汉语中介语动态追踪有声数据库建设的基本设想

袁丹　吴勇毅[①]

［摘　要］ 本文介绍了汉语中介语动态追踪有声数据库不同于以往中介语语料库的特点，并讨论了数据库的建设流程，包括字表与词表的设计、信息库的建立、语音数据采集、语音标注、语音分析、建立声学参数数据库等。该数据库的建设有利于完善二语语音习得理论，并为开发留学生语音测试软件用于语音诊断及语音矫正提供研究基础。

［关键词］ 汉语中介语；动态追踪；有声数据库

The Innovation of Chinese Inter-language Dynamic Tracing Audio Database

Yuan Dan　Wu Yongyi

［**Abstract**］ This paper discusses the construction process of the Chinese inter-language dynamic tracing audio database, which with different characteristics from the previous inter-language database, includes word list designing, information database building, sound data collecting, sound labeling, sound analysis and acoustic parameter database building. It is believed that construction of such a database is beneficial to improve the second language phonetic acquisition theory, and to provide the research basis for the development of foreign students' pronunciation test software for speech diagnosis and speech correction.

［**Key words**］ Chinese inter-language; dynamic tracing; audio database

1　引　言

国内对语音数据库建立最为关注的要数民族语学界，鲍怀翘等(1992)就已提出了建立藏语拉萨话语音声学参数数据库的设想，而后蒙古语(呼和等，1997)、安多藏语(于洪志等，2007)、哈萨克语(娜孜古丽·吐斯辅那比，2015)等少数民族语言的声学数据库项目也随之

① 作者简介：袁丹，华东师范大学对外汉语学院讲师，研究方向为语音学、方言学、社会语言学。吴勇毅，华东师范大学对外汉语学院教授、博导、院长，研究方向为语言学及应用语言学、第二语言习得、对外汉语教学理论与教学法、教师发展等。

开展，这些研究除了有效保存了少数民族语言的语言资源外，还为少数民族语言的语音合成、语音识别和语音教学打下了研究基础(廖艳莎等，2010)。2007 年，国家语言文字工作委员会提出了建设中国语言资源有声数据库项目。目前在方言库建设方面，江苏、北京、上海等地已经率先完成，山东、浙江、湖北、福建等地的建库工作也正在开展之中，民族语库的建库工作也在逐步展开。中国语言资源有声数据库建设目的是为了保存中国境内的语言资源，当然也涉及一些与此相关的项目开展，如语音标注软件的开发、语言文化网站的建设和维护等(曹志耘，2015)。王韫佳等(2001)提出了建立汉语中介语语音语料库的基本设想，文中详尽地阐述了数据库建立的五个主要过程：(1) 发音人和发音素材的确定；(2) 录音；(3) 数据库系统和数据库管理系统的建立；(4) 原始资料的登录；(5) 对部分录音的声学分析和声学参数的登录。非常可惜的是，这个中介语语音语料库的材料未能共享，因此一般无法查到这个数据库。另外，近年来，北京语言大学、南京大学、苏州大学、南京师范大学等高校都在建立中介语语料库。这些中介语语料库收集了较大规模的中介语语料，但是对于录音质量却不加以考虑，不能运用于实验语音学的分析与研究。事实上，国内对外国学生汉语语音偏误的教学大多基于经验性的判断，并没有一个大规模的中介语语音数据库作为支撑，因此纠偏大多也是零星的、经验性的，对留学生的汉语语音偏误并未做全面的、科学的考察，也没有设计出相应的留学生语音测试软件。而建立大规模的中介语有声数据库可以为语音教学、语音测试、语音诊断、语音矫正等打下基础，同时也为完善二语习得理论提供了大数据支撑。

2 基本特点

不同于以往的中介语口语语料库，我们要建设的汉语中介语动态追踪有声数据库，旨在建立一个可进行动态追踪和语音实验分析的语音偏误有声数据库。以往的语料库是静态的，只能体现某个留学生在某个阶段的语言面貌，本数据库要求对留学生的语音偏误进行动态追踪，分析其语音习得的过程。以往的中介语口语语料库，对语料采集的声音质量要求并不高，但本数据库要求采集的语料必须是高保真、非压缩的语音信号，以利于后期的语音实验分析。具体来看，有以下几个特点：

