基于古音系统的汉语方言语音合成研究*

黄小明[1]，熊子瑜[2]
(1. 中国社会科学院研究生院语言学系，北京 102488；2. 中国社会科学院语言研究所，北京 100732)

文　摘：采用 HTS 语音训练合成工具和 STRAIGHT 语音合成器，尝试在未知某方言（或土语）的实际语音系统的条件下开展相应的语音训练合成工作。采用古音系统来转写汉语字音，并在此基础上设计相应的问题集以实现语音训练合成。设计了一套面向汉语方言语音合成的通用发音文本；录制了一些汉语方言点的合成语音语料库；搭建了基于古音系统的汉语方言语音合成平台。普通话的合成实验结果表明：基于古音系统训练合成出来的语音，在可懂度和音质上跟基于普通话拼音系统训练合成出来的语音非常接近。这表明基于古音系统进行汉语方言语音合成的方法是有效的、可行的。

关键词：古音系统；方言；语音合成
中图分类号： H113；H17；TN912.33

尽管基于大语料库的波形拼接技术在语音合成领域中仍然处于主导地位，但因其存在音库构建周期长、合成系统可拓展性差等一系列的技术难题，近年来，基于 HTS 工具的语音频谱参数训练合成技术在学界和业界得到了越来越多的关注[1,2]。目前已有一些研究利用 HTS 工具进行汉语普通话和方言以及情感语音的合成[3,4]。

通常而言，详细掌握某方言的语音系统并编写出相应的字音转写程序是开展其语音合成研究的前提。为了准确真实地反映某方言的语音系统，通常依赖于方言学家的田野调查。然而田野调查不仅费时费力，而且还要求调查者系统掌握音韵学和方言学的知识，并接受严格的听音和记音训练。对于普通话和某些汉语方言来说，其语音系统已经得到了较为充分的研究，有坚实的字音转写数据基础。但对于大量的次方言（或土语）来说，要想通过田野调查方法去分析和描写其语音系统、整理同音字表并设计出相应的字音转写程序，则无疑会制约方言语音合成的进程。

传统音韵学中对语音的许多分类方法与现代音系学中的“区别特征”的对立或对比的分类有很多相似之处[5,6]。图 1 给出了“英雄源自人民”这句话中 6 个音节的字音转写结果，第一行是汉字，第二行是按照普通话拼音方案系统转写出来的汉字读音形式，第三行是按照汉语古音系统转写出来的汉字读音形式。例如：“雄”字在拼音系统中的声、韵、调分别是“x、iong 和阳平”，在古音系统中的声、韵、调分别为“云母、东韵和平声”。

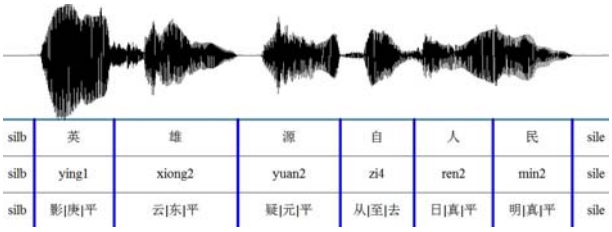

图 1 “英雄源自人民”的拼音转写与古音转写的对比

当前针对普通话的语音训练合成系统无一例外的都采用汉语拼音方案系统来转写汉语字音，而本研究将尝试采用汉语古音系统来转写出汉语字音并进行语音训练合成。该方法的好处是：针对不同的方言或次方言，若能采用一套共同的汉语古音系统来转写其字音，则训练某方言或次方言的语音合成模型时可在一定程度上降低对方言田野调查结果的依赖。其主要难点在于：古音系统和今音系统之间不存在一一对应关系，有时甚至无规律可循。因此用古音系统来转写字音进行训练合成时可能会存在着不能如实反映实际读音的现象，从而影响到合成语音的质量。

尽管古音与今音之间存在着很大差别，但已有大量的方言学和音韵学的研究成果表明，方言语音演变具有很强的规律性和系统性，基于古音系统有可能在一定程度上把握某方言或次方言的实际读音。本研究将通过语音合成实验来探索基于古音系统来开展汉语方言语音合成的可能性。

*收稿日期：2013-04-27
基金项目：中国社会科学院哲学社会科学创新工程项目经费支持。
作者简介：黄小明（1989-），女（汉族），福建省，硕士研究生。
通讯联系人：熊子瑜，副研究员，xiongziyu@163.com.

