中国科学: 信息科学 2015 年 第 45 卷 第 7 期: 849–868
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

汉语神经分析系统研究现状与展望

张少白*, 王勇, 何利文, 成谢锋

南京邮电大学计算机学院, 南京 210046
* 通信作者. E-mail: adzsb@163.com

收稿日期: 2014–07–05; 接受日期: 2014–11–05; 网络出版日期: 2015–05–20
国家自然科学基金 (批准号: 61373065, 61271334) 资助项目

摘要 在神经生理学和神经解剖学的基础上仿真和描述大脑中涉及语音生成和理解区域的相关功能是目前人工语音合成系统的重要研究领域. 波士顿大学语音实验室 Guenther 教授及其所带领的研究小组成功研制出了一种称之为 "神经分析系统 (Neuralynx System)" 的仪器. 这种仪器可以让使用者将自己头脑里想象的东西用语音合成系统正确地表述出来, 其所依赖的语言背景为英文的 29 个基本音素. 能否将中国人大脑里想象的东西也 "阅读出来" 呢? 汉语与英语的发音区别很大, 加工脑机制也颇为不同, 仅基本发音音素就多于 70 个. 那么, 要想构建适用于中国人思维过程的汉语神经分析系统 CNS(Chinese Neuralynx System), 需要在 Guenther 的研究基础上做些什么样的补充和修改, 或者说, CNS 本身有哪些需要关注的特殊问题, 其发展趋势、重点和难点是哪些? 这是本文要加以叙述和探讨的主要问题. 本文内容包括: (1) Neuralynx System 研究现状; (2) 国际、国内有关 CNS 的研究现状及存在问题; (3) CNS 发展趋势和展望. 通过本文的介绍, 期望从事语音生成与获取以及汉语脑机制等领域研究工作的研究者们能有所启迪和收获.

关键词 神经分析系统 汉语 音素 DIVA 模型 语音生成与获取

1 引言

1.1 简介

将人脑里的思维过程 "阅读" 出来, 然后将其转换为正常语言进行实时表述, 这样一种新型仪器已经由波士顿大学 Guenther 带领的科研小组研制成功 [1]. 这种被称为 "神经分析系统 (Neuralynx System)" 的仪器, 让使用者只需简单想一想自己所希望表达的内容, 语音合成系统就能将其直接转换成语音. 这里所说的语音合成系统, 实际上就是一种模拟人类语音生成和获取过程的具有生物学意义的机器人语音控制系统. 通过与脑 — 计算机接口 BCI (brain-computer interface) 相结合, 使用者可以直接控制声音输出, 其速度比著名科学家霍金 (Hawking) 目前正在使用的电脑发声系统快了许多. 此外, 研究人员还能利用这个系统进一步了解大脑在表述语言的过程中神经系统的工作过程和状况, 其原理如图 1 所示.

Neuralynx System 由两部分组成: 脑 — 计算机接口和语音合成系统 DIVA (directions into velocities of articulators) 模型. BCI 中, 脑电信号的产生方式是一种无线神经电极 [1,2], 用于长期植入患

引用格式: 张少白, 王勇, 何利文, 等. 汉语神经分析系统研究现状与展望. 中国科学: 信息科学, 2015, 45: 849–868, doi: 10.1360/N112014-00187

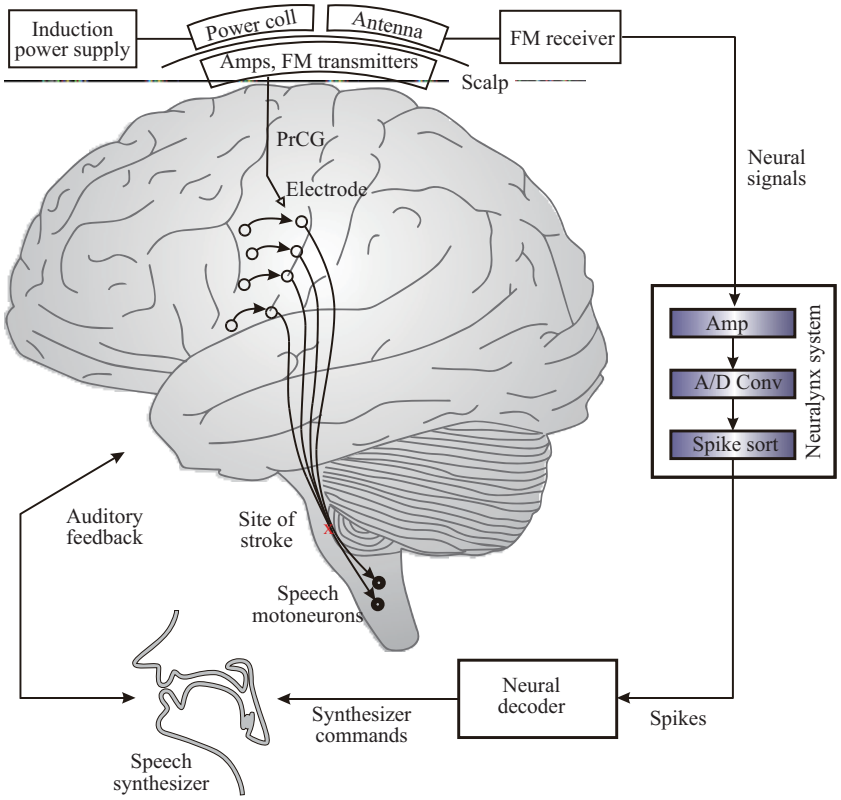


图 1 (网络版彩图) 实时语音生成合成系统脑 – 机接口原理图

Figure 1 (Color online)Schematic diagram of the brain-computer interface of the system for synthesizing and generating real-time speech

者的大脑皮层, 而检测到的神经信号则被用于驱动语音合成器的连续 “运动”, 为患者提供实时语音输出; DIVA 模型则是一种具有生物学意义的关于语音生成和获取的神经网络 [3,4].

据文献 [5,6] 报道, 近年来, Guenther 的研究在系统原有的基础上有了不少改进, 除了在实时性方面有所突破 (语音输出在 50 ms 之内, 基本满足实时性要求) 以外, 还包括在辅音处理方面也有了进展, 这比系统在研制成功时还只是停留在元音的发音和处理上前进了一大步.

但 Neuralynx System 以英语发音为基础, 其研究和处理对象是英语的 29 个基本音素 (Phoneme)[3]. 由于不同语种的发音和处理对大脑相关区域的激活程度和范围都有所不同, 脑机制区别也很大 [7], 可以说, Neuralynx System 对于以汉语为母语的人来讲适用性极差.

要想 “阅读” 汉语思维过程, 需要对汉语语音加工脑机制进行深入的理解和研究. 并在此基础上, 从相关的语音合成系统中找到汉语发音音素对应的感兴趣区域, 再根据不同的设计方案解决汉语音素串的任意组合问题.

从语言学的角度来讲, 语音可以分为音素和音节. 音节是语言的自然单位, 音素则是语音的最小单位. 语言认知研究表明, 音素与音节是两种不同性质的表征, 其加工过程表现出的脑机制具有很大的不同 [8]. 那么, 如何将这些不同在汉语思维 “阅读” 过程中恰如其分地表征出来, 这是汉语神经分析系统 CNS (Chinese Neuralynx System) 要面临的一个重要问题. 如果能依据 MNI (montreal neurological institute) 标准参照系 [9], 将 CNS 中的模型组件与大脑皮层区域有机地关联起来, 使得: (1) 神经解剖学和以言语运动控制为基础的神经过程形成统一的理论框架; (2) 可以动态分析、比较模型关于汉语

语音生成与获取过程中相关机能在神经解剖学上的临床和生理学发现, 那么 CNS 的实现就迈开了第一步.

