

基于音系理论的变调自动处理模型*

贺俊杰

南开大学外国语言文学博士后流动站 天津 300071 陕西师范大学外国语学院 西安 710062

E-mail: junjie_he@126.com

摘要: 汉语各方言表现出来的变调现象十分复杂。本文以在变调模式和机制方面具有代表性天津话和丹阳话等方言为例,说明基于管辖音系学的变调自动化处理模型具有跨方言的算法内核一致性、算法代码精简性以及方便的可移植性等特点。这些优点都源于管辖音系学特有的声调表征模式及声调音系过程解释机制。

关键词: 变调, 自动处理, 音系学, 管辖音系学

Phonology-based Automation of Tone Sandhi

He Junjie

Post-doctoral Station on Foreign Languages and Cultures, Nankai University, Tianjin 300071

School of Foreign Languages, Shaanxi Normal University, Xi'an 710062

E-mail: junjie_he@126.com

Abstract: Chinese exhibits complex tone sandhi phenomena across its dialects. This article sets forth a discussion on phonology-based automation of sandhi processes of Tianjin and Danyang Chinese, with a conclusion that automation algorithm based on tone theories in Government Phonology boasts several advantages—the sandhi algorithm has a principled and unified core; the sandhi code is simple and concise; the algorithm can be easily ported to other languages or dialects. The above-mentioned advantages come from the unique explanatory device of the Government-phonological tone theory.

Keywords: Tone sandhi, Automation, Phonology, Government Phonology

1 引言

本文在管辖音系学 (Government Phonology) 声调理论^[1-2]框架内,以在变调模式和机制方面具有代表性的天津话和丹阳话等方言为例,运用计算语言学方法,对变调的计算机自动处理算法进行研究,提出基于管辖音系学声调理论的声调表征和变调处理算法,以检验该理论的精简性和可行性,并证明此算法模型具有跨方言的算法内核一致性、算法代码精简性以及方便的可移植性等特点。

2 声调系统生成模型

2.1 管辖音系学声调理论简介

管辖音系学因其分析极具特色,在国外音系学界影响颇大。自1980年代中期产生以来,它在音段结构、音节结构和韵律结构等领域取得了重要的进展。它严格遵循生成语言学的理论目标和研究方法,具有论证严谨、体系简约等特点,是一种非常重要的音系学研究范式。

* 此文承国家自然科学基金项目《管辖音系学声调理论研究:模型构建及应用》(项目号:09CYY001)的资助,同时为作者在南开大学外国语言文学博士后流动站期间的研究成果。

为了有效制约语法过强的生成能力,管辖音系学试图通过有限的允准制约原则来规范有限的表征元素的组构形式,使冗余的表达形式数量降到尽可能低的水平。通过界定相邻音系成分间有限的几种管辖关系,音系表达式的合格性(well-formedness)评估就有章可循——某音系域内除该域主位以外其他位置上的音系元素的出现都需要得到允准,而允准的基础是管辖关系。

Kaye等^[1]提出了音系的表达系统,采用独值特征系统对音段进行描写。Kaye等^[4]以及Kaye^[5-6]提出了首音、韵音和核心等三种音节成分,明确了音节成分结构上的严格二分原则,并界定了成分内部以及成分之间的几种管辖关系;提出了投射原则(Projection Principle),从而排除了如音节重新划分(resyllabification)等一些需要改变词库表达层面业已确立的管辖关系的音系推导过程;还提出了空范畴原则(Empty Category Principle)、尾音允准原则(Coda Licensing Principle)等重要音系原则,进一步制约了音系表达。

在声调音系现象的解释上,Kaye^[1]提出了面向曲线声调表达的声调理论模型,贺俊杰^[2]对这一模型进行了理论拓展,针对主位设定方式提出了理论修订,认为主位设定只有位置型。取消内在型主位设定后,延伸方向就仅仅由主位位置蕴含,无需独立设定。扩展后的声调理论能为声调音系过程提供解释。