1 古音系统

《广韵》是中国历史上完整保存至今并广为流传的最重要的韵书之一，能较好的反映中古时期的汉语语音系统。《广韵》的声母系统包含 41 个声母 [6]。古今声母的演变可从发音部分和发音方法两方面来看。从发音方法上看，最突出的一条就是全浊声母的清音化，这使得现代普通话声母系统大大简化。从发音部位来看主要变化有两条：1）普通话从重唇音中分化出轻唇音来，但在方言中有的保留了“古无轻唇音”的痕迹；2）古声母知章庄 3 组合流为现代卷舌音。另外，古声母清浊与今声母送气与否也大有关联[7]。

就韵母而言，《广韵》的韵母系统中有 206 个韵，293 个韵类，142 个韵母[8]。韵类和韵母都要求介音、主要元音和韵尾相同，但韵类相比韵母还区分了声调。根据“韵尾相同，主要元音相近”的原则，206 个韵可归并为 16 个摄。另外，古韵母还区分开合口及一二三四等，这体现的是古韵母的介音系统，其在古代声母和韵母的配合上以及古今语音演变中都起到了重要作用。

现代普通话和方言的字调分类主要是由古四声和古声母的清浊决定的。古音系统中表示声调的平、上、去、入四声，跟现代普通话中的阴平、阳平、上声和去声四个调类之间的对应关系主要有三条：平分阴阳、浊上变去、入派四声。

所以，通过考察古音系统与现代普通话和方言的语音系统，可以知晓其对应关系是多方面的，同时也可以看出古今语音的演变具有较强的规律性和系统性。这也表明利用汉语的古音系统有可能在一定程度上驾驭现代普通话和方言的实际读音，从而实现方言语音训练合成。

2 语音库设计

开展语音训练合成研究，离不开用于训练的语音材料。语音材料的好坏直接关涉到后期训练出来的语音模型的效果[9]。为了能够比较系统全面地反映出汉语古音系统中的各种声韵调范畴，本研究首先以单字为主要切入点，通过一定的算法挑选出比较常用的、能够尽量均衡覆盖所有古声、古韵和古调的 1040 多个汉字，然后再根据这些挑选出来的汉字去进一步挑选词语和句子。

2.1 通用字

在挑选汉字的过程中，本研究主要用到以下材料：《现代汉语方言音库》[10] 764 个常用字和 147 个特字，《方言调查字表》[11] 3768 个汉字和 447 个典型例字，《古今字音对照手册》[12] 6939 个汉字，以及基于大规模文本语料库统计出来的包含 11301 个汉字的“汉语字频表”。

在挑选单字时，本研究首先将《现代汉语方言音库》、《方言调查字表》和《古今字音对照手册》中的汉字古音信息统一整理成“汉字(简体)”、“古音韵地位信息”以及“优先级别”三类信息。“古音韵地位信息”中包含了古声、古韵、古调、摄、呼、等、是否属于多音字等 7 类信息。“优先级别”从高到低分为“调查字表中的典型例字”、“方言音库特字”、“方言音库常用字”、“调查字表中的汉字”和“对照手册中的汉字”五个级别。处理时，首先排序并剔除了重复数据，即重复出现的汉字仅保留级别最高的记录。为保证字表数据的可靠性，本研究还专门聘请了一位方言学博士研究生对字表数据进行了认真细致地审核校对。基于整理后的字表，本研究采用贪婪算法一共挑选出 1040 个常用汉字，能够比较全面地覆盖汉语古音系统中的各种分类信息。

在这 1040 个单字中，单音字 1003 个，多音字只有 37 个。图 2 给出了在这些汉字中所出现的各类古声母的个数，从中不难看出，这些汉字能够较好地覆盖汉语的 41 个古声母。