1.2 目的和意义

人类所执行的各种感觉运动控制任务中, 语音生成和获取恐怕是最为复杂和神奇的了. 在处理单词、音节或者音素时, 除了必要的速度, 感知过程涉及不同参考系统中诸多信息之间复杂的相互作用, 包括听觉、体觉以及感觉运动参考系统等等. 一般来讲, 语音生成与获取是一个涉及人类脑组织诸多部位的复杂认知过程. 这个过程包括一种从依照句法和语义组织句子或短语的表述一直延伸到音素产生的分层结构, 需要根据发声时大脑中各种感官和运动区域的交互作用建立相应的神经网络模型.

因此, 在神经解剖学和神经心理学层次上仿真和描述大脑中涉及语音生成和理解区域的相关功能, 成为近年来人工语音合成系统所追求的主要思想. 本文的主要目的, 即围绕这样的主题思想, 在介绍国内外学者和研究机构的研究现状和发展趋势的同时, 探讨如何以 DIVA 模型为基础, 提出符合汉语语音发声规律、具有真正生物学意义的语音生成与获取神经计算模型的方法, 为进一步构造具有中国人思维特征的汉语神经分析系统 (CNS) 奠定理论和实践基础.

本文具体组织形式如下: 首先介绍 DIVA 模型; 其次, 在此基础上介绍汉语神经分析系统 CNS 的结构和方法, 以及当前国内外的研究现状; 最后, 对 CNS 的发展趋势做出展望.

2 DIVA 模型

1994 年, Guenther 首次提出了一种称为 DIVA 的神经计算模型 [3]. 这个模型主要依据有关语音生成及感知心理物理学实验的行为数据、fMRI (functional magnetic resonance imaging) 和 PET (positron emission computed tomography) 实验的神经成像数据以及对动物所做的运动控制实验的神经生理学数据等而建立, 目的是为了生成音素串而学习控制模拟声道的运动. 主要特征是反映神经解剖学与大脑相关区域的关联性. 从 1995 年到 2003 年, 不同版本的 DIVA 模型粗糙地反映了神经解剖学与大脑区域的关联性. 这些模型主要由如下几个部分组成: (1) 声道模型 (vocal tract model)[10], 该模型以 8 个发音器官 (3 个舌、3 个唇的形状, 以及颚、喉的高度) 的有关参数来定义声道的形状, 并将声道区域函数转换为用以综合声音信号的数字滤波器; (2) 发音器官的位置及方向向量; (3) 规划位置向量和规划方向向量; (4) 语音发声神经元组; (5) 变换 (映射) 学习机制; (6) 控制机构.

目前看来, 这个早期的模型还存在一些缺陷 [11], 主要表现在: (1) 模型假定所有在给定点给出的关于系统状态的信息, 对于系统而言都是瞬间可用的; (2) 系统使用瞬时反馈控制, 并假定没有神经延迟; (3) 用于控制的基准框架, 不是口腔感觉空间就是听觉空间, 不能两者同时并存; (4) 关于皮层与子皮层处理过程的分割以及大脑区域成分的关联性的描述相对粗糙.

针对这些问题, Ghosh[12] 提出了一种更接近神经解剖学的 DIVA 模型. 在这个模型中, Ghosh 引入了针对语音生成的双感官 (听觉和体觉) 基准结构. 除了建立模型神经元与实际神经解剖学之间的对应关系外, Ghosh 还解决了模型各成分之间实际传输的延迟问题. 在此期间, 其他一些学者也针对这些问题作了研究. Tourville 等 [13], Max 等 [14], Civier 等 [15], Nieto-Castanon 等 [16] 随后又提出了一系列修订版本或新的思想.

特别是 Tourville 和 Guenther[17]2011 年提出的 DIVA 模型, 对大脑皮层以及小脑中包括预运动 (premotor)、运动、听觉、体觉的几个区域所涉及的成分做了精确的定义, 并对发音期间模型对嘴唇和

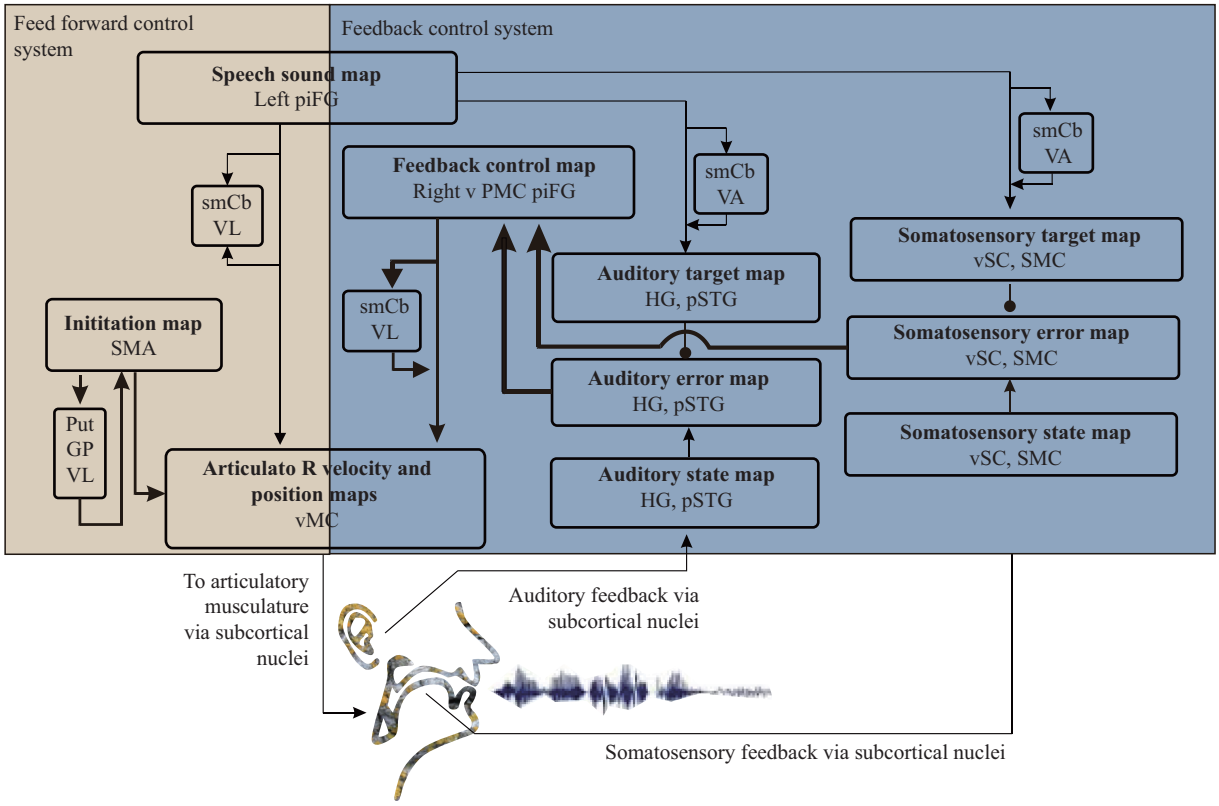


图 2　DIVA 模型示意图

Figure 2　Schematic of the DIVA model. Modified from figure 1 of Ref. [17]

下颚摄动的补偿能力进行了计算机仿真.

该版本的模型由前馈控制子系统、反馈控制子系统以及前田 (Maeda) 模拟声道所组成. 训练中, 模型通过语音的共振峰频率作为输入的同时, 产生一个发音速率以及发音器官位置变化的时变序列. 应用这个序列, 模型可以得到所需要的理想发音, 其原理如图 2 所示.

图 2 中, 每个方框代表构成某一神经表述的神经元集合. 方框中黑体字下方的文字代表集合所对应的大脑皮层区域; 箭头则表示一种神经元表述到另一种表述之间的映射 (变换), 而且这种映射被假定是某一集合中所选中的细胞活度 (activation) 通过突触映射到另一集合的过程. 每个映射集中只能有一个细胞的活度达到最大值, 模型通过这个细胞单独对输入空间的当前状态进行编码. 突触权值则是在模型两个阶段之一的咿呀学语阶段所获得的. 发音器官的随机运动提供触觉、本体感受 (proprioceptive) 以及听觉反馈信号, 并通过这些信号学习不同神经元表述之间的相互关系. 咿呀学语后的执行阶段, 模型可以快速地利用音频采样学习产生新的发音.