2.1 不同母语背景

本数据库建立的目的之一是要获得二语语音习得普遍性原则和差异性特征研究的大数据支撑，因此大规模跨语言的数据采集是本数据库的重点。数据采集必须针对不同母语背景者，希望最大限度地保证每种语言都能建立一个分数据库，这样才能保证提取普遍性规则的有效性，保证母语迁移差异性特征的多样化。不同母语背景的分数据库建设，应采取“从严划分，灵活合并”的原则，如英语有美国英语、英国英语、澳大利亚英语等之别，在建立分数据库时不能将其合并，因为它们在元音音质上并不能完全对应。这样的差别会对学习者的语音产出和感知产生影响。但是在某些语音特征上，它们又几乎一致，如没有送气和不送气塞音的对立，在考察这样的语音特征时可以灵活合并。在保证语言多样性的同时，还需保证每种语言具有足够大的样本量。一般来说，日韩来华留学生较多，样本量能够有所保证，但其他语言样本量的采集就会遇到困难。由于此数据库还有个人语音诊断的作用，因此在样

本量的采集上，我们将采取“多多益善”的原则，即对不同母语背景的样本量的均衡性不做过多的控制，只要有一个学习者，就为他建一个档案库，以利于做动态追踪性的研究。

2.2 动态追踪性

数据库所采集的语音资源并非某个留学生在某个阶段的静态语料，而是从这个留学生学习汉语（或入学）之初，就给他建档，动态采集其在不同学习阶段的语料，以达到动态追踪的目的。实验研究表明，二语习得者在语言学习之初要掌握母语音系中缺失的音位对立会比较困难，但随着学习时间的增加，这种困难会得到改进，甚至完全克服。如 Liu 和 Jongman(2012)对美国学生习得汉语[ts]和[tsh]的研究结果表明：初等一级的美国学生能够掌握汉语[ts]和[tsh]的时长对立（包括爆破时长和擦音段时长），不能掌握重心（center of gravity）的对立；而初等三级的学生则两者都能掌握。因此，对学生的语音习得过程进行动态追踪，可以发现哪些是学习者在习得过程中易于矫正的错误，哪些是学习者容易“化石化”难以改变的错误。另外为每个学习者建立一个语音数据库，可以对每个学习者进行追踪诊断，每隔一段时间即为学生进行测试，开出诊断报告，并提出建议性的矫正方案。

2.3 有声性

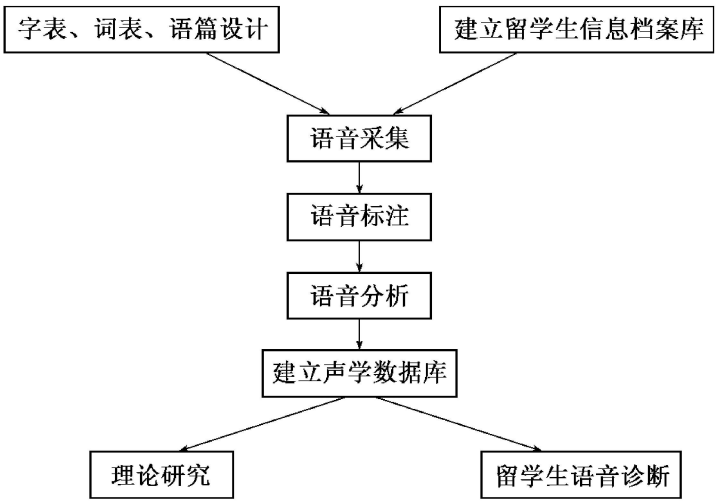
不同于以往的中介语口语语料库，本数据库十分注重录音样本的采集方式。以往的中介语口语语料库虽然也强调了有声性，但是对录音质量的控制却并不十分严格，大多采用录音笔录音。当然这和语料库的建库目的相关。以往的语料库一般用来做词汇、句法、语篇的分析，对录音质量的要求并不高，本数据库致力于对语音材料做实验语音学的分析，在此基础上建立不同母语者汉语语音的声学参数库，因此对录音质量有相当高的要求。录音质量的优劣会直接影响到声学数据的采集，如在录音中经常会遇到的“削波”问题和“喷麦”问题。削波会对元音共振峰的提取造成影响，而喷麦又会对送气声母的测量造成干扰，录音笔等非专业的录音设备往往无法对削波、喷麦等技术问题进行控制，而且录音中已进行过滤波的处理，实际上无法进行准确的声学测量。基于此，本数据库的录音样本全部要求在专业的录音棚中完成，信噪比要求控制在－65 dB 左右，录音声卡使用 Sounddevice Usbpre2，录音话筒使用头戴式指向性话筒 AKG C520。