2.2 计算机实现及生成系统示例

基于扩展后的声调理论^[2],声调语言各方言间声调系统上的差异可以通过各系统间在调元数目、主位、延伸和允准制约条件选择等参数设置上的差异得到解释。下面我们通过构建跨语言的声调表征算法,以汉语普通话、天津话、德州话、厦门话、长沙话、广州话、西安话、丹阳话和镇江话为例,借助语音合成实验来说明管辖音系学声调理论的工程学意义和应用价值。

2.2.1 跨语言声调系统生成参数设定

下面先以天津话的声调系统为例说明某声调系统中允准制约条件、主位和延伸等参数设置在系统生成过程中的作用。天津话有四个调类,阴平、阳平、上声和去声。它们的语音表现分别为低平调、高平调、低升调和高降调,调值描写分别为11、55、24和53。它们在音系上的价值分别是LL、HH、LH和HL。根据管辖音系学声调表征理论,这四种调类调型的音系表达应该是阴平“L”、阳平“H”、上声“LH”和去声“HL”。其中下加着重号的为调型主位,位于主位上的调元强制性延伸至无调的非主位,从而形成平调,这样的声调表征方式是依据音系理论中的一条强制曲拱制约条件而产生的。这条制约条件要求相邻位置不允许出现相同的调元,即OCP1。由于天津话属于复杂I型声调系统,该型系统有两个调元H和L,将它们和占位元素进行两两组合,共有 3^2 种组合。¹其中四种合法组合须通过OCP1以及另一条允准制约条件的共同作用得以生成。(3)是天津话声调系统的管辖音系学模型示例。

(3) 天津话声调系统

- a. 声调类型: 复杂I型;
- b. 声调延伸: “H”, “L”, 强制性;
- c. 声调主位: 主位居尾;
- d. 允准制约条件: 1. OCP1(*XYX..., $X \in \{H, L\}$, $Y \in \{H, L, \phi\}$; $X \neq Y$);
2. 主位必有调, 即X, 其中 $X \in \{H, L\}$ 。

表1给出了更多语言(方言)的参数设置。从该表我们可以观察到模板型语言的允准制约条

¹ 模板型语言中,最小的音系范域由 $O_1N_1O_2N_2$ 模板定义^[7-8],每个调型与两个核心联接。

件、主位设定以及延伸设定等参数设置与相应语言（方言）之间的关系。需要说明的是，表中所标“？”说明该声调系统参数组合尚未找到其对应的语言实例。这是因为其相应的允准制约条件与管辖音系学声调理论中高调主位性理论的要求存在较大差距。也就是说，高调主位性理论对语言事实有一定的预测能力。