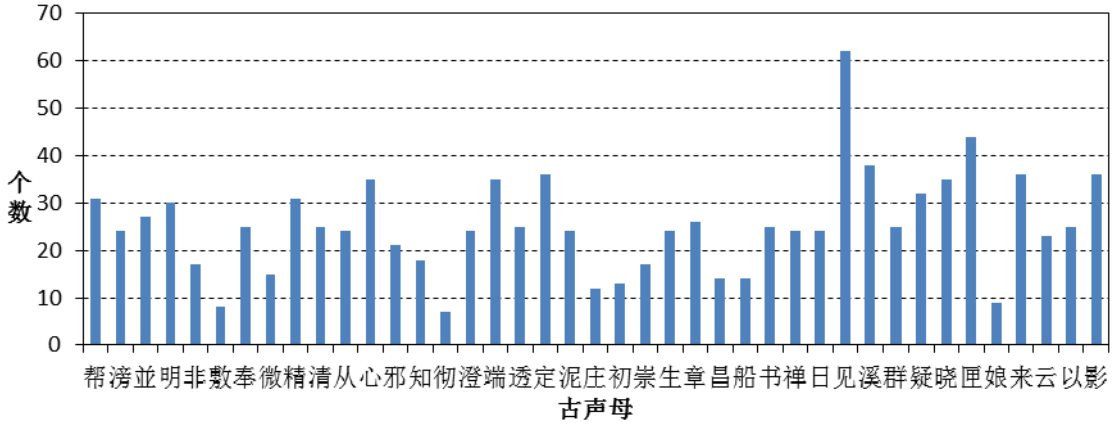

图 2 通用字表的古声母分布情况

古韵母的覆盖情况也非常好，1040 个汉字能全面覆盖 206 个古韵母。图 3 给出了在这些汉字中所出现的“摄”的个数，从中可以看出，这些汉字能够较好地覆盖汉语的 16 个摄。

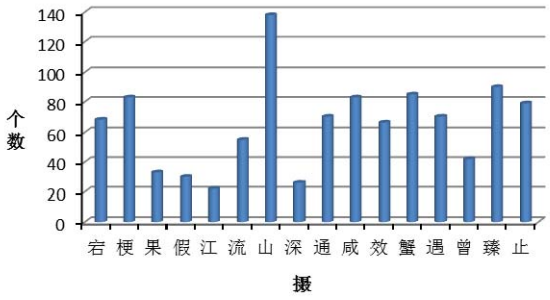

图 3　通用字表古韵母的“摄”分布情况

等和呼的信息可以说明古韵母的介音系统。通用字表中“呼”的分布情况如下：属于开口呼的有 660 个汉字，属于合口呼的有 380 个汉字。表 1 给出了通用字表中“等”的分布情况。由表 1 及“呼”的分布情况可以看出，挑选出的 1040 个汉字能较好地反映出古韵母的介音系统。

表 1　通用字表古韵母的“等”分布情况

等	一	二	三	四
个数	293	171	495	81

在声调方面，表 2 给出了通用字表 1040 个汉字的古声调分布情况。可以看出，平上去入四个古声调的分布情况较为均衡。

表 2　通用字表的古声调分布情况

古调	平	上	去	入
个数	312	237	270	221

2.2　通用词

挑选词语时参考的主要依据是：（1）二字词中的汉字都在通用字表中；（2）尽量均衡覆盖前字韵摄和后字古声的搭配以及前后例字之间的古声调搭配；（3）词频越高越好。本研究采用贪婪算法最终挑选出了 808 个常用二字词，能够比较均衡地覆盖前字韵摄和后字古声的搭配以及前后例字之间的古调搭配。

2.3　通用句

挑选句子的原始语料取自 2006-2011 年的《人民日报》。挑选句子时参考的主要依据是句中出现的汉字必须能够在通用字表中查到，而且语句的长度应该适中。本研究采用了贪婪算法最终挑选出了 999 个句子，总共包含 16484 个汉字，平均长度为 16.5 个汉字，而且能够全面覆盖所有的通用字。

