第16卷 第6期 声 学 学 报 Vol. 16, No. 6
1991 年 11 月 ACTA ACUSTICA Nov., 1991

快报

# 汉语普通话语音数据库

孙金城 陈希清 李昌立 莫福源 倪 宏 李 彤
(中国科学院声学研究所,北京 100080)

1991 年 5 月 22 日收到

**摘要** 中国科学院声学研究所建立了一个汉语普通话语音数据库，这个语音数据库由声母、韵母、1282 个单音节、几百个双音词和三音词、语音试验句、短文及数字 0—9 等构成. 该语音数据库的发音人有六位(三男三女),他们是广播学院的教师和职业播音员，讲标准的汉语普通话. 语音材料录制在高质量的磁带上，其中有一部分已数字化. 已有许多汉语语音研究部门使用该语音数据库.

# A Chinese speech database

SUN Jincheng CHEN Xiqing LI Changli MO Fuyuan NI Hong and LI Tong
(*Institute of Acoustics Acadimia Sinica*)

Received May 22

**Abstract** A Chinese speech database has been built in The Institute of Acoustics Academia Sinica. The speech database consists of initials, finals, 1282 monosyllables, hundreds of dissyllables and trissyllables, some speech test sentences and short passages which are selected to provide a wide range of phonemic structure, and the digits zero through nine, etc. The number of speakers is 6(3 male and 3 female speakers), who are teachers from the broadcast college and professional announcers, and speak standard Chinese. Speech material are recorded on the high quality tape, and are partially digitaled, so far it has already been used bp many Chinese language reaserch department.

## 一、引 言

语音是传达语意的声音，研究语言要从语音入手. 随着语言信号处理、人机对话(语言合成、语言识别、语言理解)和计算机的发展,建立语音数据库的必要性日益为人们所认识. 目前许多国家都设专项支持本国科学工作者,以建立本国语言的语音数据库,形成了普遍建立语音数据库的趋势[1],如美国、日本、法国等[2-4]. 国内以往从事语言信号研究的单位，或是因为找不到现成适用的语音材料而感到工作棘手；或是搞一些只适合自己研究课题的语音材料. 这种各不相同的语音材料使得彼此的研究成果没有统一的评价依据而不能对比,它既浪费人力、物力和时间，也有碍科学的发展. 尽快建立汉语普通话语音数据库已是科学发展的客观需要和人们的共识,其目的有三:

1. 促进汉语语言信号处理研究工作的深入发展.

当前我国语言信号处理领域的研究、开发正方兴未艾，第五代人工智能计算机又向言语工程学提出了新的挑战．国内许多科研单位、大专院校、部队系统、公司企业等都纷纷投入这一领域的研究与开发，并取得众多可喜的成果．如何科学地评价对比这些研究成果并找出今后改进设计的依据，是一个有待解决的重要研究课题．汉语普通话语音数据库可为各种语言研究、语言信号处理系统提供既科学实用又广泛典型的语音试验材料，同时也为客观地评价对比各种语音研究成果提供科学依据，并促进学术竞争和学科的发展．

2. 促进国际学术交流，提高汉语研究水平．

汉语是世界上最发达的语种，许多国家都有研究汉语的单位和个人．语言信号处理领域的国际学术会议频繁召开，是世界上引人瞩目的国际学术会议之一．当一些语种已建立语音数据库时，尽快建立汉语普通话语音数据库，必将促进语言信号处理领域的国际学术交流，提高汉语语音研究水平．

3. 避免浪费，加快研究进程．

汉语普通话语音数据库会给语音研究工作者带来方便，减少不必要的重复性劳动，避免人力、物力和时间的浪费，人们可以利用现成的语音材料和数据，直接深入各自的研究领域，加快汉语语音研究的进程．

## 二、对汉语普通话语音数据库的要求

1. 选用的语音材料要符合汉语普通话规范，以北京语音为标准语音，以北方话为基础方言，以典范的现代白话文著作为语法规范．

2. 选用的语音材料要既科学又实用，既广泛又典型，能概括所有汉语语音现象．

3. 选用的语音材料要考虑语音平衡，声母、韵母、声调都要有一定的出现频率．

4. 选用的语音材料要兼顾语音基础与应用研究的需要，尤其要满足语音信号处理系统、人机通讯工程(语言合成、语言识别、语言理解、语言压缩编码、语音增强等)的需要．

5. 选取典型的语音样本，设计有代表性的试验句．

6. 语音材料的发音人应受过良好的语音训练，能讲标准的汉语普通话．要有若干发音人，男女各半，年龄不等．

7. 发音人朗读语音材料的速度应与正常讲话速度一样(特殊语音材料除外)，有些语音材料要重复朗读多遍．

8. 使用高质量的录制设备，保证录制的语音材料有高保真度和低噪声．

9. 语音材料应以模拟和数字两种方式生成．模拟方式用开盘与盒式磁带存储，数字方式用高密软盘、硬盘或光盘存储(应有不同的采样率)．

## 三、汉语普通话语音数据库的内容

1. 声母、韵母、声调[5]

2. 轻声

3. 语音音变

4. 汉语全部单音节

5. 数字

6. 语音试验句

7. 字词(双音词、三音词)[6]

8. 单句(汉语三百句)

9. 短文 13 篇(现代语言大师的作品：含政治、经济、哲学、散文、诗歌等内容)

10. 键盘符号(26 个英文字母、10 个数字、72 个符号键)

## 四、汉语普通话语音材料的录制

已完成六人(男声、女声)的全部汉语普通话语音材料的录制与编辑. 其中单音节以两种方式读音：

(1) 读单个音节——全部音节随机排列，每十个音节为一组，每读一个音节后要有停顿，使所读音节不受前、后音节的影响，整个音节读的充分、完整.