但是, 与 Ghosh 提出的模型一样, 尽管他们做了诸多方面的改进, 但有关时间同步以及传输延迟中感官运动的学习问题仍然没有很好地解决. 2012 年, Guenther 又对 2011 年版本的 DIVA 模型做了修正, 重点对语音生成与获取的另外一些性能因素 (例如发音速率、协同发音、定序以及模拟等) 进行了关注, 主要方法是对模型原有的小脑控制方案做了较大的修改, 基本解决了时间同步与传输延迟等问题 [6].

3 基于 DIVA 模型的汉语语音生成与获取

3.1 关于脑区域映射关系

DIVA 模型的语言基础是英文的 29 个基本音素 [3] 对于母语为中文的人来说 DIVA 模型是否也能完成其语音生成和获取的任务呢? 这个过程涉及不同语言持有者在说话时其发音过程对大脑皮层中布洛卡 (Broca) 区以及相关区域的不同影响.

文献 [18] 曾对高度流利中英文晚双语者词频相关性语义判断任务做过专门的功能磁共振脑成像 fMRI 技术研究, 探讨语言加工相关脑功能区域、词频效应及中英文对脑功能区域可能存在的差异和影响. 实验的影像学结果表明, 中文任务词频效应出现在左侧 Broca 区, 英文任务词频效应出现在双侧 Broca 区, 低频任务均不引起额外脑区的激活. 其他区域 (如前额叶背外侧 (BA9, 46)、额叶内侧面 (BA6, 32, 24)、左侧颞中回后部 (BA21)、角回 (BA39) 和右侧缘上回 (BA40) 等) 激活的程度和范围也有不同.

文献 [19] 中, Chee 等所做的实验也证明了这样的问题. 他们采用中文、英文和图片的语义判断任务对中英文高度流利早双语者进行研究发现, 中文和英文任务激活左侧前额叶皮层 (BA9, 44, 45)、左侧颞叶后部 (BA21, 22)、左侧梭状回 (BA37) 和左侧顶叶, 两者激活的脑区的空间分布相似, 但中文任务激活的脑区范围更大, 中英文两种任务激活脑区的分布、强度及范围都存在一定程度的差异.

另外, 四声是汉语语音加工中所独有的特征. 声调范畴性知觉涉及快速、自动地从高度变化的物理声学信号中提取有意义的语音信息, 使用范畴内和范畴间两种刺激类型, Xi 等 [20] 的研究表明, 在早期 MMN (mismatch negativity) 窗口, 基本感觉机制和语音功能机制能够共同影响言语听觉刺激的加工. 语音加工的激活脑区主要在左侧颞中回 (L-MTG), 而声学信号的激活脑区主要在右侧颞上回 (R-STG)[21]. 汉语语音产生时, 声调比元音加工激活更强的脑区在右脑额下回 [22].

文献 [8] 指出, 语音不仅可以从听觉通道中提取, 也可以从视觉通道中提取. 阅读中 script to sound 的过程, 由于涉及视觉字形表征与听觉语音表征的转换, 因而可以被认为是视觉通道中的语音加工. 汉语与英语者语音加工过程中脑机制的不同主要表现在: (1) 汉语者通过视觉通达语音过程中左侧颞上回不被激活; (2) 汉语与英语者左侧额叶存在功能上的差别. 导致这些差异的主要原因可能是: (1) 中文与拼音文字语音加工方式不同; (2) 不同的语言和文化背景使得大脑的可塑性产生了不同; (3) 汉语者听觉通道中负责语音表征的脑区与英语者不同 [8].

所有这些结果表明, 如果应用 DIVA 模型对中文发音人群进行语音合成, 其已有的音素 — 脑区域映射关系已经不能满足要求, 需要予以重新考虑或更正.

DIVA 模型与 fMRI 之间联系非常紧密 [23]. 由 DIVA 模型所设定的各种假设可以应用 fMRI 相应的实验来加以测试和论证; 由 fMRI 所获得的数据也可以由 DIVA 模型加以分析和解释. 这样, DIVA 模型实际上就构成了一个用以解释来自各种研究的相关数据以及对关于语音神经处理过程进行一致性描述的基本框架. 这样的框架为我们进行音素 — 脑区域映射关系的重排或更正提供了良好的基础.

3.2 关于音素建模单元集的构建

为表征新的基于汉语者的脑区域映射关系, 语音生成和获取系统中建模单元集的构建就成为一项重要任务. 这是因为建模单元集的构建是连续语音识别中声学建模需要面临的首要问题之一, 其合理与否直接影响到识别系统最终的性能.

过去的几十年里, 在中文语音识别系统中, 研究人员分别考虑用不同粒度的建模单元, 包括词

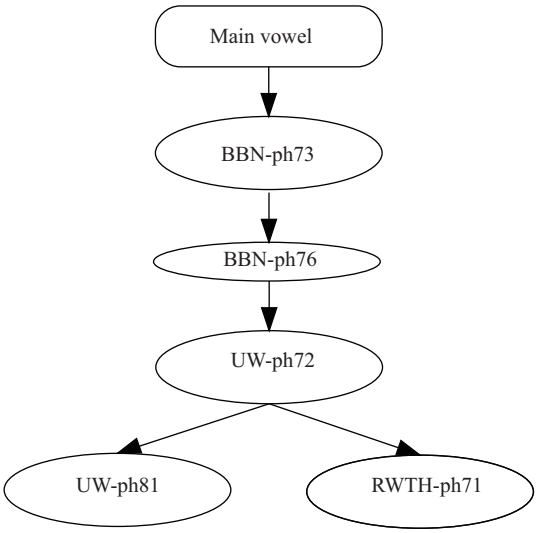


图 3　各音素建模单元集间关系示意图

Figure 3　The relations among various phoneme modeling unit sets

(word)、音节 (syllable)、声韵母 (initial/final, IF)、音素 (phoneme) 等等. 以词或者音节为粒度构建建模单元集, 往往会造成建模单元数目过于庞大, 从而出现训练数据稀疏的问题, 导致模型参数得不到充分准确的估计, 而且还会使解码的搜索空间增大, 大大降低解码效率. 因此一般只适合用于一些小词汇量的中文识别系统.

以声韵母构建建模单元集 [24], 在一定程度上反映了中文语音学的知识和特点, 并且被成功地用于搭建大词汇量连续语音识别系统, 也是目前被广泛认可的建模单元集. 但是与英文音素建模单元集相比, 声韵母建模单元集的建模单元数目还是太多, 带调的情况下更是如此.

鉴于音素建模单元集已在英文系统中被广泛应用, 并且取得了良好的性能. 长期以来, 有许多研究机构包括香港大学 (HKU)、宽带网机构 (BBN)、华盛顿大学 (UM)、亚琛大学 (RWTH) 等, 也倾向于在中文大词汇量连续语音识别系统中使用以拼音音素为粒度构建的建模单元集 [25~29]. BBN, UM, RWTH 音素建模单元集之间主要关系如图 3 所示. 其中, "ph73" 表示音素集里含有 73 个建模单元, 其他类推. 表 1 则列出了这个音素集的具体内容. 在构建新音素建模单元集 (newPS) 时, 主要参考 BBN, UW, RWTH 音素建模单元集系列的依赖和变化关系, 结合 DIVA 模型的具体情况, 构造出新的单元集来.

虽然音素没有声韵母那么清晰的中文语音学特点和背景, 但其建模单元数目却比声韵母单元集模式少了很多. 在同等训练数据量的情况下, 音素建模单元的参数能够得到更为充分和准确的估计. 结合主元音准则 (main vowel principle)[30], 音素建模单元集的性能比较声韵母建模单元集更具有优势.