3　语音库制作

3.1　语音库录制

方言语音数据库的所有录制过程都是在专业的录音室中进行的，并使用专门设计的语料库录音软件及 AKG C1000S 话筒，录音质量较好，能够满足语音分析和语音训练合成的需求。录音文本材料一共包含了 1040 个通用字、808 个通用词和 999 个通用句，每个发音人的录音过程耗时两天，大约 13 个小时。

在挑选方言发音人时，尽可能选取那些能够熟练使用方言、年龄在 20-40 岁的方言母语者。普通话的发音人是一名普通话等级为一级甲等的电视台女主持人。录音前让发音人熟悉录音材料。在录音过程中，每半小时休息一次，并要求发音人朗读时语速适中、发音饱满。

目前的方言语音数据库已包含 12 位发音人的语音数据，涵盖普通话及 8 个方言点，分别包括上海、天津、西安、烟台、长沙、成都、南昌和南京。

3.2　标注与转写

根据发音文本，本研究使用 Praat 脚本对每个声音文件统一标注出 7 层信息：第 1 层为实际发音文本，第 2 层为单字，第 3 层为间断指数（初始状态仅包含音节边界、词语边界和语句边界等分类），第 4 层为古音信息，第 5 层为古音信息编码，第 6 层为普通话拼音，第 7 层为普通话声韵母（含声调信息）。

根据自动生成和校准后的 TextGrid 标注文件可以进一步生成训练合成用的标记文件（*.lab，包括 Mono Label 和 Full Label 两类）。按照古音系统转写的标记文件对于所有的方言来说都一致。

4　训练合成

4.1 问题集

模型训练前一项非常重要的工作就是问题集的设计，设计中需要考虑到古音系统中的各种分类关系，如古声母的发音部位和发音方法等。

为了比较，本研究分别设计了古音系统和拼音系统的问题集。古音系统中，古声母共 41 个，根据发音部位分为 4 系 12 组，根据发音方法分为清浊、送气与不送气、塞音、塞擦音、鼻音等；206 个古韵归为 16 个摄，142 个韵母，并根据其呼和等信息归并介音系统。拼音系统中，声母也同样根据发音部位和发音方法归类，韵母根据主要元音、韵尾、介音等归类。

针对基频参数变化特性，本研究设计了针对上下文环境的连调模式。古音系统中，由于古调

四声和古声母清浊都会影响声调的演变，因此可将二者的信息合并处理。拼音系统则是在原来四声的基础上加入了轻声，共五个声调。

另外问题集中还包含分词信息，字在词中的位置信息、词语的长度信息、词在句中的位置信息等。实验结果表明，对于某些不符合语音演变规律的汉字，基于问题集的训练合成方法能在一定程度上依照上下文相关信息解决。

4.2 韵律信息

由于自动转写的韵律间断信息在初始状态下只包含了词语边界（根据分词结果）和语句边界（根据标点符号）等基本信息，难以真实地反映出每个句子的实际韵律层级。为了能够在一定程度上解决这一问题，本研究编写了 Praat 脚本，利用 HTK 工具的音段自动切分结果来捕捉较大的无声停顿，并在此基础上增加短语层级的标注信息。

4.3 合成效果对比

接下来从数据和听感两个方面来对比分析基于拼音系统训练合成的普通话和基于古音系统训练合成的普通话，以考察这两种方法的合成效果有无显著差异。

首先，根据数据统计分析的结果，这两种训练合成方法对字、词、句的时长预测基本一致，特别是在句子时长的预测方面，二者都与原声较为接近，参见表 3。据统计，这两种训练合成方法对音节时长的预测相似度达到 91.65%。

表 3　两种合成音的字、词、句平均时长

	平均时长/s		
	字	词	句
普通话原声	3.080	2.826	6.815
古音系统合成音	2.829	3.166	6.801
拼音系统合成音	2.833	3.133	6.823

其次，根据数据统计分析的结果，这两种训练合成方法对字、词、句的基频预测基本一致，见表 4。可以看出两种方法下的基频预测效果相差不大。据统计，这两种训练合成方法对音节基频的预测相似度达到 81.12%。

表 4　两种合成音的字、词、句平均基频

	平均基频/Hz		
	字	词	句
普通话原声	236.623	236.679	239.629
古音系统合成音	214.180	216.981	227.381
拼音系统合成音	217.069	217.629	226.458