2.4 开放性和共享性

一个庞大的数据库，需要有庞大的数据样本量作为支撑，而数据库的开放性则是数据库样本量不断扩充的保证。本数据库将和本院留学生的入学测试和阶段性测试挂钩（华东师范大学对外汉语学院的各类长期进修和学历留学生每年在千人以上的规模），尽量做到采样的最大化。张宝林(2015)认为，“国家资助语料库建设的目的就是促进学术发展，推动国家教育事业与科学技术的进步，而实现最充分的资源共享是达此目的的前提”。实现共享性的手段有很多，但最好的方式应是搭建后台数据库，制作专门网页，并提供多维度的检索方式，让所有人都可以在网上检索并下载语音数据，真正实现数据共享（而共享者也可以无条件或有条件为本数据库提供符合要求的语音资料）。

3 数据库主要内容及建库流程

建立一个如此庞大的中介语有声数据库，首先要制定录音的字表、词表以及语篇，并为入库的留学生建立信息档案库。当这些准备工作完成后，才是最重要的录音采集阶段，采集来的录音材料都需要进行标注，然后进行声学分析，建立声学参数库，最后将这些数据用于理论研究以及应用于留学生语音诊断(具体流程参见下图)。以下我们将分步骤详细阐述建库流程。



3.1 字表、词表、语篇的设计

字表理论上要包括普通话声、韵、调配合的所有音节，但需要排除某些音节可能只有生僻字的音节，扣除后再经过挑选，确定普通话 1 000 个音节为数据库的字表。考虑到初级学生在辨识汉字上还有很大的困难，1 000 字音节只列出拼音不列汉字。

词表只设计两字组的词组，但即使是两字组在设计上困难也比较大。目前来看，似乎只能根据经验来判断哪些语音项目需要设计两字组的词组，如上上变调、送气塞音和不送气塞音的对立、塞擦音声母等语音项目。以汉语普通话送气塞音 p[p^h]和不送气塞音 b[p]为例，日语、英语、法语中的辅音为清浊对立，清辅音并没有送气和不送气的对立，但却是条件变体。三种语言送气辅音和不送气辅音出现的条件并不相同：日语清辅音处于词首时通常送气，不在词首时不送气；英语的清塞音在词首时送气，在 s 后不送气；法语的清塞音在词首时不送气，在词尾时送气不送气皆可。那么，日语、英语和法语学习者在习得汉语普通话的送气和不送气声母时在出现位置上是否有习得差异？这就需要大规模的中介语语音数据库来检验。

语篇设计包括朗读部分和讲述部分。朗读部分，分等级设计统一的语篇让学生朗读，初级学生只给出拼音不给出汉字。朗读部分的语篇设计要尽量做到声、韵、调搭配的全面性，可考虑设计多个较短小的语篇让学生朗读。讲述部分可确定几个话题，如：个人爱好、个人经历、家庭情况等让学生自由发挥讲述，时长为 10～15 分钟。朗读部分的语篇可以考查学生在语篇中的语音习得情况，可以进行不同母语背景者语篇中语音习得情况的比较；讲述部

