# 863 语音识别语音语料库 RASC863——四大方言普通话语音库

*李爱军，王天庆，殷治纲*

中国社会科学院语言研究所
liaj@cass.org.cn

## 摘要*

语音库的制作和分发应该是一个系统工程，每一步都应该遵从特定的规范，以便数据交换。本文按照数据库制作规范，介绍 863 四大方言普通话语音语料库（包括上海、广州、重庆和厦门普通话）--RASC863 （863 annotated 4 regional accent speech corpus）。RASC863 包括自然口语部分、朗读部分（语音平衡和常用口语句）和方言词汇。口语设立了 160 个话题，由发音人任意选择一个，然后讲述相关内容 4-5 分钟。朗读语料是挑选的语音平衡的句子共 2200 句和 600 个常用口语句。每个方言点的发音人为 200 人，共 800 人。覆盖不同年龄、性别和教育背景。

## 1 引言

在NOKIA中国研究中心的支持下，我们在2001-2002年期间，完成了SPEECON汉语语音数据库的制作(www.specon.org)，并在2000年和2003年与清华计算机系和德意电子科技有限公司参与了JHU汉语口语库和方言普通话库的标注工作。此外，在国家863基金和973项目基金的支持下，完成了863普通话识别语音库设计[1]和电话语音库的标注工作。通过这些工作，我们积累了一定的经验，制定了国际上认可的韵律标注规范C-ToBI[4]和音段标注规范SAMPA-C[5].

纵观国际上著名的语音语料库网站(见附录所列网站)，我们觉得汉语语音库应该遵从一套严格的评价、制作和分发规范。我们将试图提倡和采用这种思想，应用于这次收集和制作的863方言口音库中。

1996年863语音识别数据库以朗读语体为主，考虑了语音的音段平衡。随着语音识别技术的发展，制作口音和口语化的语音库变得重要起来。所以，我们在国家863项目基金支持下，从去年开始准备收集上海、广州、重庆和厦门口音的普通话语音语料库，突出口语化的特点，加大语料覆盖范围。本文将介绍整个语音库的语料设计和制作过程。