(2) 读嵌有音节的引导句——引导句是：我发 x 字. x 是欲读音节，此时 x 处于自然语言环境中. 从引导句中截取的 x 音节，其长短、高低、强弱等特性接进自然语音的音节特性. 全部音节按字母、声调的顺序依次读完.

如：我发 mā 字. 我发 má 字. 我发 mǎ 字. 我发 mà 字.

1. 发音人

发音人受过良好的语音训练，能讲标准的汉语普通话. 他们是北京广播学院的教师和省、市级电视台、电台的播音员，男女各三人，年龄在 27 岁到 50 岁.

2. 录音设备

传声器选用美国 TURNER SE14 动圈传声器，频响特性满足广播与电视技术标准.

放大器选用丹麦 2636 测量放大器

录音机选用波兰 M3401SD 型 HIF 开盘录音机、日本 PIONEER 型 CK-W3 盒式录音机.

3. 录音地点，中国科学院声学研究所语音测听室，测听室内混响时间 0.5s.

4. 录音框图示于图 1.

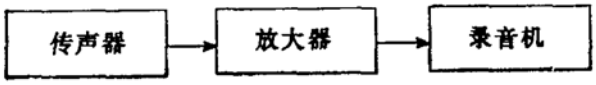



图 1 录音框图

## 五、语音材料的编辑

1. 编辑录制在开盘磁带上的语音材料，剔除错误发音和消除噪声后，转录于盒式磁带上.

2. 盒式磁带上的语音材料按清单顺序录制，依发音人分类，用户可选用 1 人或多人的汉语普通话语音数据.

# 六、语音材料数字化

语音材料数字化后存储在 5 寸低密或高密软盘上，文件按字母顺序排列，用户可从软盘调数据使用。现已完成一个男声(高扬)的 8k 和 10k 采样率的数字语音数据。内容有：

(1) 汉语普通话全部单音节。

(2) 语音试验句。

(3) 短文有“茅以升”、“北风与太阳”。

(4) 键盘符号含 26 个英文字母、10 个数字、72 个符号键。

**1. 采样设备**

放大器选用丹麦 2636 测量放大器

录音机选用波兰 M3401SD 型 HIFI 开盘录音机

滤波器选用中国科学院声学所 AFII 型模拟滤波器 (截止频率 3.4k、4k、5k、6.4k；通带起伏＜±1dB；阻带衰减＞60dB)

A/D 转换器选用中国科学院声学所 TMS320C25 DSP 开发扳(模数转换长度 12 位；采样频率6.3，8，10，16k)

**2. 采样框图示于图 2**

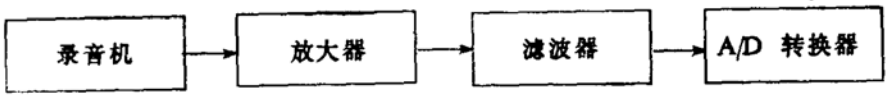



图 2　采样框图

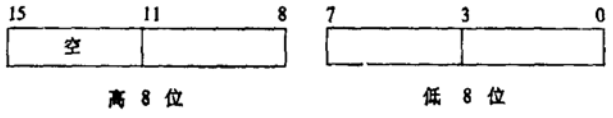
**3. 采样**

10k 采样时，12bit 量化，滤波器的截止频率 4k。

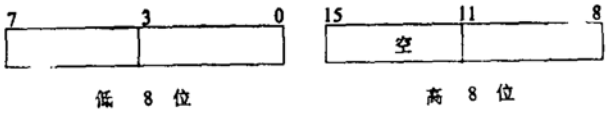
8k 样时，12bit 量化，滤波器的截止频率 3.4k。

**4. 数据格式示于图 3**(第 11 位是符号位)。

**5. 数据调用**




(a) 量化后的数据格式




(b) 数据调入 PC 机后的格式

即：低 8 位的数据放入内存的低 8 位，高 4 位的数据放入内存的高 8 位的低 4 位

图 3　数据格式

(1) 用 BASIC 语言的 BLOAD 语句将数据调入计算机内存

例如：把A驱动器中软盘的 ta1 (塌他它她)音节的数据调入计算机内存.

```
BASICA
DEF SEG = &H  4000
BLOAD "A: ta1. dat", 0
S  YSTEM
```

完成上述操作后，ta1 音节的数据便存放在内存 &H 4000:0 地址处.