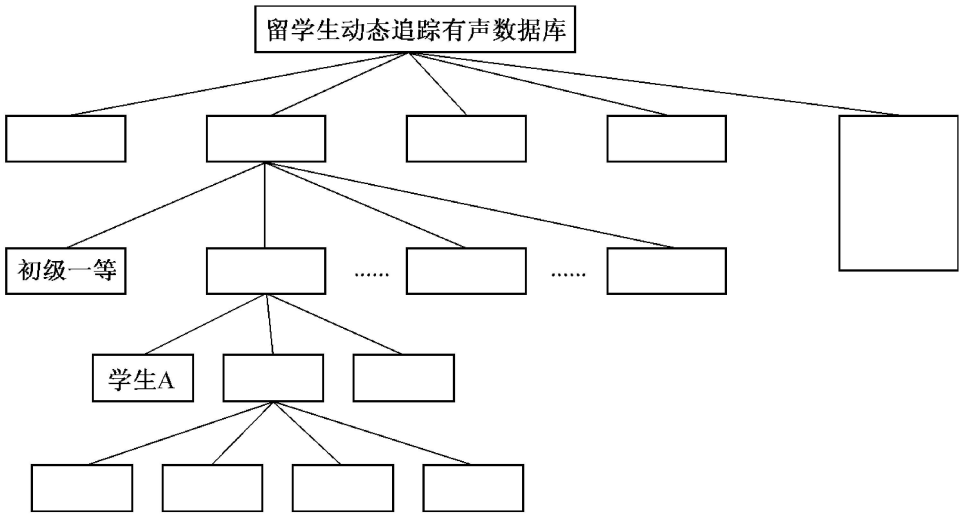
分的语篇可以看出每个学生个体的语音问题。

3.2 建立留学生信息档案库

要对留学生的语音习得情况进行研究或诊断，首先要对每个留学生的情况有大致的了解，所以建立留学生信息档案库就显得十分必要。留学生的信息档案包括个人基本情况、学习经历、工作经历、家庭成员背景等几个大块。要了解学生的母语、掌握其他语言的情况以及家庭成员的语言情况等对语言习得有影响的信息。这部分信息不仅是对学习者做二语语音习得研究时的分类根据，而且也可以用来进行二语语音学习的社会语言学分析。留学生信息档案库的建立，先要按照不同母语背景建立分库，然后再按照不同等级建立次分库，最后再为每一个留学生建立一个信息库。

3.3 语音采集

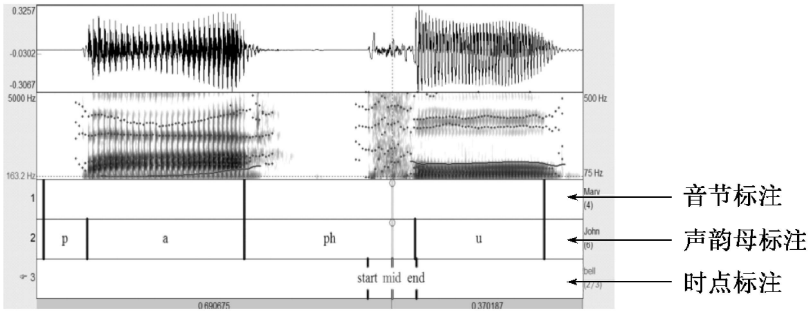
为了达到动态追踪的目的，每学期分三次对留学生进行语音采集工作：入学考试一次、期中一次、期末一次。零起点的学生入学初不做采集，在入学一个月后做一次采集。事实上，一个长期在本院学习的留学生（尤其是学历留学生）从初级到中级甚至到高级，会有很多次的语音采集，当然到中高级以后或许可以减少语音采集的频率。除了对留学生的语音进行采集外，我们还将进行母语者的语音采集，选择汉语普通话标准的 15 男、15 女录音，作为和留学生语音偏误进行比较的参照库。数据库将首先按照不同母语背景建立分库，然后再按照不同等级建立分库，为每个留学生建立一个数据库，数据库里包括字库、词库、语篇库以及个人信息库，以便于日后的检索。具体如下图。



3.4 语音标注

采集到的所有录音都需要进行初次标注和二次标注。初次标注包括两层：第一层是音节标注，也就是切分出音节来；第二层是声韵母标注，也即将所有音节都标注出声母段和韵母段。初次标注的目的是为了便于后期跑数据。如此大规模的语音有声数据库，如果完全用人工来进行测算，耗费的时间是不可想象的。在 praat 中做完初次标注后，就可以运用其他的软件（如 voicesauce）或者在 praat 中编写脚本来运行程序，由计算机自己进行大规模数

