汉语普通话的管辖音系学特征及提取方法

刘娇蛟,贺前华,韦 岗
(华南理工大学电子与信息学院,广东广州 510641)

摘 要: 语音识别中多采用音素作为识别单元,因其数量较多,对神经网络训练复杂度的要求高,在多语言语音识别中需要针对不同语言分别建立识别模块.然而,管辖音系学提出了适用于多语言的语音学特征.本文根据英语和汉语发音的相似性,确定汉语普通话声韵母的 GP 特征表示形式,并应用神经网络实现特征提取.实验表明,GP 特征同样可作为汉语语音的语音学特征.

关键词: GP 特征; 特征提取; 普通话; 多语种识别
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 0372-2112 (2006) 10-1917-03

Mandarin GP Features and Extraction Method

LIU Jiao-jiao, HE Qian-hua, WEI Gang
(*School of Electronic and Information, South China University of Technology, Guangzhou, Guangdong* 510640, *China*)

Abstract: Phonemes are often used in speech recognitions. The amount of phonemes makes the RNN's training complicated. Moreover, different modules need to be built up for different languages in multi-language speech recognition. Government Phonology (GP) provides new primes, which can be used to describe the speech features of different languages. According to the similarity of the pronunciation of English and Mandarin, we introduced GP expressions of Mandarin phonemes and used RNN to extract mandarin GP features. As the experiment results show, GP primes can also be used in describing Mandarin speech features.

Key words: GP features; feature extraction; mandarin; multi-speech recognition

1 引言

音素可以看作构成语音的基本单位——原子,因此常常作为自动语音识别的基本单元.但是,构成不同语言的音素差异很大,要实现多语言识别往往要针对不同语言建立多个语音识别模块.然而,正如原子不是构成物质的最小单位一样,在音系学领域中,人们大都认为音素还可以分成更小的元素(prime).这些元素的数量不多,相互独立,而且对于不同的语种来说都是不变的,它们通过不同的聚合或分裂方式可推导出不同的音素.管辖音系学(Government Phonology,简称 GP)就是这样的一种音系理论.

2 管辖音系学基本理论

管辖音系学是生成语法框架内的一种非线性音系学理论,其理论研究始于 1982 年,经过 20 多年的不断发展,已成为生成音系学的主要理论之一.该理论扬弃了经典生成音系学的规则部分,保留了生成音系学从底层到表层的推导过程,并指出音系表征是按照一组原则和参数从一个固定的元素集推导而来的,从而发展了以音节成分之间的管辖关系为核心的管辖音系学基本理论[1,2].在修正后的 GP 理论中,这个固定的特征集只包含了六个元素,在众多跨语种的语言差异面前,如此少量的元素怎么能够满足所有人类语音的识别呢?事实上,在语音中含有各种信息,但只有一部分和语音学相关.比如,语音信号还传递了像性别,年龄、群体关系等信息.多语种语音识别要求能够滤除语音中不相关的部分并把精力集中在语音学特征上,即本文提出的管辖音系学特征.

在自动语音识别中,应用源于音系学特征和元素的概念并不鲜见[3~8],管辖音系学利用六个特征元素来描述发音器官的形状,这种全局特性使其格外适合多语种识别模型.

3 GP 元素及其音系表达

3.1 GP 元素集合

根据 Kaye 的 GP 理论,GP 特征集包含六个元素[8,9]: $E=\{A, I, U, H, L, ?\}$.具体含义如表 1 所示,其中"-"元素作为标记元素(identity element)使用.

表 1 Kaye 元素集

元素名称	含义	对应表达式	英文举例词
A	非高位性	*A*	father
I	腭音化	*I*	me
U	唇音性	*U*	too
H	紧声门	({*H*}, -)	horse
L	松声门	({*L*, ?}, -)	sing
?	闭塞性	({?}, -)	go
-	空	({}, -)	kisses

3.2 GP 音系表达式及组成结构

根据 GP 理论,所有的语音均可以表示为 GP 特征的音系表达式:(*O*, *H*).其中,*O* 被称为操作数(Operator),*H* 被称为头或主位(Head),主位起允许制约其操作数的作用.