## 2 语音库制作过程和一般规范

语音库从准备到制作完成并进行分发，需要经过一系列的工程，如图 1 所示。

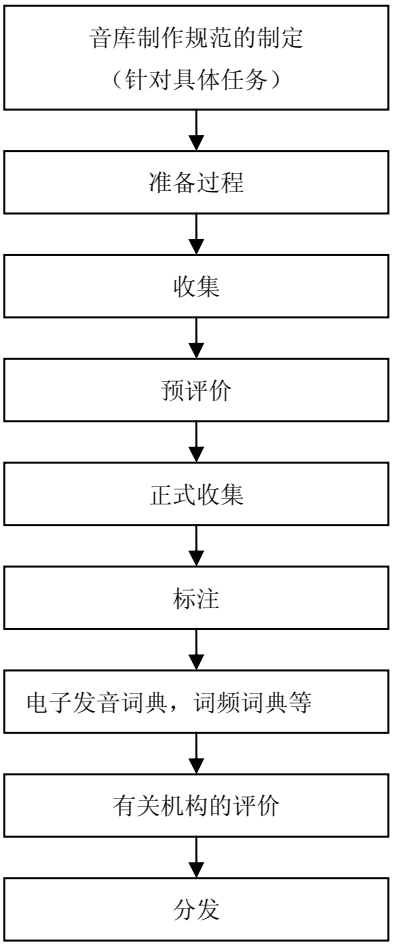


图 1：语音库的制作过程

具体涉及的制作规范和含义如表 1 所示。

* 863 项目子课题编号：2001AA114012，由于一部分内容在 2003 年 10 月的天津第 6 届全国现代语音学学术会议上宣读，这里省略了很多内容。

| 规　范 | 具 体 内 容 说 明 |
|---|---|
| 发音人规范 | 描述发音人的具体要求。如年龄、性别、教育背景和方言背景的要求 |
| 语料设计规范 | 描述语料的组织与设计内容。如发音方式是口语、朗读、回答问题还是情感语音等；语音学和语言学方面的考虑。 |
| 录音规范 | 包括录音设备、录音声学环境等技术指标，录音软件 |
| 数据存储技术规范 | 包括采样率、语音文件存储的格式和描述文件格式等技术规范 |
| 语料库标注规范 | 标注内容和标注系统说明 |
| 法律声明 | 发音人录音之后签署的有关法律条文 |
| 语料库评价规范 | 由分发机构进行应用和使用价值评估 |
| 分发规范 | 分发计划、原则、存储介质（CD/ DVD） |

表 1：　制作语音语料库的一般规范

## 3 863 方言普通话语音语料库制作介绍

本节按照上述规范，简单介绍863方言普通话语音库。

### 3.1 发音人

每个地区发音人200个，年龄、性别以及口音和文化程度分布如下，允许误差5%。

口音按照普通话水平测试标准分级，分为三级，每级又分甲乙两等。首先由录音人判断发音人的普通话级别，最终由专家抽样检查。我们在上海、重庆、广州和厦门，除了选择市区发音人以外，还选择一定比列的郊县发音人，适当扩大口音范围。以收集中等口音为主，三级口音为辅，一级乙等口音占很少一部分。

| 年龄 | 16-30岁 | 31-45岁 | 45-55岁 |
|---|---|---|---|
| | 45% | 45% | 10% |
| 性别 | 男女各一半 | 男女各一半 | 男女各一半 |
| 口音 | 中度二级口音80%，一级乙等5%，三级15% | | |
| 文化程度 | 90% 高中以上学历，10%高中以下学历 | | |

表2：　发音人要求和分布

### 3.2 录音语料

录音语料包括口语和朗读两种体裁。每个发音人的录音语料具体内容见表3。

独白3-5分钟，由发音人从160个话题中任意选择一个适合自己的话题，然后用自然的口语讲述。

15个问题是让发音只回答一些问题，包括工作单位、个人爱好、联系电话、网址、数字等问题。

常用口语句子，我们收集了约500个，每个发音人读20个。

本地方言词汇各地收录的数量不尽相同，涉及日常口语特别是和普通话说法不同的词汇，如厦门地区将“便利”说成“利便”；“不要紧”说成“无要紧”等。这部分是要求发音人用方言说出来。

语音平衡的句子，选自访谈对话、口语对话以及人民日报等语料，句长小于30个音节，尽量覆盖所有的音节间的三音子音联。整个挑选的句子有2200个，原始语料覆盖音节间三音子89%，挑选语料覆盖音节间三音子84%。覆盖所有音节间两音子和几乎所有音节。同时兼顾2-3音节词的声调搭配。详见[2]。

| 录音项 | 发音方式 | 内容说明 |
|---|---|---|
| 0 | Spontaneous speech | 自然独白口语叙述：3-5分钟 |
| 1-16 | Spontaneous speech | 回答15个问题 |
| 16-35 | Read speech | 常用口语句子 20个 |
| 36-50 | Read speech | 本地常用词汇 15个 （方言） |
| 51-165 | Read speech | 语音平衡的句子 110句左右 |

表3：每个发音人的发音语料（ prompt sheet ）

### 3.3 录音设备和录音软件

我们配置两套录音设备，每套包括：笔记本电脑一台，USB 声卡 （M-Audio）。两通道信号：分别采用德国的森海塞尔（Sennheiser）头式话筒和797厂生产的CR722电容传声器（20-20000Hz）。