考虑到系统后续模块 (如区分性训练、Tandem 特征提取等) 处理的高效和便利, 以及与 DIVA 模型的一致性, CNS 也可以考虑使用音素建模单元集的方法来替代传统的声韵母建模单元集.

3.3　关于汉语声调生成在 DIVA 模型中的应用

声调作为汉语信息处理的重要特征, 在汉语语音识别、汉语语音合成、汉语方言辨识中具有广泛的应用. 理想实验环境下的汉语孤立词声调识别, 已经取得了很好的成果 [31,32], 但噪声环境下的汉语声调识别效果还不尽如人意, 这里既有基频检测的问题, 也有声调特征的选取问题, 还有分类器设计方

表 1 BBN-ph73

Table 1 BBN-ph73

Category	Units
Consonants	b,p,m,f,d,t,n,l,g,k,h,j,q,x,z,c,s,zh,ch,sh,r
Glides	y,w,v
End syllables	W,Y,N,NG
Main vowels with tone	E(1-4),I(1-4),IH(1-4),a(1-4),e(1-4),er(1-4),i(1-4),o(1-4),u(1-4),yv(1-4)
Others	silence, fragment, breath, hesitation laugh/cough garbage

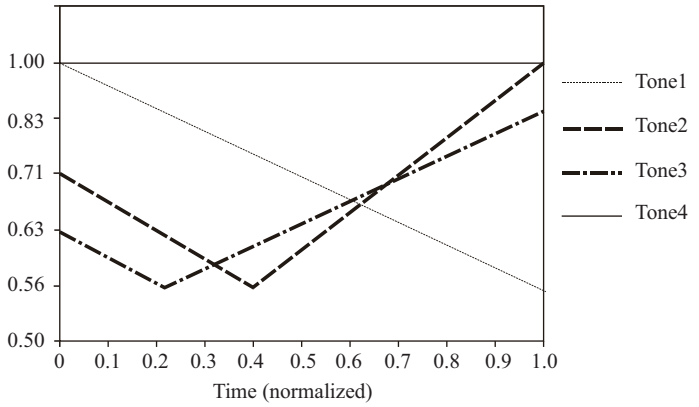


图 4 汉语普通话的基频线

Figure 4 The fundamental frequency line of Mandarin

面的问题. 由于汉语声调识别主要依赖于基频的轮廓特征, 基频检测无疑是人们关注的焦点, 很多实验和理论研究都证实了这一点 [33].

对于基频的检测, 目前已有各种容错性算法, 其中较为典型的是各种改进的自相关算法 [31,32]. 但因为噪声信号不具备相关性, 而语音信号尤其是语音的浊音信号却具有较强的相关特性. 因此, 这种自相关容错性算法仍然存在半频和倍频现象, 信噪比较低时更是如此. 如何利用线性变换提升自相关函数在噪声环境下的峰值特性, 使得基音频率的提取更为准确, 这是语音合成系统研究者们目前正在努力解决的问题.

正是在这样的理论基础指导下, 吴超民等 [34] 对现有的 DIVA 模型做了定向修改, 目的是使得改进后的 DIVA 模型能够生成汉语的 4 种声调, 并能模拟涉及声调发音时大脑相关区域的活动或反应.

吴超民采用了 Chao[35] 所建议的五度制声调符号法来表示基频和汉语声调的关系. 依据这样的方法, 用范围在 [0,1] 之间的时间轴表示汉语声调的变化, 并根据已公布的汉语标准声调曲线调整四声的升降占空比. 这样, 4 种声调就可以被具体量化 (见图 4).

仿真前, DIVA 模型中所有细胞都处于初始状态, 因而所有细胞的活度初值都被设置为零, 模型处于没有语音运动的休眠状态. 仿真过程中, 模型被转换为发音状态, 细胞变得活跃起来. 此时, 语音映射集 SSM (speech sound map, 见图 2) 中代表元音发音的细胞被激活, DIVA 模型中每一模块的细胞值都被一一计算出来, 其大小与相应脑区的激活程度成正比. 根据细胞的活动轨迹, 系统将所有细胞的最大活度值进行统一标记, 以此来补偿其在不同模块间的动态差异. 这种称之为归一化的过程可以在 MATLAB SPM 工具箱中通过应用血液动力学响应函数 (spm_hrf) 卷积的方式来实现. 最后, 仿真

结果在 DIVA 模型所定义的 MNI 标准参照系中被显示, 模块包括: 运动前区皮层 (pro_motor cortex)、听觉皮层、辅助运动区域 (SMA)、体觉皮层和小脑.

做好这些准备以后, 吴超民应用改进后的 DIVA 模型做了 3 种不同类型的仿真实验. 前两种用来验证改进后的 DIVA 模型功能, 第 3 种则用来验证改进后的 DIVA 模型是否能被用来研究汉语声调的神经关联性. 第 1 种实验主要是通过分析生成元音/a/4 种声调的语音信号过程来验证改进后的 DIVA 是否能产生声调, 用 Praat 语音分析软件显示其基频数值. 第 2 种则是应用改进后的 DIVA 模型模拟生成两种不同元音/a/和/u/所对应的大脑活动区域, 以此来验证模型是否能够维持原有的功能. 在第 3 种实验中, 吴超民将第 1 种实验所用到的语音信号用于改进后的 DIVA 模型, 以此来模拟大脑皮层相关区域的活动情况, 比较元音不同声调的发音对大脑活动区域 (范围、程度等) 的影响, 并通过与临床研究结果 [36] 的比较来验证改进后的 DIVA 模型能否用来研究汉语声调的神经关联性.

仿真的结论是: 第 1 种实验证实了改进后的 DIVA 模型根据基频可以从语音信号中清晰的识别汉语声调; 第 2 种实验表明元音/a/和/u/发音所激活的脑部区域会有差异, 但都涉及到运动前区皮层以及由 DIVA 模型所定义的运动皮层唇部区域和 MNI 空间的额下回. 也就是说, 元音/a/和/u/发音时声道形状的主要差别是唇和喉的高度. 改进后的 DIVA 模型在处理有关汉语声调的神经关联性时, 控制好模拟声道的这些高度是完全可行的; 第 3 种实验证明, 改进后的 DIVA 模型用于模拟生成元音/a/不同声调时所对应的大脑活动区域的差异与 Ladefoged[37] 的研究结果基本一致.

但是, 仿真也发现了一些问题. 主要是仿真结果和先前的神经影像学的研究数据在某些方面存在不一致. 例如, 实验仿真生成不同元音时的大脑活动区域的差别表现在双侧额下回、左额中回、颞上回、颞中回和双侧中央前回等区域, 与文献 [38] 有关神经影像学的研究结论有些不同, 这可能与语音规划的形式有关. 改进后的 DIVA 模型只能模拟语音发音的神经关联性, 不能进行语音规划. 因此, 改进后的 DIVA 模型还无法与神经影像学研究数据在语音规划方面进行很好的吻合. 这表明, 涉及语音编码和规划方面的研究仍然是 DIVA 模型在汉语元音声调识别道路上需要努力关注的问题.

另外, 吴超民只是关注了几个简单的元音声调的发音问题, 有关其他诸多汉语音素的生成与获取以及获取后的语音合成等问题, 目前看来还是一个非常艰难的系统工程.

在吴超民的研究基础上, 本文对汉语声调与英语重音在 DIVA 模型中生成和获取时所表现出的异同点进行了研究. 文献 [39] 中, 我们构建了一种能分析英语单词重音的神经计算模型. 通过利用时域基音同步叠加算法 (TD-POSAL) 对 DIVA 模型的语音韵律参数进行修改, 使得 DIVA 模型可以模拟英语单词重音的生成和获取过程, 从而解决了 DIVA 模型不能模拟英语重音的发音和对自然语言的模拟尚存在某种缺陷的问题. 通过这样的研究, 我们期望获得 DIVA 模型在处理汉语和英语这两类不同语种时本身所具有的融通率.

