,并促进语料库的开放共享。

### 1.2 云计算技术的优势

云计算技术是"通过互联网提供各种计算服务和存储服务"的新兴技术,"云服务供应商主要提供数据中心硬件和软件,利用互联网实现存储服务和计算服务。"(朱文武,2011)

对于用户来说,云计算技术将原本由单台服务器承担的数据存储、计算等任务,分散到由云计算系统管理的服务器集群,用户不再需要直接维护服务器硬件,软件的维护也非常简便。云计算环境下,服务器运营成本显著下降,维护工作得到很大简化,同时性能更加稳定,数据安全也更有保障。

在国内互联网上,已有云 Web 服务器、云存储、云语音等多种公有云计算服务可供使用,这些云服务都可以整合起来用于汉语中介语口语语料库的建设。云服务器可用于运行语料库建设及检索系统,云语音可用于语料转写,云存储可用于音频文件存储、加工和检索。

## 2 语音转写测试

口语语料库建设按流程大致可分为音频语料采集整理、音频转写为文本、正确标注及偏误标注、建立数据库、编写检索程序等几个环节,其中转写和标注的人工耗时最长。书面语语料的标注目前还只能"在总体上采用'人标机助'的标注方式"(张宝林,2013),口语语料的标注也类似。但现在汉语语音识别技术已有了很大的发展与进步,口语音频材料的转写如能借助相关技术,可显著提高效率。

目前,已有多家互联网公司推出了开放的云语音服务,并支持多语种及方言。如百度语音可支持汉语(普通话、粤语、四川话)、英语;讯飞开放平台支持汉语(普通话、川豫粤和东北方言)、藏语、维吾尔语、英语。不过,二语学习者口语音频语料的语音面貌、口语水平参差不齐,国内的云语音平台对二语学习者所说的汉语识别效果如何,能否达到实用化的水平,还需进行实际的测试。

本文选取了中高级阶段两位汉语学习者的口语录音文件,拆分出多种样本,使用自编程序连接到百度语音开放平台①,进行了批量的自动语音识别测试。

① http://yuyin.baidu.com

## 2.1 音频材料的准备

将录音转换为百度语音云服务平台所支持的格式，再从中选取一部分录音剪辑另存，生成一批单个录音时长不超过 60 秒的音频文件[①]。

### 2.1.1 选取测试录音

在中级、高级阶段各选取了一位学生的 HSKK 模拟考试录音，一位是蒙古学生（口语一般，参加中级口语模拟考试），一位是越南学生（口语较好，参加高级口语模拟考试）。

### 2.1.2 整理录音样本

使用 Adobe Audition 对录音进行剪辑。首先，删除无效的部分，包括引导音乐开始之前的录音、考生准备部分的空白录音、引导语“好，考试现在结束，谢谢你！”之后的录音等。然后，转换音频格式。整理完成后保存录音文件为 Windows PCM 格式文件（后缀名为 .wav），作为标准录音文件，再将标准录音文件另存一份，音频采样类型调整为百度语音云服务平台所支持的 16 000 采样率、16 位单声道，以用于下一步的剪辑。

### 2.1.3 剪辑测试样本

从已转换格式的 2 份录音文件中各剪辑出 12 份小文件，内容分别为：

（1）对姓名的提问和回答；
（2）对国别的提问和回答；
（3）对考生序号的提问和回答；
（4）引导语，提醒考生接下来要跟读句子；
（5）引导语说明要求、朗读、考生跟读 1 个句子；
（6）引导语朗读、考生跟读 1 个句子（与前 1 句不同）；
（7）引导语朗读、考生跟读各 2 个句子；
（8）引导语朗读、考生跟读各 3 个句子；
（9）引导语朗读、考生跟读各 3 个句子（与前 3 句不同）；
（10）引导语，提醒考生接下来要准备第二部分；
（11）考生就主题说话，2～3 句；
（12）考生就主题说话，50 秒左右；

这样剪辑主要是为了测试不同类别（只有母语者发音、只有二语者发音、母语者与二语者交替发音、短句、长句、单个句子、两个句子、多个句子、成段表达等）（有交叉）音频的语音识别效果。

① 百度语音云服务平台当前只能上传不超过 60 秒的音频文件，超出时长则报错，不能识别。

## 2.2 语音识别结果及统计

### 2.2.1 识别准确度分类

语音识别得到的文本，与人工转录的文本对照，可分成以下几类[①]：

(1) 完全正确，人工重播录音后，可直接确认通过的。

(2) 虽有少量错误但仍成句，在重播录音后可在已识别出的前后文基础上快速订正的。(语音识别错误的部分用着重号标出)