下图 4 是普通话句子“饭碗的含金量也越来越重要”的原声、拼音系统合成音、古音系统合成音三者的音高曲线的比较。不难看出，两种训练方法下的合成结果没有太大差异。

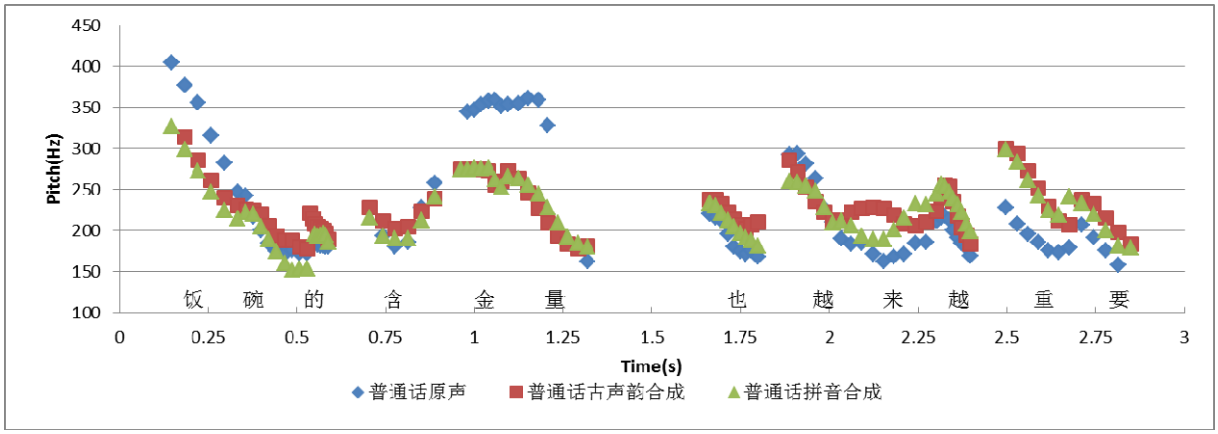

图 4　普通话拼音系统合成音与古音系统合成音比较

另外，基于拼音系统合成的普通话和基于古音系统合成的普通话在听感上的效果相差也不大。根据初步评测，在音质、流畅性、可懂度等方面，两种训练方法的合成效果也都基本相当。

鉴于这两种训练方法的合成效果并无实质差异，本研究初步认为，基于汉语古音系统的字音转写和语音训练合成方法是基本可行的。本问作者还采用长沙话和西安话的发音材料进行了试验，合成出来的语音具有明显的方言特色，而且能基本满足合成语音在可懂度方面的要求。

5 结论与讨论

本文尝试采用古音系统来转写汉语字音，并在此基础上设计相应的问题集以实现语音训练合成：（1）设计了一套面向汉语方言语音合成的通用发音文本；（2）录制了一些汉语方言点的合成语音语料库；（3）搭建了一套基于古音系统的汉语方言语音合成平台。

研究结果初步表明：基于汉语古音系统训练合成出来的普通话，在可懂度和音质等方面与基于拼音系统训练合成出来的普通话非常接近。基于古音系统进行汉语语音合成的方法是可行的。

本研究也注意到，语音训练合成的每个环节都会影响到最后的效果：如语料设计是否合理，发音人的语音系统是否稳定一致，语料标注是否正确，韵律标注是否充分，问题集是否全面等。因此提升语音合成效果将是一个不断完善的过程。进一步的研究还将考察如何利用已知的亲属方言的语音系统信息来弥补古音系统的不足，从而开展次方言的语音合成研究工作。

参 考 文 献

[1] 吴义坚. 基于隐马尔科夫模型的语音合成技术研究[D]. 安徽：中国科学技术大学，2006.
WU Yijian. A Speech Synthesis Technology Research Based on Hidden Markov Model [D]. Anhui: University of Science and Technology of China, 2006. (in Chinese).