复杂 I 型系统					
允准制约条件	主位	延伸	合法调型	非法调型	语言示例
OCPI: 调型内须有高调元	居首	强制性 (H、L)	$\underline{H}, \underline{H}, LH, HL$	$\underline{L}, \underline{L}, \underline{\quad}$	汉语普通话
OCPI: 调型内须有低调元	居首		$\underline{L}, \underline{L}, HL, LH$	$\underline{H}, \underline{H}, \underline{\quad}$?
OCPI: 主位须有调	居尾	强制性 (H、L)	$\underline{H}, \underline{L}, LH, HL$	$\underline{H}, \underline{L}, \underline{\quad}$	天津话
OCPI: 主位须有调	居首	强制性 (H、L)	$\underline{H}, \underline{L}, HL, LH$	$\underline{H}, \underline{L}, \underline{\quad}$	德州话, 西安话,
OCPI: 非主位须无调	居首		$\underline{H}, \underline{L}, \underline{\quad}$	$\underline{H}, \underline{L}, HL, LH$?
OCPI: 主位须无调	居首		$\underline{H}, \underline{L}, \underline{\quad}$	$\underline{H}, \underline{L}, HL, LH$?
OCPI: 非主位须有调	居首		$\underline{H}, \underline{L}, HL, LH$	$\underline{H}, \underline{L}, \underline{\quad}$?
OCPI: 调型内仅有一个调元	居首	强制性 (H、L)	$\underline{H}, \underline{H}, \underline{L}, \underline{L}$	$HL, LH, \underline{\quad}$	镇江话
OCPI: 调型内最多只有一个调元	居首	强制性 (H、L), 词库标注不延伸 (H、L)	$\underline{H}, \underline{H}, \underline{L}, \underline{L}, \underline{\quad}$	HL, LH	广州话
OCPI: 调型内最多只有一个调元	居首	强制性 (H、L), 词库标注不延伸 (L)	$\underline{H}, \underline{H}, \underline{L}, \underline{L}, \underline{\quad}$	HL, LH	长沙话
OCPI: 调型内最多只有一个调元	居尾	强制性 (H), 词库标注不延伸 (H、L)	$\underline{H}, \underline{H}, \underline{L}, \underline{L}, \underline{\quad}$	LH, HL	武鸣壮话
OCPI: 低调不居主位	居首	强制性 (H、L)	$\underline{H}, \underline{H}, HL, \underline{L}, \underline{\quad}$	\underline{L}, LH	厦门话
OCPI: 高调不居非主位	居首	强制性 (H、L), 词库标注不延伸 (L)	$\underline{H}, HL, \underline{L}, \underline{L}, \underline{\quad}$	\underline{H}, LH	玉林话
OCPI: 低调不居非主位	居首		$\underline{L}, LH, \underline{H}, \underline{H}, \underline{\quad}$	\underline{L}, HL	?
OCPI: 高调不居主位	居首		$\underline{L}, \underline{L}, LH, \underline{H}, \underline{\quad}$	\underline{H}, HL	?

简单型系统					
允准制约条件	主位	延伸	合法调型	非法调型	语言示例
OCPI	居尾	强制性(H), 词库标注不延伸(H)	$\underline{H}, H, \underline{\quad}$		丹阳话

表1 跨语言(方言)声调系统参数设置

2.2.2 跨方言声调生成与合成

根据管辖音系学声调系统, 可以设计出跨语言(方言)声调系统的生成算法, 并构建跨语言(方言)声调生成系统, 如图1所示。图2所示是通过该系统的声调生成与合成机制, 对字符串“南开大学外语学院”的声调层进行跨方言置换, 说明了该算法的合成效果具有很高的可懂度和跨方言区别度。图中显示所生成的各示例方言的声调层在示例字符串上的音高曲线走势。将依据各方言声调系统所生成的示例字符串声调层与原始发音的声调层进行置换, 即可获得目标方言的合成语音。经过听辨实验, 合成语音的方言声调特征明显, 所属方言归属辨识率高。