据的测算。二次标注是时点标注，也就是标注一些重要的时点，如元音起始点和元音中点，或者擦音段的 0%、25%、50%、75%、100%处。二次标注的目的是为了保持数次测算的一致性，就如哲学家赫拉克利特所说的那样——“人不能两次踏进同一条河流”。在选择某些重要的时点时，我们也不能保证每一次都能选到同一个时点。只有将这些重要的时点标注出来，才能确保不管是谁，不管什么时候测量的都是同一个时点。如下图：第一层为音节标注，切分出了[pa]和[p^ha]两个音段；第二层为声韵母标注，将[pa]切分出[p]和[a]，[p^ha]切分出[p^h]和[a]；第三层我们对[p^h]的送气段做了时点标注，标注出了起点(start)、中点(mid)和终点(end)。



音节的初次标注和二次标注

3.5 语音分析

大规模的录音采样完成后，就可以对一些留学生汉语中介语的语音偏误项目进行声学测量参数的设定，建立留学生二语语音声学参数库。什么样的音素应提取什么样的声学参数，一般来说是有针对性的。元音一般提取第一共振峰(F1)和第二共振峰(F2)这两个声学参数，考察二语学习者的元音声学空间和母语者的元音声学空间的差异；送气声母可以提取送气段声母的时长这个声学参数，如果是塞擦音声母[ts^h]/[ʈʂ^h]/[tɕ^h]/，除了测量送气段时长外，还应测量擦音段不同时点的 COG 的参数；声调提取不同时点的基频值，考察二语学习者四声的读音和母语者的差异。声学数据库的建立是我们开展后续学术研究和留学生二语语音偏误诊断工作的基础。

3.6 语音诊断

建立起不同母语背景者汉语中介语语音声学参数数据库后，就可以对不同母语背景学习者在习得汉语时的语音偏误进行研究和分析，列出偏误要点，开发留学生汉语语音测试软件，为留学生进行语音诊断，并且进行语音矫正。目前来看，还没有一个留学生汉语语音测试软件被应用于留学生汉语语音诊断。虽然国内在普通话水平测试上已经开发了软件，可以进行电脑测评，但是普通话水平测试是针对不同方言区的，并不适用于留学生，测试的效果会大打折扣。也有学者提出可以使用讯飞或百度的语音识别系统来测试留学生的语音，但事实上效果也不尽如人意。一方面语音识别的正确率不是百分之百的，且会随着环境噪音的增加而降低识别率，会把正确的识别为错误的；另一方面，从原理上来看，讯飞和百度的语音识别并非用来判断发音正确与否，而是尽可能地排除个人口音、环境噪音等因素对识别结果的影响，所以对语音失误的宽容度比较高，因此只能作为判断语音的辅助工具，而不能

作为一个语音测试的软件。

3.7 理论研究

留学生汉语中介语有声数据库的建立可以为完善二语语音习得的理论提供大数据支撑。母语迁移是二语习得研究中讨论最多的。Best(1995,2007)的 PAM 理论以及 Flege(1981、1988、1991、1992)的 SLM 理论系统阐述了二语语音习得感知和产出中的母语迁移,建立留学生汉语中介语有声数据库,分不同母语背景对留学生汉语的语音习得偏误进行考察,可以为这两个理论提供汉语的例子,进一步完善这两个理论。虽然学者们普遍承认母语迁移在二语语音习得中的重要性,但是也有一些学者指出,母语者和非母语者相关音位产出的差异并不能完全归因于非母语者的母语迁移,如 Garnica and Herbert(1979)的论文《Some Phonologicac Errors in Second Language Learning》。在这以前已有其他的学者也提出过类似的观点(如 Briere 1966,1968;Tarone 1978;Wode 1977,1978)[①]。Bohn(1995)的研究则进一步指出,不仅在二语的言语产出中母语迁移不能解释所有的问题,在跨语言的言语感知中母语迁移也同样不能说明所有的问题。Bohn(1995)对德国学习者、中国学习者以及西班牙学习者感知英语的[ɛ]—[æ],[i]—[I]进行了实验研究,英语的[ɛ]—[æ],[i]—[I]具有复杂的声学线索,既有元音音质的差异,也有时长的差异。实验前的预测为:德语只有[ɛ],可以和英语的[ɛ]、[æ]形成对立,且有时长对立,预测德语母语者会以元音音质差异和时长差异来区分英语的[ɛ]和[æ];西班牙语有[i],而英语有[i]、[I],没有时长差异,预测西班牙学习者仅以音质差异来进行区分,忽略时长的作用;汉语有[i],而英语有[i]、[I],没有时长对立,但是时长是四个声调的伴随特征,预测母语为汉语的学习者和西班牙语学习者一样,以音质差异来区分[i]和[I],时长的作用较小。实验结果表明:英语为母语者主要依据元音音质的差异来区分[ɛ]和[æ],时长的作用很小;德语为母语的学习者主要依靠时长而不是元音音质来区分[ɛ]和[æ];汉语为母语的学习者几乎完全依靠时长来区分[i]和[I],音质差异的作用很小;西班牙语为母语的学习者也主要靠时长来区分[i]和[I]。由此,Bohn 提出了“去敏化”假说,即当元音音质的差异不能够满足听话人区分元音对立的要求时,时长差异便会用来区分非母语的元音对立,不同母语的二语学习者会表现出共性。因此,通过对不同母语背景者汉语中介语语音数据的大规模采样,除了考察不同母语背景者汉语语音习得的差异性外,也可以考察产出和感知中的共性问题。