(2) 用提供的 CAN. exe 程序将数据调入计算机内存

例如：把A驱动器中软盘的 ta1 (塌他它她)音节的数据调入计算机内存.

```
(执行程序)        CAN
(语音文件名?)     NO. 1 SYLLABLE NAME
(回答)            A: ta1
(文件长度?)       ITS LENGTH
(回答)            6120
```

完成上述操作后，ta1 音节的数据便存放在内存 &H4000:0 地址处. 如果想继续调入数据文件，敲 ENTER 键，重复上述过程后，数据便装入前次数据 (ta1) 之后. 可连续装入 64k 数据，但不能超过 64k. 当不需要继续调入数据时，键入N则屏幕显示:

```
(文件名?)    THE SENTENCE NAME?
```

回答一个自定意的文件名，例如 AAA，此时生成的 AAA. dat 数据文件是将前面调入的文件合并成一个数据文件. 如果不需要合并文件，则键入 Ctrl + C 退出.

**6. 数据显示与编辑**

用提供的 CUR.exe 程序显示与编辑语音数据文件. 例如，截取某个音节的辅音段、逐渐切去某个音节的一部分、将某些音段重新组合等. 如果您使用了中国科学院声学研究所语言与通讯研究室研制的 TMS320 系列 DSP 开发板，还可以回放试听您所编辑的语音的变化情况. 随 TMS320 系列 DSP 开发板提供了调入语音数据、显示与编辑语音数据文件的 CAN. exe、CUR.exe 系列软件，其用法与本文所述相同.

```
(执行程序)                          CUR
(显示几个数据段数据?
每段 64k，最多选 5 段)              SEGLECT SEGG. ADDR. (DEFAULT 5)
(原数据的采样频率?
回答 10 或 8)                       SAMPLING FREQUENCY (kHz)
(显示数据的地址?)                   SEGMENT ADDR. OF DATA DISPLAIED=?
```

一般回答 &H4000，因为 CAN. EXE 由该地址开始装数据.

下面以截取 TA1 音节的辅音T为例，介绍如何编辑语音数据. 本程序定义如下功能键:

| | |
|---|---|
| F-光标线向前步进 | B-光标线往后移动 |
| N-向前进一页(每页 600 个样点) | R-向后退一页 |
| Y-认可选项 | N-取消选项 |

将光标移至辅音T的起始处，要敲Y键认可，然后敲N键前进一页，敲F键移动光标线至

辅音T结束处，再敲Y键认可，至此已完成辅音T的截取．如果您有中国科学院声学研究所研制的 TMS320 系列 DSP 开发板，当屏幕提示：

DO YOU WANT LISTEN THIS PART OF SPEECH (Y/n)

回答 Y，则可回放您刚刚截取的辅音 T．否则就回答 N，进入下一选项：

DO YOU WANT SAVE A FILES (Y/n)

如果要存刚刚截取的辅音 T，就敲 Y 键，屏幕提示：

FILE LENGTH 1200　　（文件长度 1200 字节）

NAME OF FILE TO BE SAVED

回答存入硬盘或软盘的驱动器名(默认为C盘)、自定义的文件名．例如回答：

B:T

此时在B驱动器的软盘上存入了一个取名为T的 T.dat 辅音数据文件．

## 七、今后改进方向

1. 增加语法和语意方面的词汇和句子．
2. 选取有代表性的样本，设计语调试验句．
3. 增加发音人数量．
4. 增加同一语音材料的同一发音人的不同时间，不同地点的发音次数．
5. 考虑不同传输条件和不同声级噪声的语音材料录制．
6. 增加多种采样率的、多个发音人的数字化语音数据．
7. 改用数字录音设备和介质．
8. 增加语音评价材料和测试软件．

## 八、结　　论

几年来，我们经过充分调研，认真听取各方面专家的意见，投入大量人力和物力，从事这一课题的研究与实践，撰写了有关论文[7]，建立了汉语普通话语音数据库．该语音数据库的数据可靠、内容丰富、材料典型、存储合理、使用方便．尽管它还不完善，还有待改进，但是，自语音库建立之日起，就得到了学术界的关心与支持，并在很多研究领域中得到应用．今后，它必将在汉语语音的基础与应用研究；言语通讯工程和系统的设计中得到更广泛的应用．

## 参　考　文　献

[1] 吴宗济，实验语音学概要，高等教育出版社，1989.
[2] Shuichi Itahashi: "A Japannese Language Speech Database", Proc. ICASSP' 86, (7.4)
[3] LtColonel Michael、F. Gayote et al.: "A Speech Data Base at the United States Air Force Academy", Proc. ICASSP' 84, (7.2)
[4] R. Carre, et al.: "The French Language Database: Defining, Planning and Recording a Large Database", Proc. ICASSP' 84, 42. 10
[5] 马大猷，语言信息和语言通信，知识出版社，1987.
[6] 北京语言学院，现代汉语频率词典，北京语言学院出版社，1986.
[7] 陈希清等，标准汉语语音数据库，SICS-3/87.