录音时，记录录音的声学空间面积和背景噪音大小。

4-5分钟的口语语料用Cooledit录制。

语句用我们编制的录音软件录制，同时录制近距话筒（离嘴角2-3cm）和中距离话筒（20-30cm）两个通道语音信号。

### 3.4 数据存储

以16KHz采样16bit精度，Wave格式存储。每个文件至少存贮在不同的两种存储介质上，每位发音人对应一个描述文件，记录发音人的信息：

Speaker ID
Sex
Age
Dialect
Educational background
Recording date

每一个声音样本对应一个描述文件：

Session ID
Speaker ID
Date of Recording
Recording place
Speaking style
***** acoustic and technical description
Recoding sound name
Environmental Conditions
Microphones
Sampling rate
Bits per sample
***** Annotation part
Annotation Convention
***** Annotation should be like this
Orthographic annotation
Prosodic Annotation
Segmental Annotation
***** Or like this
Corresponding annotation file

## 3.5 语料库标注规范

口语部分，包括4-5分钟口语独白和回答问题，全部转写成对应汉字，包括副语言和非语言现象的转写。每个方言点标注20人的语音，包括语音音段和韵律标注。标注系统采用中国社会科学院语言所语音室的C-ToBI3.0[4]和SAMPA-C标注规范[5]。标注软件使用Praat。

朗读发音的标注，包括在时间点上标注正则的发音和实际发音，并且标注由于方言口音引起的“错误发音”。如下图中“招呼小刚”，招呼的’zhao’念成’zao’，用‘#’在音节层和SAMPA-C层分别标出。

方言普通话“音系”是方言和普通话的混合体。我们在制定各个方言的声韵母以及对应的SAMPAC时，主要参考汉语方言词典系列[6]，同时去除老派发音影响。这项工作还没有彻底完成，正在制定过程中。

## 3.6 法律声明

录音之后的法律声明非常重要，以免引起使用和分发过程中的不必要的法律纠纷。只有发音人同意法律声明之后方可录音。每一位发音人录音之后，都必须签署一份法律声明，同意他/她录制的语音的所有内容可以归863项目组使用、复制和分发。我们在录音中也遇到一些发音人，由于不同意这些声明，而放弃录音的。

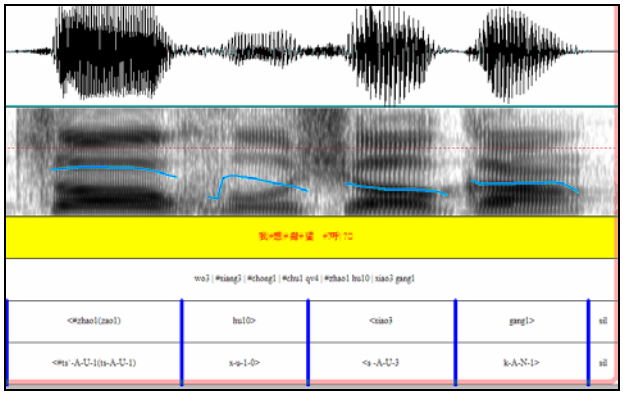

图 2：音段标注例子

## 3.7 语料库评价(validation)和分发(release)规范

在正式录音之前，我们应该进行试录音，然后进行预评测(pre-validation)，目的是发现和解决问题，为大规模实际录音做准备。

录音之后的评价工作是分发单位在数据库分发之前必须进行的工作。“validation”的定义有几种，欧洲语言资源联盟ELRA (the European Language Resources Association)对它的定义是：“the term validation in ELRA is normally used in reference to the activity of checking the suitability for the market, the adherence to standards, and the quality control of the LR product.”

所以，对即将分发的数据库进行评测，有几个方面的目的：

1） 质量控制：由制作者的自我评价和相关机构对数据库的质量进行的评估，最后确定语音库的质量等级分。
2） 可控制性：购买者通过评价可以确定是否购买一个他真正需要的数据库。
3） 可扩展性：通过评价，可以产生更广泛的使用价值。
4） 可比较性：通过一套统一的评价标准，为不同数据库之间对比提供依据，可以通过评分进行对比。但是，需要说明的是，这种对比是有前提的，要和应用结合。

一般评价过程包括，确定评价参考标准、产生错误报告文件，然后进行手工或自动评测，最后产生一个评价报告。

其中，制作者需要按照分发商的要求填写、制作可供分发共享的数据库信息，如数据库内容，大小，存储在介质等，然后真正可以分发之前，分发商和制作者、使用者之间还需要制定和签署一整套的相关法律条款与合同。