收稿日期:2005-11-10;修回日期:2006-03-16
基金项目:国家自然科学基金(No. 60572141),广东省自然科学基金(No. 36562)

GP 理论中没有音节的概念,但是存在类似音节的成分,即音系组成结构[9],具体包括:首音(Onset)、核心(nucleus)和韵音(rhyme).在骨架图 1 中可以看到,这些音系组成的投影是用骨架层 x 来描述的,而骨架层直接与不同 GP 音系表达式相对应.

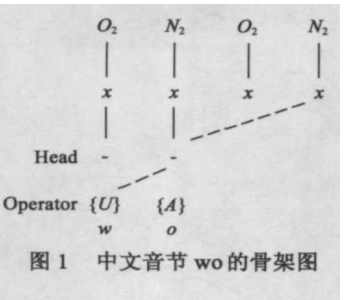

图 1 中文音节 wo 的骨架图

3.3 GP 音系表达式的制约规则

Kaye 提出可以通过一组参数化的准许制约原则对元素的组构加以制约,从而生成特定语音的各种音素.到目前为止,语言学家们对首音的制约原则了解不多,但对于核心成分的制约原则却有比较充分的研究,汉语普通话核心成分内的允许制约规则[9]如表 2.根据 GP 理论,其他韵母(在 GP 理论中由核心集描述,下同)的发音是通过各组成结构之间的管辖关系实现的.图 1 展示了中文音节 wo 的 GP 特征形成原理.通过组成成分间的管辖关系及其制约规则,可以推导出中文 GP 表达式的所有核心集.但是,语言学家在这方面的研究还处于探索之中,完全罗列出所有韵母的 GP 表达式在目前具有一定困难.

表 2 普通话核心内制约规则示例

({}, *I*)	ni
({}, *U*)	mu
({}, *A*)	ma
({*A*}, .)	k7

3.4 汉语韵母的 GP 表示

Simom King 和 Paul Taylor 使用延时递归网络[10],给出了三种特征系统的英语音素和特征之间的一一映射,将其整合到 Kaye 的六元素理论上,可以得到 GP 特征和英语音素之间的对应关系.英语和汉语中的音素虽然有差异,但部分音素发音基本相同或相似.本文从主观听觉上对中英语音音素进行比较[11],根据相似性给出汉语中部分韵母的 GP 特征表示,其中 1 表示含有特定元素,0 则反之,如表 3 所示.

表 3 汉语部分韵母的 GP 特征表示

GP feature system		1	2	3	4	5	6	Head		
ID	phone	*A*	*I*	*U*	?	*H*	*L*	*a*	*i*	*u*
1	a	0	0	0	0	0	0	1	0	0
2	i	0	0	0	0	0	0	0	1	0
3	u	0	0	0	0	0	0	0	0	1
4	e	1	0	0	0	0	0	0	0	0
5	ai	1	1	0	0	0	0	1	1	0
6	ei	1	1	0	0	0	0	0	1	0
7	ao	1	0	1	0	0	0	1	0	1
8	ou	1	0	1	0	0	0	0	0	1

4 实验和结果

4.1 实验原理

利用 GP 特征实现语音识别,首先必须进行特征提取,确定语音流中各音素包含的 GP 特征,然后可以采用 HMM 识别模块实现自动语音识别,具体识别流程如图 2 所示.

鉴于 Kaye 元素集中各元素的相互独立性,拟采用六维向量空间来描述首音或核心的 GP 特征空间,每一个音系组成结构映射为向量空间中的一个向量 υ_i,六个正交的单位矢量作为训练时的教师信号,即单位特征向量 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_6$.