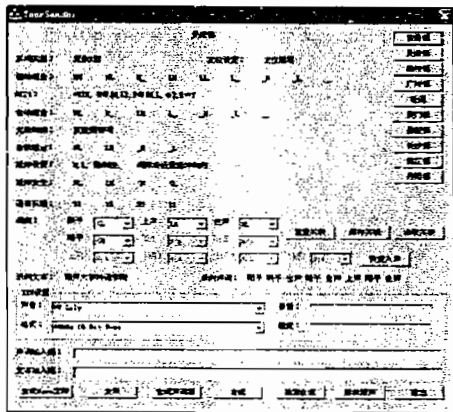


图1 跨语言(方言)声调生成系统截图

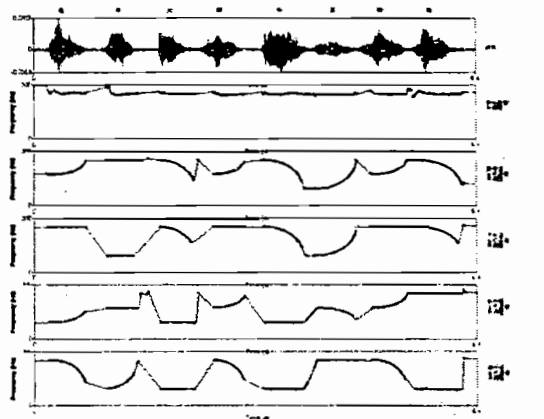


图2 跨方言声调层置换合成示例

3 变调算法举例

下面以天津话二字组连读变调和丹阳话变调为例, 提出它们的变调算法, 以此说明基于语言学理论研究成果的言语工程算法具有较好的通用性和可移植性。

3.1 天津话二字组连读变调算法

由天津话四种合法调型组成的二字组共有16种组合方式, 其中12种因未满足变调触发条件不引发变调, 其余的4种调型组合因满足变调触发条件而可能产生变调。其它是:

阴平+阴平 ($\leq L + \leq L$), 上声+上声 (LH+LH), 去声+去声 (HL+HL), 去声+阴平 (HL+ $\leq L$)。

(4)是天津话二字组连读变调系统的管辖音系学模型。以上四种变调二字组组合均因违反(4)中的变调触发条件而引发变调。变调过程中需经历若干变调音系操作, 每一操作的施用都是在变调触发条件、操作施用原则、变调制约条件以及该声调系统允准制约条件的共同作用下完成的。

(4) 天津话二字组连读变调系统

变调域: 前字, 即前字变调型。

变调触发条件: 1. OCP2(*XYXY, X,Y ∈ {H,L}; X≠Y);

2. OCP3(*LL.LL);

3. OCP4(*L.LL)。

变调制约条件: 1. 遵守允准制约条件;

2. 不违反其他的变调触发条件。

变调操作: 1. 增加; 2. 删除; 3. 邻接 (因声调系统允准制约条件有“主位必有调”)。

(4)能直接转换成计算机算法, 输出正确的二字组连读变调结果。对于符合变调条件的二字组, 通过对其变调操作空间中各种候选操作进行变调经济性评估, 然后选择其中经济性最高的变调操作对原声调组合施以变调作用, 即可得到相应变调组合的最终变调形式。

(5)是天津话三字组连读变调的管辖音系学分析模型, 它为天津话三字组的变调方向性问题提供了音系学解释。其中变调的单方向假设在工程上实现也有着简化算法的优势。

(5) 天津话三字组连读变调

变调全域: 三字组;

变调触发条件: 同二字组变调触发条件;

变调搜索方向: 右向搜索。

变调制约条件: 前二字组变调尝试过程中, 对于中字调型呈弱势管辖的三字组, 变调全域内不允许因变调而产生违反 OCP2(*XYXY, X,Y ∈ {H, L}; X≠Y)或 OCP(*L.L.LL)的调型组; 如产生, 则放弃当次变调尝试, 往下搜索变调可能性。

3.2 丹阳话二字组连读变调机制

丹阳话的变调现象十分复杂, 变调后的字调与单字调没有明显而整齐的对应关系。不同的单字调有可能变为相同的变调, 而某个单字调也有可能不同的环境下变成不同的变调。

图 3 是丹阳话连读变调的管辖音系学模型, 图 4 是该模型程序实现对由 1345 个二字组构成的测试集进行模拟测试的截图。该模型模拟了母语习得者变调字组的习得路径。其中前、后字底

