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，并探讨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、[illegible]留学生语音测试软件而为语音诊断及语音纠正提供研究基础。

［**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 [illegible] pronunciation test software for speech diagnosis and speech correction.

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

1　引　言

国内对语言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[illegible]

---

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。

[illegible] 开展，这些研究除了有效保存了少数民族语言的语音资源外，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（廖梅芬等，2010）。2007年，国家语言文字工作委 [illegible] 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 [illegible] 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 [illegible] 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音面貌，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 [illegible] 验分析。具体来看，有以下几个特点：

2.1 大[illegible]

[illegible] 据支持，因此大规模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。[illegible] 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 [illegible]

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 [illegible] 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 [illegible] gravity）的对立，而初等三级的学生前两者都能掌握。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纠正的错误，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 [illegible]

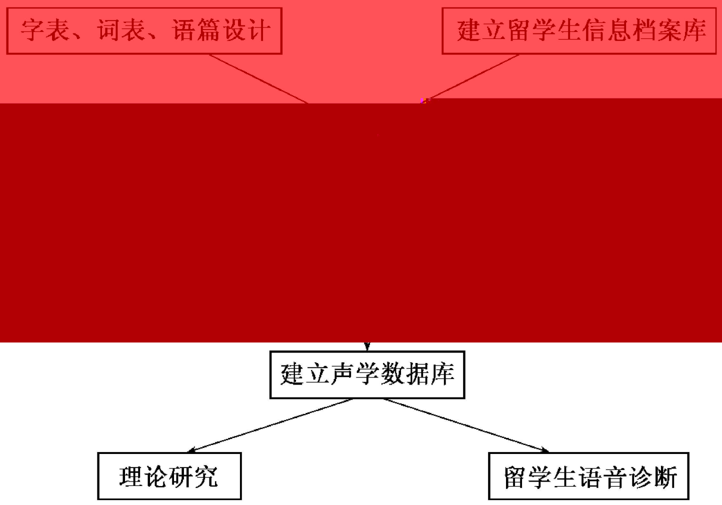
不同于以往的中介语口语语料库，本数据库十分注重录音样本的采集方式。以往的中介语口语语料库虽然也强调了有声性，但是对录音质量的控制却并不十分严格，大多采用录音笔录音。当然这和语料库的建库目的相关。以往的语料库一般用来做词汇、句法、语篇的 [illegible] 使用头戴式指向性话筒 AKG C520。

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 [illegible]

[illegible]

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。[illegible]

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、送气音和不送气音[illegible]

[illegible]

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[illegible]

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[illegible]

[illegible]

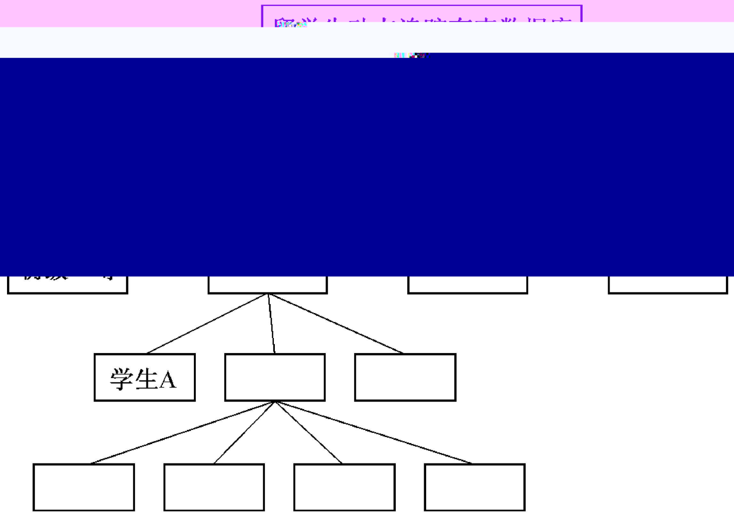
[illegible]

音习得研究时可资参照依据，而且也可以用来进行二语语音学习与社会语言学分析。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[illegible]

[illegible]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[illegible]

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[illegible]

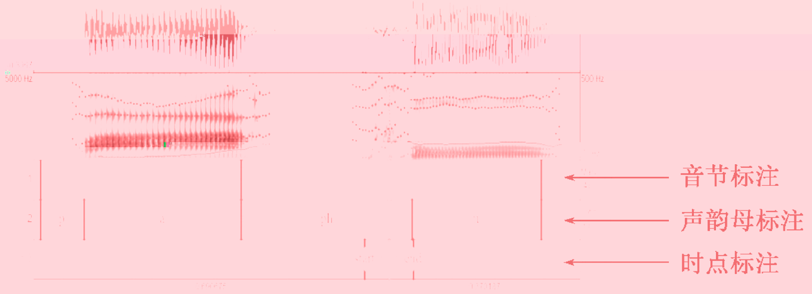
据的测算。二次标注是时点标注，也就是标注一些重要的时点，如元 [illegible]

[illegible]

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 [illegible]

[illegible]

注，切分出了[pa]和[pʰa]两个音段；第二层为声韵母标注，将[pa]切分出[p]和[a]，[pʰa]切分出[pʰ]和[a]；第三层我们对[pʰ]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 [illegible]

[illegible]

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，并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。 [illegible]

[illegible]

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。

[illegible]

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 [illegible] (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 [illegible] 《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 [illegible] 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 [illegible] 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 [illegible] 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(3).

[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).

[redacted]

[12] Best, C. T. Nonnative and second-language speech perception: commonalities and complementarities

[redacted]