3.4 脑电波 — 繁体中文转换 N200 系统

2012 年 4 月, 香港特区政府宣布: 香港中文大学成功研制出 “脑电波 — 繁体中文转换 N200”——“脑 — 机接口” 系统, 该系统能将脑电波转换成繁体中文. 这种方法的基本原理与用手机键盘进行 “笔画输入” 的方法类似: 戴上无线脑电波接收器, 面向计算机屏幕上的中文笔画输入接口, 当注意力集中到不同方位时, 就能在屏幕上先后选择横、竖、撇、捺、勾 5 种笔画, 然后将中文字逐笔写出. 形象一点说, 该方法就是利用脑电波在注意力集中到不同方位时会产生的不同 “波段”, 去对应 “打” 出不同的笔画, 从而组成一个个的汉字. 但正是由于这一突出成就, 该项目成为香港特区政府信息科技总监办公室 “无障碍辅助科技研发基金” 的 9 个重点资助项目之一.

现在看来, 支撑这个方法的理论基础是香港大学心理学系教授张学新等 [40] 发现的中文特有脑电波 N200. 实验显示, 中国人在看到汉字后大概 200 ms 左右, 会产生一个特殊的脑电波, 也就是 N200 电波, 它只在中国人阅读汉字时出现, 西方人在阅读字母文字时不存在这样的现象. 由于汉字和字母文字在视觉形态上存在巨大差异, 人们一直希望弄清两者是否具有不同的大脑加工机制, 但过去 30 多年的实验研究并没有得出明确的结论. N200 的发现, 找到了区分两种文字不同加工过程的一个关键的科学指标. 这个指标清楚地表明, 汉字是视觉文字, 其识别过程很早就涉及非常深入的视觉加工; 而字母文字作为听觉文字, 不注重视觉加工, 也就不会出现 N200 这个现象. N200 还是一项强大的科学证据, 证明汉字具有超越 "象形" 的特质. 张学新等在其实验中发现, 人们在看图画或无意义字组时, N200 不会出现. 这反映出理解汉字已超越了 "象形" 的概念, 阅读者能看到图像以外的信息.

但这种方法与我们所要讨论的汉语语音生成与获取的方法区别很大, 基本不属于同一类型. 如概述所述, 我们所要讨论的方法, 需要对汉语诸多音素的发音可能激活的单个或复合脑区的脑机制进行研究和比较, 在类似 DIVA 模型的语音合成系统中 (实际要复杂的多) 找到其对应的转换区域, 然后, 对这些区域的相互作用以及映射关系加以计算和排序, 将一个一个的音素组合成任意的音素串 (音节), 最后用语音的形式表征出来. 这一过程, 尽可能精确地反映了汉语者大脑中复杂的思维过程. 比较而言, "脑电波 — 繁体中文转换 N200 系统" 只是对笔画出现的方位脑电波进行了区分, 其复杂性应该远远小于我们所要讨论的方法.

4 声道处理神经模型 (ACT 模型)

语音生成和感知是人类所具有的重要能力, 这种能力包括认知和感觉运动. 除了波斯顿大学语音实验室提出的 DIVA 模型以外, 德国亚琛工业大学语音和交流障碍学院的 Kröger 等 [41] 提出了一种语音神经处理模型, 即 ACT (vocal tract ACTions) 模型 (其结构请参考文献 [41] 的图 1).

ACT 模型包含 3 个基本的功能模块: 语音生成、语音感知和语音采集. 一方面, 该模型基于特定的运动、感觉和语音等因素来表征发音时大脑的神经解剖结构 [42,43]; 另一方面, 该模型也能够感知和获取特定语言的相关知识和技能, 并且将这些知识和技能通过突触权值的神经映射融入到模型之中 [43].

ACT 模型是在 DIVA 模型的基础上提出来的, 两者都拥有前馈和反馈回路. 从音素表征的角度来看, 两个模型都可以产生正确的发音器官运动和声学语音信号. 而从语音感知的观点来看, ACT 模型具有对分类感知进行建模的能力, 这一点已经由 Kröger 等 [43,44] 做的仿真实验所证实.

无论是 DIVA 还是 ACT 模型, 其语音生成的过程都是以音素表征作为引导的, 发音模式 (例如一个单词或一个短语) 以音节为单位. 为了获取发音过程的运动轨迹, 对于出现频率较高的音节, 两个模型都需要通过语音映射的方式激活运动状态, 使得每一次声道运动都能引发相关神经元的运动, 然后再通过模型产生语音信号.

ATC 模型的语音感知由外部听觉信号所引导. 如果以识别音素为目的, 那么模型所感知的信号就必须是出现频率很高的音节信号. Kröger 等 [45] 在语音感知的架构上采用了 Hickok 所提出的双路径方式. 虽然 Kröger 等 [11,45] 只完成了听觉区域到运动区域一种路径的转换, 但模型已经能对元音和音节进行有效辨识.

为了证实 ACT 模型的语音获取能力, Kröger 等 [41] 做了一个仿真实验. 同 DIVA 模型类似, 实验包括咿呀学语和执行模仿两个阶段. 在咿呀学语阶段, 模型将感觉与运动状态关联在一起. 因此, 模型具有在模仿训练时产生运动规划的能力. 实验使用 V 型和 CV 型音节, 包括 5 个元音 (V=/a/, /o/,

/e/, /i/, /u/) 和 3 个辅音 (C =/ B/, / D /, / G/), 元音和辅音的所有可能组合概率相等. 仿真结果表明, 在咿呀学语阶段, ACT 模型就可以产生这些音节组成的各种序列了.

综上所述, ATC 模型的主要功能是声学技术改良, 且侧重于对语音感知中央存储库中的声道动作库进行完善, 并没有考虑临床验证的问题, 因而仅局限于发音与辨音. DIVA 模型与之不同的一个重要方面在于对感觉适应的临床验证 [5,16], 但其听觉感知机制却不十分健全. 尽管存在这些差别, 但从原则上来说, 两个模型有关语音生成与获取以及定量感知等方面的功能是相互兼容的. 也就是说, ACT 模型在很大程度上能够代替 DIVA 模型.

最后要说明的是, 这两个模型都不具备汉语语音生成与获取的能力.

5 构建 CNS 需要关注的关键问题及发展趋势

如前所述, CNS 是一个涉及 BCI 以及语音合成系统的复杂系统. 其中, 语音合成系统又涉及汉语脑机制加工模型、控制机构、声道模型、转换机制以及映射关系等诸多方面的问题. DIVA 模型尽管取得了巨大成功, 但除了因为语言背景的不同, 使其不能直接应用于 CNS 以外, 其他方面也还存在着一些问题, 有些甚至是非常重要的问题. 例如, DIVA 模型没有充分考虑关于 "感知能力与语音生成技巧发育平衡" 的问题 [11], 因而其系统自组织、自适应能力受到某种程度的影响; 再例如, 对前田声道模型发音器官位置限制的控制, 由于汉语发音与英语发音存在的差异, 当原有的伪逆算法对超出发音器官位置限制的分配解不能最终获取时, 能否考虑引入基于零空间再分配的伪逆算法对原有的分配解进行线性叠加, 从而使得其结果重新处于发音器官的可行空间内? 诸如此类的问题很多, 需要认真加以考虑.

5.1 感知能力与语音生成技巧发育平衡问题

关于人类语言能力在遗传基因中所占比重到底是多少这样的问题, 一直以来是语言习得理论所涉及的一个重要问题. 从先天论和生态学的观点来看, 这个问题一方面表明了人类语言能力有些是与生俱来的, 例如婴儿能够从所获取的少量语音信息里归纳出语法或句法规则的能力 [46]; 另一方面, 也可以解释为人类语言的胜任能力是信息处理原则的结果. 也就是说, 当最初只能获得有限输入时, 系统只是学习输入信息的基本结构而不是记住每个输入的具体内容, 这样的处理原则非常重要 [47].