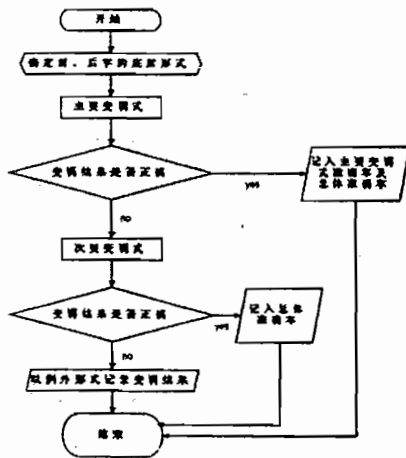


图 3 丹阳话连读变调习得路径暨算法流程

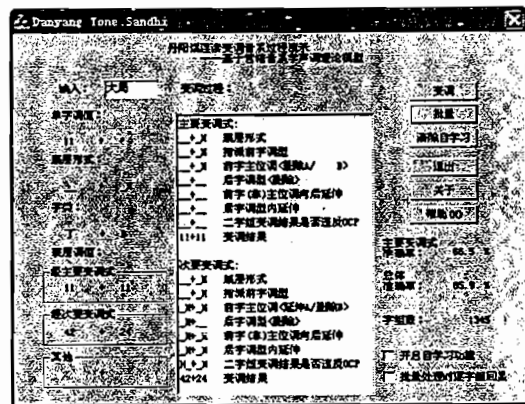


图 4 丹阳话连读变调音系过程程序演示截图

层形式的确定包括：根据前字的分类标准，指派前字类别，确定前字形式 ($\langle H, \underline{H}(E), \underline{H}(C), \underline{\quad} \rangle$)；根据后字的分类标准，指派前字类别，确定后字类别 (A、B)。主要变调式的音系过程为：A类后字致前字主位调删除，B类后字不影响前字调型；后字调型删除；前字声调向后延伸；判断是否违反 OCP，并选择相应音系过程。而次要变调式的音系过程为：A类后字致前字（甲乙丙）主位调延伸，B类后字致前字主位调删除（丁类为增加）；后字调型删除；前字声调向后延伸；判断是否违反 OCP，并选择相应音系过程。

母语者在习得以上变调“规则”后，如按主要变调式变调，产生正确结果的比例分别为：

首字平清：87.0% (42-11)、60.3% (55-55)；

首字平浊：41.7% (42-11)、85.2% (24-55)；

首字仄清：74.4% (33-33)、68.8% (42-24)；

首字仄浊：65.7% (11-11)。

若按主要变调式变调出现错误结果，则按次要变调式进行变调，这样获得正确结果的比例分别为：

首字平清：4.3% / 13.0%² (55-55)、25.5% / 39.7% (42-11)；

首字平浊：43.1% / 58.3% (55-55)、7.9% / 14.8% (42-11)；

首字仄清：11.6% / 25.6% (42-24)、25.7% / 31.2% (33-33)；

首字仄浊：27.5% / 34.3% (42-24)。

因此，只需依据简单的两种变调模式，母语者就能将连读变调的准确性达到相当高的水平：

首字平清：91.3% (42-11, 55-55)³、85.8% (55-55, 42-11)；

首字平浊：84.8% (42-11, 55-55)、93.1% (24-55, 42-11)；

首字仄清：86.0% (33-33, 42-24)、94.5% (42-24, 33-33)；

首字仄浊：93.2% (11-11, 42-24)。

以上理论解释与算法设计为丹阳话复杂的变调现象提供了简明的解决方案。其理论贡献主要表现在：(1) 通过对管辖关系作强、弱勢的区分，明确了在仄声前字向后延伸时，后字的管辖类别所起的决定性作用——不同管辖关系的后字对前字调型起到不同的支配作用。(2) 入清字作为后字时跨 A、B 两类分布也能找到理论根据。因为入清的音系表征为 $(H)\underline{\quad}$ ，呈弱勢管辖，弱勢管辖对前字仄清选择 $\underline{H}(C)$ 的实现不起作用，所以在仄声后入清字归 A 类，而在平声后则归 B 类。

(3) 丹阳话变调中只有前进型变调，而无后退型变调，也没有首尾定调或嵌入式变调。丹阳话的变调其实如吕叔湘^[9]所认为的那样，变调组的首字调与其历史调类有对应关系，后字是前字调的延伸结果，在 III 式中确实存在异化过程。异化操作在音系上的解释就是主位调被删除及自动邻接。异化操作受制于 OCP2(*XY.XY, X, Y ∈ {H, L}; X ≠ Y)，而此 OCP 在丹阳话中仅有 24-24 连调组适用。以上理论贡献表现在工程算法上就是简化算法，并为过程处理提供语言学理论支持。