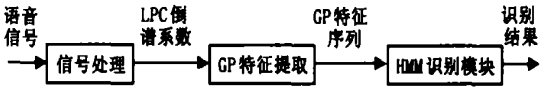

图 2 基于 GP 特征的自动语音识别流程

如果 $\forall \upsilon_i \in R^6$,并且 $\upsilon_i = (w_{i1}, w_{i2}, \cdots, w_{i6})$,$\varepsilon_j = (w'_{j1}, w'_{j2}, \cdots, w'_{j6})$,$\upsilon_i$ 与 ε_j 的相似度[12]如式(1)所示:

$$sim(\upsilon_i, \varepsilon_j) = \cos(\upsilon_i, \varepsilon_j) = \frac{\sum_{k=1}^{6} w_{ik} \cdot w'_{jk}}{\sqrt{(\sum_{k=1}^{6} w_{ik}^2)(\sum_{k=1}^{6} w_{jk}'^2)}} = \frac{w_{ij}}{\sqrt{\sum_{k=1}^{6} w_{ik}^2}} \quad (1)$$

可见,υ_i 在坐标轴上的归一化投影值就说明了它与该特征之间的相似度.如果该相似度超过一定门限(经验值),就认为含有这个坐标轴所代表的特征,具体判别过程如下:

$$\begin{aligned} &\text{if } sim(\upsilon_i, \varepsilon_j) \geq \text{threshold, then } j \in \upsilon_i; \\ &\text{else if } sim(\upsilon_i, \varepsilon_j) < \text{threshold, then } j \notin \upsilon_i \end{aligned} \quad (2)$$

4.2 实验及结果

本实验从汉语普通话的核心集和首音集中各选 8 个音素,包括:b、c、ch、n、q、ch^{w6}、sh^{w6}、zh^{w6}[9]和 a、i、u、e、ao、ou、ai、ei,每个音素选取 50 个数据.其中,28 个数据构成集内训练数据,另外 22 个数据作为集外测试数据.实验数据是从 863 中文语音库中通过手工标记获取的,并采用 12 阶 LPC 倒谱系数表示.其中,帧长为 30ms,帧移为 10ms,汉明窗长为 30ms.

实验中采用 Elman 递归神经网实现特征提取,它具有与 MLP 网络相似的多层结构,采用 BP 算法进行训练,并具有一个特别的隐层.该层从普通隐含层接收反馈信号,输出被前向至隐含层,形成一个局部反馈网络,因此该网络适合于语音等非线性动态系统的辨识[13].由于 GP 特征个数少,与直接识别音素相比,神经网络训练的复杂度要求大大降低.在本次实验中,神经网络的隐层只采用了 15 个神经元.同时,为了描述 GP 表达式中首位对操作数的控制作用,在实验中还引入了首位 a,i,u 值的编码.因此,包括 6 个 GP 元素在内,神经网共有 9 个输出.本实验在 Matlab 6.5 平台上实现,具体识别过程如图 3 的仿真所示.图 3 的横轴坐标是待识别的特征元素,纵坐标是该音素在正交坐标轴上的投影相似度.可见,(*a*)图中音素 ou 的 3 个特征元素都可以正确识别,而(*b*)图中对音素 ai 的特征提取过程中,特征元素 a 出现漏判.

类似于信息分类的评价指标,可以采用准确率 P(precision)、召回率 R(recall)作为特征提取的性能指标[14].为了减少漏判和错判的情况出现,对于一个好的特征提取模块,应该兼顾准确率和召回率.表 4 列出了对训练集内和集外的首音、核心 GP 特征提取的测试结果.考虑到 GP 特征生成音素的制约规则是非线性的,本实验采用了递归神经网来模拟音节成分间的这种非线性允许制约关系,提取测试音素中的 GP 特征.相比 Simon 等人在文献[10]对英文 GP 特征的提取结果,

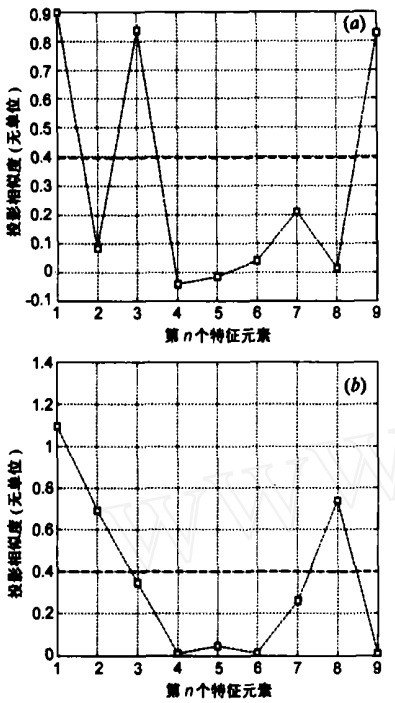

图 3　韵母 ou 和 ai 的 GP 特征提取

本文在增加训练音素规模的基础上,针对汉语普通话的 GP 特征提取也可以达到集外精确度为 89.7%的平均值.可见,GP 特征同样也适用于描述汉语普通话的语音学特征,一定程度上验证了它具有跨语言的表征能力.