## 4 总结和讨论

863 方言普通话语音库RASC863，反映了口语化、语篇化的特点。

在朗读语料的设计上，尽量覆盖所有的音节间的音段音联关系，而没有将三音子进行语音学归并， 这样做主要是考虑到在连续语流中，受到韵律因素的影响，音节音段音联表现的复杂性。同时，由于句子长度的增加，长句子就是一个小语篇，丰富了语句的韵律结构信息。有很多口语句，如问句增加很多。

独白口语部分非常自然，包含一些的情感强烈的语音，为情感语音标注和分析也可以提供一些宝贵的材料。

语音库的制作和分发应该是一个系统工程，每一步都应该遵从特定的规范。我们应该在语料库的制作中尽量使用一些国际上通用的规范，以便汉语语音库在国内外分发和共享。

当然，由于资金和时间的原因，我们大量的标注工作还要留待以后进行，目前只能对十分之一的朗读语音进行精细的语音学标注。我们希望这样的工作会为语音识别的建模提供可靠数据，同时，为方言普通话的声学语音研究提供基础，特别是客观普通话水平测试提供基础数据。[7-11]

想了解更多的信息可以和作者直接联系。

## 参考文献


[1] 祖漪清，1998，连续语音数据库设计的科学性问题，《语音研究报告》，中国社会科学院。

[2] 王天庆，李爱军，“连续汉语语音识别语料库的设计”，第六届全国现代语音学学术会议，天津，2003 年 10 月。

[3] Florian Schiel, Christoph Draxler, “production and Validation of Speech Corpora”,Bastard Verlag Munchen, Erstausgabe, 2003.

[4] Aijun Li, Chinese Prosody and Prosodic Labeling of Spontaneous Speech, Speech Prosody 2002, Aix-en-Provence.

[5] Xiaoxia Chen, Aijun Li, et. al. Application of SAMPA-C in SC, ICSLP2000, Beijing.

[6] 李荣主编，汉语方言词典系列，上海、广州、贵州和厦门方言方言辞典，中国社会科学院语言所。

[7] Aijun Li and Xia Wang, A Contrastive Investigation of Standard Mandarin and Accented Mandarin, Eurospeech2003.

[8] 于珏、李爱军、王霞，“上海普通话与普通话元音系统的声学特征对比研究”，本论文集。

[9] 于珏 、李爱军、王霞，“上海普通话与普通话卷舌元音的声学特征对比研究”，第六届全国现代语音学学术会议，天津，2003 年 10 月。

[10] 陈娟文、李爱军、王霞，“上海普通话与普通话两音节调连对比研究”，第六届全国现代语音学学术会议，天津，2003年10月。

[11] 陈娟文、李爱军、王霞，“上海普通话与普通话两音节词重音对比研究”，本论文集。


**附录：**

国际著名语料库以及相关规范、软件网站：

www.ldc.upenn.edu

www.praat.org

www.bas.uni-muenchen.de/Bas

www.icp.grenet.fr/ELRA/home.html

www.nist.gov/speech

www.phon.ucl.ac.uk/home/sampa/home.htm

www.speechdat.org

www.w3c.org

# RASC-863__AN ANNOTATED SPEECH CORPUS WITH 4 REGIONAL ACCENTS


*Aijun Li, Tiangqing Wang, Zhigang Yin*

Institute of Linguistics, CASS


## ABSTRACT


A large speech corpus with 4 regional accents is introduced in this paper. It includes 800 speakers, half female and half male, from Chongqing, Shanghai, Guangzhou and Xiamen. Close talk microphone and middle distant microphone are used in two channel recording. Data recorded for each speaker includes 5 minutes spontaneous speech, 110 phonetic balance sentences/ discourses, answering 15 questions, 23 frequently used sentences, some dialectically specific words (different for 4 regions). Additionally, some recording specifications and their importance are emphasized in this paper.