我们可以将婴儿学习语言的能力比喻成一个有限存储容量, 但需要处理大量信息的神经计算系统. 为了模拟信息处理过程、学习信息处理的根本原则, 我们必须要知道学习内容以及学习这些内容所需要的必要条件. 也就是说, 研究人类语音生成与获取的发展过程, 需要一个类似婴儿生理能力和生态背景的系统, 系统的学习过程应该是自组织和自适应的, 但 DIVA 模型恰恰在这个方面不够完善.

感知磁效应 (perceptual magnet effect) 现象在语音学上是一种已知的特殊现象, 也是影响婴幼儿语言发展的重要因素 [48]. 这种效应的特色是听觉感知空间会受到扭曲, 造成一个音素周围的声音都会被归成同一类. 举例来说, 在辨认英文辅音/r/与/l/时, 美国人的辨认结果在听觉感知空间中明显的分成两种不同的分类, 而日本人由于没有受到类似的训练和积累, 分辨不出两种发音的本质区别, 因而只能观察到一种分类. 事实上, 这种现象是一种心理语音学上的实验结果. 为了结合神经科学上的反应, 有研究学者利用竞争学习方式构建了能模拟感知磁效应的神经网络模型 [48].

在 DIVA 模型中, 倘若要发出英文元音的声音, 模型得通过声学空间 (共振峰 F1~F2) 来对元音进行分类. 但 DIVA 模型在这部分却存在两个问题: 第一, DIVA 模型的元音分类是根据先前文献的统

计数据用人工标记的, 并非经由后天的语音环境所形成. 显然, 这种方法并不符合人类语音获取能力的描述过程; 第二, DIVA 模型虽然可以通过人工标记的元音分类来发出声音, 但由于这些分类并非是通过感知机制所学习到的. 因而, 针对某些音素, 模型不能发出正确的声音就完全可以理解了.

其实, 这个问题与人类大脑中语音系统或者语音神经传导机制有着重要的关系. Hickok 等 [49] 曾经指出, 人类听觉与说话的动作, 好比控制系统的输入与输出. 发音器官所产生的语音共振峰, 应该如同人类所感知到的语音共振峰一样, 成为一种反馈信息; 大脑会经由这些信息去指挥发音器官产生相关的声音来. 但由于 DIVA 模型缺少语音感知机制, 是一个非完整的语音系统, 所以在 DIVA 模型绘制的声学空间中, 有些共振峰向量是人类发音器官无法产生的.

总之, 纵观 DIVA 模型对每个语音体觉目标的学习过程, 可以看到其基本假定如下: 在婴儿正确可靠地生成给定语音之前, 他能够正确可靠地感知这个语音, 并假定模型具有感知所有即将生成语音的能力. 然而, 这与婴儿咿呀学语的过程并不完全相符. 因为婴儿在感知语音时, 其过程与这些语音发生的瞬时环境密切相关, 是一种多感觉输入 (听觉、视觉、触觉等) 的融合体 [50]. 在将这些融合体信息映射到脑皮层语音区域的过程中, 逐步形成了听觉部件渐进式的反应机制. 因此, 音素表征这种形式就可以被看作是能将若干输入候选词相互区分开来, 最终形成婴儿早期语言获取过程中一种附带的衍生品的功能部件. 可惜的是, DIVA 模型没能有效地表征这种功能部件, 因而没有充分实现感知语音的自组织自适应过程.

因此从这一点来看, DIVA 模型还不完全具备神经生理学意义上的控制功能, 对其进行完善或重构使其具有充分的语音感知的自组织和自适应能力, 是非常重要且非常有意义的一项工作.

5.2 语音加工脑区映射关系的确定

要研究语音加工脑机制, 大脑皮层和小脑分割系统的定义非常重要. 例如, DIVA 模型有关听觉状态的映射对应于听觉皮层区域 (BA41, 42, 22)(参见图 5, 图中 M 表示运动皮层, P 表示预运动皮层, Au 表示听觉皮层, S 表示体觉皮层, Δ 表示误差, 箭头表示映射关系等), 这符合 Caviness 经典定义 [51]. 但 Caviness 分割机制不是专门为语音和语音障碍研究而设计的, CMA (center for morphometric analysis) 系统中多个感兴趣区域 ROI (regions of interest) 的定义不是很细致, 并不十分适合于语音神经影响的研究. 为了获取基于语音皮层相互作用的更细粒度的功能映射集, 需要将皮层语音相关区域划分成更小更细致的功能单元, 也就是说, 需要重新定义 CMA 系统.

在重新定义了感兴趣区域 (ROI) 后, 再将与汉语发音有关的新的脑区映射关系反应在新的 CNS 系统中. 这涉及系统的输入模式、控制机制、存储方式以及搜索策略等诸多方面的问题, 需要统筹考虑.

功能磁共振成像的研究结果证实了语音生成与获取过程中大脑一些区域的重要性, 这些区域相互作用从而生成语音并输出. 这些区域中的一些已经是 DIVA 模型的一部分, 例如前区运动皮层 (BA6, 44)、主运动皮层 (BA4)、体感皮层 (BA1, 2, 3)、缘上回 (BA40)、主听觉皮层 (BA41, 42)、高阶听觉皮层 (BA22) 和小脑等区域, 有些则还不是. 构建 CNS 的一项重要工作就是将剩下的一些与汉语发音有关的区域加入到模型中, 并阐明它们之间的相互作用和影响. 特别是要对辅助运动区 (supplementary motor area)、基底神经节 (basal ganglia)、脑岛 (insula) 和盖骨 (opercula) 等区域的作用和反应机制加以描述和定义.

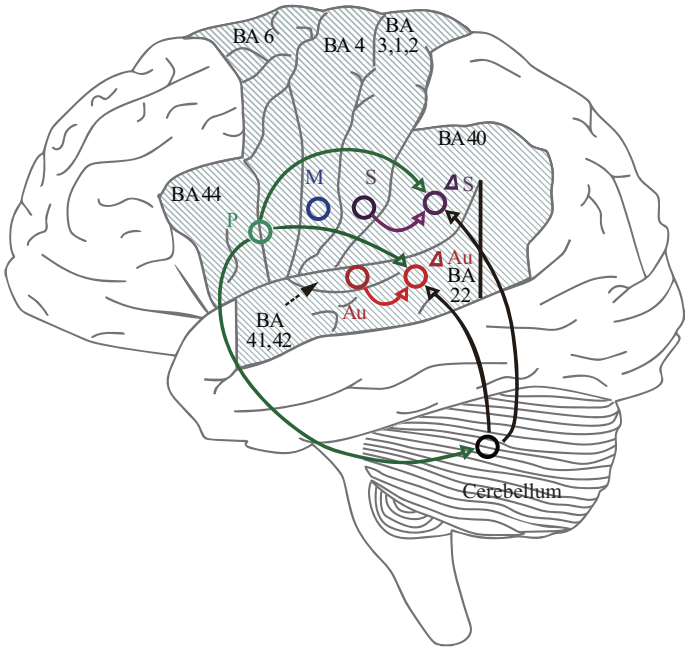


图 5　DIVA 模型对应脑区示意图之一

Figure 5　One of the corresponding brain areas schematics of DIVA model

5.3　脑电信号特征提取与分类研究

一般说来, 脑电信号获取后特征提取方式及其模式分类方法是 BCI 系统研究的重点和难点. 有效的特征提取及分类方法对提高脑电信号分类的正确率以及加快系统运行速度意义重大, 对整个系统具有重要意义.

基于脑电信号的 BCI 系统, 以实时或瞬时的方式对脑电中反映大脑不同状态的特征信号进行提取. 常用的方法有 P300 诱发电位、视觉诱发电位、皮层慢电位、自发脑电信号、植入电极法、事件相关同步化或去同步化 (ERS/ERD) 法等 6 种方式, 且一般都具有信号采集、信号分析和控制机构等 3 个功能模块. 其中, 信号分析模块利用 FFT、小波分析、Butterworth 低通滤波等各种算法, 从经过预处理的 EEG 中提取与受试者意图相关的特征量, 如诱发电位的幅值、EEG 节律或单个神经元的触发率等. 这些信号被提取后就交给分类器分类, 分类器的输出就是控制器的输入.