表 4　普通话 GP 特征提取结果

		集内		集外	
		Precision(%)	Recall(%)	Precision(%)	Recall(%)
操作数主位	A	82.74	95.86	82.83	84.54
	I	97.62	94.25	86.27	91.67
	U	79.29	94.87	85.92	96.83
	H	100.00	100.00	100.00	97.18
	L	100.00	96.55	100.00	92.31
	?	95.92	95.92	97.09	97.09
	a	96.43	85.99	73.86	90.28
	i	98.47	97.97	95.00	96.94
	u	90.48	80.85	86.36	78.08
	平均	93.44	93.59	89.70	91.66

5　小结

相比于直接识别音素而言,利用 GP 特征可以简化语音识别的复杂性,并为跨语言识别提供了可能,通过对识别算法的进一步改进可以有效提高特征提取的精确率和召回率.如果能确定不同语言的语音在形成时的内在允许制约规则,该特征在跨语言识别方面具有明显的优越性.

参考文献:

[1] Ma Qiuwu. Government phonology: its theoretical framework and recent development [J]. Contemporary Linguistics, 2000, 2(4): 218 - 226. (in Chinese)

[2] Ma Qiuwu. Government phonology: a constraint-based phonological theory [J]. Journal of PLA Foreign Languages University, 2000, 23(1): 15 - 20. (in Chinese)

[3] L. Deng, K. Erler. Hidden markov model representation of quantised articulatory features for speech recognition [J]. Computer, Speech and Language, 1993, 7(3): 265 - 282.

[4] K. Hübener, J Carson-Berndsen. Phoneme recognition using acoustic events [A]. Proceedings of ICSLP'94 [C]. Yokohama, 1994. 1919 - 1922.

[5] M A Huckvale. Tiered segmentation of speech: opportunities, methods, problems and challenges [J]. Speech, Hearing and Language Work In Progress, Phonetics and Linguistics, University College London, 1993, 7: 133 - 152.

[6] K Kirchhoff. Phonetic features in speech recognition: a delayed synchronisation approach [D]. Masters' thesis, University of Bielefeld, 1995.

[7] A Kornai. Formal Phonology [D]. PhD Thesis, Stanford University, 1992.

[8] Williams G, G Martindale, AM Terry, JD Kaye. Multi-lingual speech recognition using phonological primes [A]. ASA-AJA conference [C]. Honolulu, 1996.

[9] Kaye J D. A Users' Guide to Government Phonology (GP) [DB/OL], Ms., University of Ulster, http://www.unice.fr/dsl/tobweb/scan/Kaye00guideGP.pdf, 2000.

[10] Simon King, Paul Taylor. Detection of phonological features in continuous speech using neural networks [J]. Computer Speech and Language, 2000, 14(4): 333 - 353.

[11] LIU Rui. Speaker Conversion Technology based on HMM [D]. Guangzhou: Masters' thesis, South China University of Technology, 2004. (in Chinese)

[12] Apte C, Dunerau F, Weiss S. Automated learning of decision rules for text categorization [J]. ACM Transactions on Information System, 1994, 12(3): 233 - 251.

[13] CONG Shuang, GAO Xue-peng. Recurrent neural networks and their application in system identification [J]. Systems Engineering and Electronics, 2003, 25(2): 194 - 197. (in Chinese)

[14] HUANG Xuan jing, WU Li de, Ishizaki Hiroyuki et al. Language independent text categorization [J]. Journal of Chinese Information Processing, 2000, 14(6): 1 - 7. (in Chinese)

作者简介:

刘娇蛟　女,1976 年 4 月生于江西南昌,讲师,主要研究领域:语音信号处理、信息安全等.
E-mail: jjliu@scut.edu.cn