例 1：请在“哔”声后重复这个句子。(人工转录：请在“滴”声后重复这个句子。)

例 2：你别望了带照相机。(人工转录：你别忘了带照相机。)

例 3：亲，他对自己在收入很满意。(人工转录：七、他对自己的收入很满意。)

(3) 部分识别，但不成句，需重播录音订正较多内容的。

例 4：你别问了等将相机。(人工转录：你别忘了带照相机。)

(4) 基本不成句，大部分内容必须重听录音由人工转录的。

例 5：兴趣系指好了两声。(人工转录：兴趣是最好的老师。)

### 2.2.2 识别准确度统计

按照上述标准，对录音样本中二语者发音部分的识别准确度进行了分类，列出每个句子的识别准确度等级。(见下表)

**表 1 语音识别准确度统计**

| 录音样本 | 样本说明 | 蒙古学生（HSKK 中级） | 越南学生（HSKK 高级） |
|---|---|---|---|
| 1 | 交替发音(共 2 个短句，二语者 1 句) | B | B |
| 2 | 交替发音(共 2 个短句，二语者 1 句) | A | B |
| 3 | 交替发音(共 2 个短句，二语者 1 句) | D | A |
| 4 | 只有母语者发音(3 个句子) | — | — |
| 5 | 交替发音(共 2 个句子，二语者 1 句) | C | A |
| 6 | 交替发音(共 3 个句子，二语者 1 句) | C | A |
| 7 | 交替发音(共 4 个句子，二语者 2 句) | AD | BA |
| 8 | 交替发音(共 4 个句子，二语者 2 句) | DAD | ABA |
| 9 | 交替发音(共 6 个句子，二语者 3 句) | CDC | ABA |

① 这里忽略标点的差异，因为语音识别得到的文本基本上全用逗号。

（续表）

| 录音样本 | 样本说明 | 蒙古学生（HSKK 中级） | 越南学生（HSKK 高级） |
|---|---|---|---|
| 10 | 只有母语者发音(3 个句子) | — | — |
| 11 | 只有二语者发音(2 个句子) | BD | BC |
| 12 | 只有二语者发音(成段表达,50 秒左右,若干个句子,包括长句) | CCDD | BBBBBBBB |
| 统计 | 句数(二语者发音) | 共 19 句 | 共 23 句 |
| | A | 15.8%(3 句) | 34.8%(8 句) |
| | B | 10.5%(2 句) | 60.9%(14 句) |
| | C | 31.6%(6 句) | 4.3%(1 句) |
| | D | 42.1%(8 句) | 0%(0 句) |

## 2.3 可行性分析

从表 1 的准确度等级统计[①]来看，HSKK 高级部分的大部分句子都在 B 级及以上(合计 95.7%)，即大部分句子不需要人工转录(但需要听后确认通过)或可以在已识别文本的基础上快速录入。HSKK 中级部分的音频，因为考生的口语水平一般，所以识别效果不佳，但 C 级及以上的也超过半数(57.9%)。

该测试说明，当口语音频语料的发音人语音基本标准、语法没有大的错误时，自动语音识别可完成大部分的转写工作。即使是口语水平一般的音频，自动语音识别也能减少相当一部分的人工工作量。而且，自动语音识别中的错误对语料建设也并非没有价值。语音识别的错误在与人工听录结果对照时，可以帮助发现一些偏误：

例 6：你敬礼的办公室在对面。（人工转录：李经理的办公室在对面。）

例 7：对不起，这里仅举抽烟。（人工转录：对不起，这里禁止抽烟。）

例 8：如果名片不下雨就好了。（人工转录：如果明天不下雨就好了。）

例 9：可是两个人给一切物件打电话的时候。（可是两个人给汽油站打电话的时候。）

例 6 的识别错误可帮助判断两个语音偏误，一是 l、n 分得不够清(“李”“你”)，二是阴平读得像去声(“经”“敬”)。例 7 的“禁止”和例 8 的“明天”也存在发音的问题。例 9 的识别错误也可能是词汇偏误引起的(汉语说“加油站”不说“汽油站”)。

总的来说，自动语音识别的准确率已达到较实用的水平，其正确的部分可以直接替代人工，错误的部分对偏误标注也仍然是有价值的。因此，将云语音应用于汉语中介语口语语料转写是可行的。