考虑到本申请课题拟采用通过头皮表面电极采集 EEG (electroencephalogram) 信号 (DIVA 模型采用植入电极法 [1,2]), 因而与其对应的特征提取分类算法需要专门被研究.

(1) 特征提取与分类. 脑电信号的特征提取方式一般涉及频域或时域, 方法有多种, 如空间滤波、功率谱权值、自回归模型、小波变换以及单神经元分离等. 模式识别系统对提取的特征信号进行分类以产生对外界装置进行控制的命令输出, 目前常用的技术有线性分类器、隐马尔科夫模型、独立分量分析及人工神经网络等. 可重点考虑的特征提取和模式分类方法有: (i) 功率谱与 Bayes 分类器; (ii) 多变量自回归模型与神经网络分类器 MAR (multivariate autoregressive); (iii) 自适应自回归 AAR (adaptive autoregressive) 模型与线性分类; (iv) 共同空间模式与线性分类 CSP (common spatial pattern) 等.

这些识别和分类算法各有各的特点, 主要表现在: 功率谱分析可以反映信号的能量变化, 但是常规的功率谱分析损失了时间信息, 加窗傅里叶 (Fourier) 变换中信号的时间和频率分辨率相互制约; AAR

模型参数是从记录的所有 EEG 信号中估计出来的, 因此它的重要特点就是不需要相关频带的先验信息, 适用于在线分析, 但对伪迹现象很敏感; CSP 方法可以发现和处理多电极阵列 EEG 信号中蕴含的空间信息, 从而提高识别和分类准确度, 缺点是需要大量的电极以及对 EEG 的多通道分析, 具体介绍可参考文献 [52].

需要特别指出的是, 由于受神经科学发展的限制, 目前人们对脑电特性的认识仍然存在许多悬而未决的问题, 甚至还不清楚系统是线性的还是非线性的, 因而现在要找到一种普遍适应的脑电信号识别模型几乎还不可能. 所以, 我们所要做的工作, 就是结合 CNS 系统的特点, 设计出一种相对最优的特征提取和模式分类的方法来.

(2) 脑电信号去噪方法研究. 脑电信号获取过程中, 工频噪声干扰现象往往会使所获取的信息产生多种多形态瞬时结构波形, 这种现象影响到 DIVA 模型对语音的正常处理.

目前, 脑电信号的去噪方法主要有陷波滤波器、自适应滤波器、小波变换等. 陷波滤波器方法参数与信号的采样频率直接相关, 当频率发生改变时, 其陷波滤波器参数也相应改变, 从而容易造成脑电波形失真 [53]. 自适应滤波器能够自动跟踪工频干扰的频率变化且能最大限度地减少有用信息的损失, 但是其频率跟踪范围很窄 [54]. 上述所说的方法都是在傅里叶变换的基础之上提出来的, 而应用最为广泛的小波变换也存在一定的缺陷, 例如计算过程较为复杂, 小波基的选择、小波阈值的设定都需要一定的先验知识等 [55].

Mallat 等 [56] 于 1993 年提出了基于过完备原子库 (over-complete dictionary) 的稀疏分解 (sparse decomposition) 思想. 这种方法能够根据信号自身特点, 自适应地选择合适的基函数, 从而完成信号的分解. 在此过程中, 过完备原子库起着决定性的作用. 沿着这样一种思路, 本文作者提出了一种专门适用于脑电信号结构的过完备原子库构建方法 [57], 并采用匹配追踪算法 [58] MP (matching pursuit) 对信号进行稀疏分解再重构. 经过这样一个过程, 脑电信号的稀疏性被增强, 从而达到去噪目的, 提高了 DIVA 模型语音处理的能力.

此外, 针对汉语神经分析系统研究中, 非侵入式脑机接口采集到的脑电数据存在的分辨率低、干扰大的问题, 本文作者提出了对脑电信号进行约束处理的方法[1)]. 该方法首先利用独立成分分析方法剔除原始信号中的噪声, 提取有效事件的相关电位 ERP (event-related potentials) 成分; 然后将 DIVA 模型模拟生成的功能性磁共振成像数据的激活点的空间信息作为限制条件, 对提取出的 ERP 成分进行精确定位; 最后, 通过对实验数据的处理分析, 实现了对实验中受试者的激活脑区的精确定位, 符合语音生成获取实验的生理学事实, 验证了方法的正确性和有效性.

清华大学生物医学工程系的高上凯带领的研究团队, 针对多信道 EEG 信号, 提出了一种基于稀疏时空分解的分层贝叶斯方法 [59]. 该方法能够解释多信道 EEG 数据中常见的实验间振幅的变异性, 是对多条件下记录的 EEG 数据时空分解的一种自然描述. 通过强制稀疏数据源信号, 依据该方法建立的模型使得每个估测源更有可能代表的是物理或生理过程, 而不是随机噪声, 从而大大提高了模型的去噪能力.

但是, 上述每一种方法都有其具体应用的背景和条件. 由于 CNS 所要获取的汉语语音 EEG 信号十分复杂, 究竟采用什么样的方法有效地获取符合汉语语音生成特征的脑电信号, 这样的研究恐怕还有很长的路要走, 自然这也成为了构建 CNS 所面临的重要课题.

1) 张少白, 陈彦林. 基于 DIVA 模型的脑电信号处理方法研究. 系统科学与数学. 已录用.

5.4 控制机制的完善和更新

(1) 关于扰动前馈补偿与反馈技术相结合的控制方法. 由于 CNS 映射关系的变化很大, DIVA 模型原有的控制机制当然也需要进行较大的调整, 这主要涉及反馈和前馈两个并行控制子系统 (参见图 2).

DIVA 模型的反馈控制子系统具有两个主要功能: (i) 激活与运动皮层相对应的语音映射细胞并输出听觉和体觉状态; (ii) 通过感官反馈, 将当前的听觉、体觉状态与目标值进行比较. 如果当前感官状态在目标区域以外, 系统会出现一个误差信号, 通过从感知误差到运动皮层的映射, 大脑皮层会将这些信号矫正为适当的运动命令.

DIVA 模型前馈控制子系统的主要功能表现在运动前区皮层到运动皮层的映射, 也就是对运动命令进行预编程. 理想情况下, 期望值与实际感官表征结果基本一致. 因此, 前馈控制起作用时, 反馈控制一般不再发挥作用. 但实际控制过程中, 前馈控制模型的精度会受到多种因素的限制和影响, 被控对象的特征会因此产生偏移. 因此, 考虑前馈与反馈相结合的复合控制方案, 既能发挥前馈调节及时的优点, 又能保持反馈对各种扰动因素都能得到抑制的长处, 其难点在于如何将二者恰如其分地结合起来. 什么时候什么地方采用什么样的控制方式, 这是新的系统设计中需要认真考虑的一个重要问题.

(2) 小脑控制机制的重新定位. 根据神经生理学, 人或动物的小脑除了司职运动控制 (包括身体姿态和平衡) 外, 还具有许多非运动控制的功能, 其某些区域与视觉和听觉机能紧密联系. 在解决问题、辨查错误以及语言发声等方面, 小脑的作用非常重要. 实际上, 人类的语言发声也是一种运动, 也涉及运动平衡的问题. 人或动物的运动平衡以及语言发声等方面的控制技能就是在这样的感觉运动系统自组织和自学习的过程中渐进的形成、发展和完善的. 其中, 操作条件反射 (operant conditioning) 扮演着重要角色, 是人或动物的小脑最为基本和重要的学习机制.