[2] 刘磊. HTS 汉语合成及其自然度的研究[D]. 大连：大连理工大学，2006.
LIU Lei. Mandarin's Synthesis by HTS and Research on its Naturalness [D]. Dalian: Dalian University of Technology, 2006. (in Chinese).

[3] 胡琼. 基于隐马尔科夫模型的天津方言语音合成[D]. 上海：上海交通大学，2011.
HU Qiong. An Hmm-Based Speech Synthesis System Applied to Tianjin Dialect [D]. Shanghai: Shanghai Jiao Tong University, 2011. (in Chinese).

[4] 苏庄銮. 情感语音合成[D]. 安徽：中国科学技术大学，2006.
SU Zhuangluan. Affective Speech Synthesis [D]. Anhui: University of Science and Technology of China, 2006. (in Chinese).

[5] 吴宗济. 中国音韵学和语音学在汉语言语合成中的应用[J]. 语言教学与研究，2002，6（1）： 1-14.
WU Zongji. Chinese Historical Phonology and Phonetics in Chinese Speech Synthesis [J]. Language Teaching and Linguistic Studies. 2002, 6(1): 1-14. (in Chinese).

[6] 吴宗济，林茂灿. 实验语音学概要[M]. 北京：高等教育出版社，1989.
WU Zongji, LIN Maochan. Summary of Experimental Phonetics [M]. Beijing: Higher Educaion Press, 1989. (in Chinese) .

[7] 丁声树，李荣. 汉语音韵讲义[J]. 方言. 1981， 4：241-274.
Ding Shengshu, LI Rong. Chinese Phonology teaching materials [J]. FangYan. 1981, 4: 241-274. (in Chinese).

[8] 唐作藩. 音韵学教程[M]. 北京：北京大学出版社，2002.
TANG Zuofan. A Course in Historical Phonology [M]. Beijing: Peking University Press, 2002. (in Chinese) .

[9] 曹建芬. 语音处理上如何逐渐减少对具体语料的依赖？[J]. 清华大学学报（自然科学版），2009，S1（49）： 1380-1387.
CAO Jianfen, How to gradually decrease the dependence on specific speech materials in speech processing? [J]. Tsinghua University(Sci & Tech). 2009, S1(49): 1380-1387. (in Chinese).

[10] 侯精一. 现代汉语方言概论[M]. 上海：上海教育出版社，2002.
HOU Jingyi. A brief description of Modern Chinese Dialect [M]. Shanghai: Shanghai Education Publishing House, 2002. (in Chinese).

[11] 中国社会科学院语言研究所. 方言调查字表[M]. 北京：商务印书馆，2011.
The Language Institute of the Chinese Academy of Social Sciences. Word Tables in Dialect Survey [M]. Beijing: The Commercial Press, 2011. (in Chinese).

[12] 丁声树，李荣. 古今字音对照手册[M]. 北京：中华书局，1981.
Ding Shengshu, LI Rong. Ancient and Modern Pronunciation Comparison Manual [M]. Beijing: Zhonghua Book Company, 1981. (in Chinese).

(原载于《清华大学学报（自然科学版）》 第 53 卷 第 6 期)

Speech synthesis for Chinese dialects based on the phonological system of ancient Chinese

HUANG Xiaoming[1] , XIONG Ziyu[2]

1. Department of Linguistics, Graduate School of Chinese Academy of Social Sciences, Beijing 102488, China

2. Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China

Abstract: The HTS speech training synthesis tools and STRAIGHT speech synthesizer are used for speech training synthesis of an unknown dialect phonetic system. This study applies these methods to various Chinese dialects by designing a set of universal pronunciation texts for Chinese dialect speech synthesis with recording of a synthesized speech corpus for some Chinese dialects. A Chinese dialect speech synthesis platform is then built based on the phonological system of ancient Chinese. Results for Mandarin speech synthesis show that the synthesized sound based on the phonological system of ancient Chinese is quite close in terms of intelligibility and quality to the synthesized sound based on the pinyin system. Thus, the speech synthesis based on the phonological system of ancient Chinese for Chinese dialects is effective and feasible.

Key words: phonological system of ancient Chinese; dialect; speech synthesis

（原载《清华大学学报》（自然科学版）第 53 卷第 6 期）