4 结 语

本文初步提出了建立汉语中介语动态追踪有声数据库建设的设想,归纳了本数据库不同于其他中介语语料库的特点,并系统阐述了建库流程。我们认为,汉语中介语动态追踪有声数据库的建立具有深刻的理论意义和实用价值,能为留学生汉语语音教学、语音测试、语音诊断以及语音矫正等提供大数据支撑。我们认为,通过将留学生的语音采集纳入到留学生入学测试、期中以及期末测试中这样的方式,大规模的数据采样完全可以做到持续而有序的进行。

① 转引自 Bohn(1995)。

参考文献

[1] 曹志耘.中国语言资源保护的理论与实践[C].IACL-23 会议论文,2015.

[2] 鲍怀翘,徐昂,陈嘉猷.藏语拉萨话语音声学参数数据库[J].民族语文,1992(5).

[3] 呼和,鲍怀翘,确精扎布.关于蒙古语语音声学参数数据库[J].内蒙古大学学报(人文社会科学版),1997(5).

[4] 娜孜古丽·吐斯辅那比.哈萨克语语音声学参数数据库研制方法[J].民族翻译,2015(2).

[5] 廖艳莎,安亚彬,杨阳蕊,何向真.藏语单音节声学参数数据库结构设计[J].陇东学院学报,2010(4).

[6] 王韫佳.建立汉语中介语语音语料库的基本设想[J].世界汉语教学,2001(1).

[7] 王韫佳,上官雪娜.日本学习者对汉语普通话不送气/送气辅音的加工[J].世界汉语教学,2004(3).

[8] 于洪志,李永宏,索南楞次,仁青多杰,李毛吉.安多藏语单音节声学参数数据库研究探讨[C].民族语言文字信息技术研究——第十一届全国语言文字信息学术研讨会论文集,2007.

[9] 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).

[10] Liu. J. and Jongman. A. American Chinese learner's acquisition of L2 Chinese affricates /ts/ and /tsh/ [C]. Proceedings of 164th Meeting of the Acoustical Society of America, 2012.

[11] Best. C. T. A direct realist view of cross-language speech perception [J]. In Winifred Strangeed. *Speech perception and linguistics experience: issues in cross-language research*, 1995: 171-206.

[12] Best. C. T. Nonnative and second-language speech perception: commonalities and complementarities. *Language experience in second language speech learning: in honor of James Emil Flege*. Amsterdam: Benjamins Publishing Company, 2007:13-34.

[13] Bohn. O-S. Cross-language speech perception in adults: First language transfer doesn't tell it all. In Winifred Strangeed. *Speech perception and linguistics experience: issues in cross-language research*, 1995: 273-304.

[14] Flege. Second language speech learning: Theory, findings, and problems. In Winifred Strangeed. *Speech perception and linguistics experience: issues in cross-language research*, 1995: 233-272.