① 全部样本识别结果和人工听录文本下载：http://www.hanyu123.cn/html/yuliaoku/。

# 3 转写方案设计

## 3.1 总体目标

通过云语音批量自动识别音频语料，辅助口语语料转写工作，显著提高工作效率，减轻人工工作量并对偏误标注有所帮助。

## 3.2 开发及运行环境配置

基础文件格式：以 XML 为基础文件格式，用于保存音频文件信息、语料转写文本等。

“XML……是 Web 服务领域的‘世界语’”，“通用且容易解析的 XML 将会成为主流的数据交换格式。”(单东林、张晓菲、魏然，2012：183)使用 XML 有利于程序和数据维护，便于以后对语料的检索和数据挖掘。

开发环境：编程工具为微软 Visual Studio 2013 集成开发环境(编程语言为 C#4.5)，数据库使用能很好地支持 XML 的 SQL Server 2008。

运行环境：一台云 Web 服务器，工作时接入到云存储和云语音开放平台。云存储使用七牛云存储[①]，云语音开放平台为百度语音。七牛云存储可免费存储 10GB 的文件，每月提供 10GB 免费流量，百度语音则完全免费。

预加工好的音频文件存放在云存储平台，相关文件信息存放在 Web 服务器的数据库中。自动转写时，Web 服务器同时连接到云存储和云语音开放平台，从数据库中批量读取文件信息，根据文件信息，将云存储中的音频文件提交到云语音平台进行识别，并将返回的识别结果保存到数据库中。

这样的配置，可以使自动转写过程中的数据传输都在云服务器间进行，运行更加稳定。

## 3.3 转写流程

### 3.3.1 音频预加工

使用 Adobe Audition 音频编辑软件对原始音频内容做无损的加工，仅删除较长时间的空白或纯噪音的部分，不改变音频格式、采样率。(为便于叙说，将加工后的音频文件称为文件 A)

将文件 A 转换为语音云平台指定的音频格式，并拆分出若干个小文件(文件系列 B)。拆分出的单个文件，其时长不能超过云语音平台的限制(百度语音以 60 秒为限)。拆分时记录相对原文件(文件 A)的时间起止。拆分并不影响文件 A 本身。

### 3.3.2 上传音频文件

将文件系列 B 通过 Web 服务器上传至云存储，同时将相关文件信息保存在 Web 服务器的数据库中。

① http://www.qiniu.com/.

3.3.3 批量语音识别

在 Web 服务器选定要识别的文件范围(文件系列 B),并从数据库中批量读取文件信息。根据文件信息,调用云存储,将存储在其中的音频文件(文件系列 B)提交到云语音平台进行批量识别,并将返回的识别结果文本以 XML 格式保存到数据库中。

3.3.4 人工核对

人工听录音(文件系列 B)并校对识别结果文本,完成语音转写为文本的工作,并可顺带进行部分偏误标注。

3.3.5 合成语篇文本

将文件系列 B 的转写文本,根据在音频预加工步骤时记录的、相对于完整录音(文件 A)的时间起止,合成为完整的语篇文本。

此转写方案不仅可以提高转写效率,而且在完成转写的同时,也自然形成了一个在线口语语料数据库,稍加扩展即可实现在线检索,可以更快投入应用。另外,将音频保存在云存储平台上,可减少语料库服务器的运算负荷,从而大大降低带宽负担,减少运营费用,对促进语料库的开放共享也有一定的作用。

# 4 小 结

不断扩大的语料库建设规模,与仍主要依靠人工的建设方式是一对矛盾。提升语料库建设的计算机技术含量有助于减轻建设人员负担,提高建设效率和建设质量。

经本文小规模测试,在口语语料库建设中应用云计算技术可以起到较好的效果,能节省可观的转写时间,对促进口语语料库建设有积极的作用。要进一步提高此转写方案的实用化水平,还需要继续扩大测试规模,并相应完善程序。

## 参考文献

[1] 张宝林,崔希亮. 谈汉语中介语语料库的建设标准[J]. 语言文字应用,2015(2).
[2] 朱文武. 多媒体云计算[J]. 电子产品世界,2011(9).
[3] 张宝林. 关于通用型汉语中介语语料库标注模式的再认识[J]. 世界汉语教学,2013(1).
[4] 单东林,张晓菲,魏然. 锋利的 JQuery(第 2 版)[M]. 北京:人民邮电出版社,2012.
[5] 张宝林. 汉语中介语语料库建设的现状与对策[J]. 语言文字应用,2010(3).
[6] 百度语音识别服务常见问题. http://yuyin.baidu.com/asr/qa.
[7] 百度语音开发文档. http://yuyin.baidu.com/docs/asr.
[8] 七牛云存储音视频/流媒体在线处理. http://www.qiniu.com/feature#data-process.
[9] 讯飞语音云开放平台. http://www.xfyun.cn/index.php/services/voicedictation.