4 结语

在计算机上模拟实现根据语言理论构建的音系处理模型能让我们了解该理论的可计算性和

² 斜线后的比例是除主要变调以外总的剩余比例。

³ 括号内前者为主要变调，后者为次要变调。

算法的复杂度。以丹阳话变调为例,变调后的调值与变调前的单字调没有整齐的对应关系,所以二字组及多字组中变调字的调值似乎无法用规则推导。这种单字调与变调之间多对多的对应关系在某些声调音系理论^[10]中需要建立与单字调分离的变调集(按包智明^[10]的处理就是词调)。字组进入变调过程需抛弃单字调,而赋予词调再经历变调音系过程。若依这种音系模型构建变调自动模型,势必在建立单字调值库之外,还需将所有不同环境下变调字的调值也建库以供检索。这种处理方式不够经济,也没有充分运用语言学的研究成果。若以管辖音系学声调理论为理论依据,则仅需依据普遍制约条件并辅之以有限的变调操作就能建立起丹阳话的变调算法。

天津话三字组连读变调涉及到的变调方向性问题引起了音系学界的广泛讨论。如果认为变调确实存在方向性,并且不同的变调式可能采用不同的变调方向,这就要求变调算法采取条件分支处理的方式。但如果认为变调仅有一个方向,也就是只需要设定单方向处理机制的话,那么变调算法将不再需要为不同的变调式人为设定条件分支处理模块,仅仅需要核查变调过程是否违反OCP(Obligatory Contour Principle, 强制曲拱原则),进而采用不同的处理方式。显然,这种朝不同处理方式流向的判断机制不是来自语言系统的外部(即根据不同的字组进行人为指定),而是来自具有语言学普遍意义的OCP原则。

最后,本文虽未提及汉语普通话的连读变调,但我们仍可将其纳入本研究的跨方言声调生成和变调系统。普通话的连读变调中具有音系学意义的仅有“上上”变调。其变调过程虽不算复杂,但如果计算机处理模型对此变调过程进行模拟时,采用将前字上声调直接向阳平调整体转换的算法,就体现不了当代音系学关于声调结构的研究与自然语言计算机自动处理的有机结合。因为那样并没有回答某底层调为何恰恰变为某表层调,而非其他调的问题。如果能从音系结构出发为其找到理据并在计算机算法中得以简明实现,那将大大减少算法的随意性,并将提高算法的通用性。也就是说,在普通话的变调算法与天津话的变调算法,乃至其他方言的变调算法中,将重用相同的变调算法内核。

参 考 文 献

- [1] Kaye J. A short theory about tones. Unpublished manuscript, 2001. Available at: <http://134.59.31.7/~scheer/scan/Kaye01shorttoneMs.pdf>
- [2] 贺俊杰. 管辖音系学的声调理论: 发展及应用. 当代语言学, 待刊.
- [3] Kaye J, Lowenstamm J, Vergnaud J.-R. The internal structure of phonological representations: a theory of charm and government. *Phonology Yearbook*, 1985, 2:305~328.
- [4] Kaye J, Lowenstamm J, Vergnaud J.-R. Constituent structure and government in phonology. *Phonology*, 1990, 7(2):193~231.
- [5] Kaye J. 'Coda' licensing. *Phonology*, 1990, 7(2):301~330.
- [6] Kaye J. Government in phonology: the case of Moroccan Arabic. *The Linguistic Review*, 1990, 6:131~159.
- [7] Goh Yeng-Seng. *The Segmental Phonology of Beijing Mandarin*. Taipei: The Crane Publishing, 1997.
- [8] 吴英成. 北京话是单音节语言的质疑. *当代语言学*, 1997, 4:231~247.
- [9] 吕叔湘. 丹阳方言的声调系统. *方言*, 1980, 2:85~122.
- [10] Bao Zhi-Ming. *On the nature of tone*. Ph.D. dissertation, MIT, Cambridge, MA, 1990.