但直至 2006 年, Guenther 在其提出的 DIVA 模型 [16] 中, 除了对大脑皮层以及小脑中包括预运动、运动、听觉、体觉等几个区域所涉及的成分作了精确的定义, 并对发音期间模型对嘴唇和下颚摄动的补偿能力进行了计算机仿真以外, 对有关时间同步以及传输延迟中感官运动的学习问题却没有很好地解决; 有关语音生成与获取的另外一些性能因素 (例如发音速率、协同发音、定序以及模仿等), 在没有合理引入与小脑认知过程密切相关的控制机制的情况下, 模型的瞬时反应特征仍然具有很大的局限性.

针对这些问题我们进行了专门研究, 并取得了一些初步成果 [60~64]2), 但小脑认知机制的探讨应该是无止境的. 如前所述, 由于我们希望构建的 CNS 映射关系变化很大, DIVA 模型原有的控制机制需要进行较大的调整, 这其中, 小脑控制机制的重新定位就是一个非常重要的问题. 应该说明的是, Guenther 等 [3,4] 后来自己也认识到了这些问题, 并在后来的版本中对这些问题一一做了更正和改进, 效果非常好. 可以说, 在这个问题上我们与 Guenther 是殊途同归.

(3) 扰动作用下声道模型鲁棒性研究. DIVA 模型采用前田声道模型作为发音器官, 其物理限制以解剖学为基础. 在表征语音生成的数学模型中, DIVA 模型应用突触权值以及近似 Jacobi 行列式的伪逆算子来计算出与不同语音信号相对应的发音器官位置. 但该伪逆控制算法并没有直接考虑发音器官的位置限制, 这样做的结果可能会导致出现发音器官过早进入饱和状态的问题, 对声道模型的鲁棒性带来挑战. 我们认为解决办法有两种: (i) 引入基于零空间的再分配伪逆算法对原有算法进行修正. 设模型发音器官的控制效率矩阵为 $\boldsymbol{B}$, $u \in \mathbb{R}^m$ 为发音器官的变化量, 模型的状态向量为 $x \in \mathbb{R}^n$, 描述模

2) Zhang S B, Zhang Z, Zhou N N. A new control model design for the temporal coordination of arm transport and hand preshape applying to two-dimensional space. Neurocomputing, in press. DOI: 10.1016/j.neucom.2015.05.067

型的线性最小扰动方程可表示为

$$x\& = \boldsymbol{A}x + \boldsymbol{B}_u u, \tag{1}$$

$$k = \mathrm{rank}(\boldsymbol{B}_u) < m. \tag{2}$$

从矩阵分析的理论可知, (2) 式的成立表明 $\boldsymbol{B}_u$ 具有 $m-k$ 维的零空间, 且在此零空间内, 动态特性不受控制输入的扰动影响, 也就是模型的动态特性不受不同控制输入扰动影响. (ii) 借鉴经典 Hopfield 神经网络利用电子电路求解非线性方程或超越方程, 其求解过程无需计算自动完成的方法, 通过研究网络中参数的约束条件对神经元组所起的作用, 研究参数扰动对声道模型的影响规律.

6 大数据分析环境下个体化 CNS 到统计意义上 CNS 映射的实现

一般说来, 脑电信号获取后特征提取方式及其模式分类方法是 BCI 系统研究的重点和难点. 有效的特征提取对提高脑电信号分类的正确率以及加快系统运行速度意义重大. 除此以外, 个体化的语音生成与获取系统模型能否全面有效地反映群体脑机制加工的特征和规律? 这是 DIVA 模型所没有考虑的问题.

从脑科学的角度来看, 目前的物联网、云计算以及大数据处理等技术所组成的并行网络正在构成与人类大脑高度相似的组织结构 — 互联网虚拟大脑. 经由物联网底层各种传感系统所获取的可能为汉语神经分析系统 CNS 所需要的虚拟声音、视觉、触觉、运动感知数据以及 EEG 信号等信息, 在互联网中枢神经系统中经过云端存储、大数据分析、信号加工、图像处理等过程, 最后模拟复原为 CNS 所必需的输入信号. 因此, 如何充分利用这样的虚拟大脑开展与认知科学相关且符合脑科学研究机理的研究, 代表了汉语神经分析系统 CNS 研究的一种全新方向.

脑电信号由于其数据的规模性、瞬时性以及多样性等性质决定了其信号特征提取与分类过程的复杂性. 针对 CNS, 由于语言习惯、教育程度以及思维方式的不同, 在不同 “脑个体” 对象上所获取的脑电数据或者图像信息存在差异是必然现象. 因此, 在建立了具有显著个体化特征的 CNS 以后, 如何将来自不同 “脑个体” 的脑电数据和图像信息推广到具有统计意义的 CNS 上去, 为探索可能具有的共同特征或趋势提供研究基础, 是一件非常有意义的工作. 这些工作主要包括: (1) 脑区映射关系的可视化及图像处理机制的建立; (2) 不同统计意义上 “脑个体” 之间的语音加工脑区映射关系的差异分析; (3) 统计意义上 “脑个体” 数据的存储、管理和分析, 以及汉语语音加工脑机制数据挖掘和处理.

7 需求分析与应用前景

著名天体物理学家霍金因患肌肉萎缩症丧失了说话能力, 只能利用眼球的移动发出指令, 通过电脑语言合成器, 发出独具特色的 “电脑声”. 据霍金助理称, 之前霍金通过其电脑系统, 每分钟能讲出 15 个英语单词, 而如今因脸部肌肉恶性萎缩, 每分钟只能讲出一个英语单词, 速度比正常交流慢很多.

如果应用 Guenther 研究成果来帮助霍金与外界交流, 其交流质量和速度都会得到极大的改善. 如果我们将 Guenther 的研究成果进一步改造, 使之符合中国人的 “思维过程” 和发音规律, 会使患有与霍金类似疾病或者因其它原因造成失音的汉语者人群, 能重新 “开口讲话”, 并能正常与他人进行交流, 其应用前景十分广阔.

Guenther 的研究获得成功, 包括实时性方面所取得的突破 (语音输出在 50 ms 之内, 基本满足实时性要求), 这对于霍金来说无疑是个福音. 如果应用 Guenther 的语音合成系统, 上述提到的准确性以

及交流速度问题都会得到解决. 我们将 Guenther 的研究成果加以改造, 使之符合中国人的 “思维过程” 和发音规律, 最终结果就是使患有与霍金类似疾病或者因其他原因造成失音的汉语者人群, 能重新 “开口讲话”, 并能正常与他人进行交流, 其应用前景是十分广阔的.

除此以外, 通过对汉语神经分析系统的开发和研究, 科研人员可以利用 CNS 这个平台进一步了解大脑在表述语言的过程中神经系统的工作过程和状况, 因而会对其他相关学科和领域的研究带来启迪和推动. 这正如 DNA 测序方法的出现, 大幅度提高了公安系统的破案能力一样.

实际上综合国内外研究现状可知, 语音发声所属的运动控制和运动平衡问题与神经生理学特别是脑机制以及小脑机能的研究之间的联系涉及到一个越来越为人们所重视的交叉学科领域, 也就是所谓的神经机器人学 (Neurorobotics) 领域.

神经机器人学的研究目标是模拟和复制生物的神经生理结构和神经生理机能将其应用于机器, 特别是自主机器人系统, 使机器或机器人表现出类似人或动物的行为.

神经机器人学的研究, 既具有重要的科学和理论研究意义, 又具有重要的工程和技术应用价值, 是一个前景广阔的研究领域, 我们在这个领域里将研究与 “中国特色” 关联起来, 更是具有十分重要的科学意义.

8 总结与展望

在神经解剖学和神经心理学层次上仿真和描述大脑中涉及语音生成和理解区域的相关功能, 特别是仿真和描述符合汉语思维过程的相关功能, 是控制科学、机器人学以及神经生理学相互融合的一个非常前沿的研究课题. 本文首先对 “神经分析系统” 的研究现状进行了概述, 接着对相关的研究成果进行了分类和评析. 总体来讲, 现有的 “神经分析系统” 很多都是在 DIVA 模型的基础上发展起来的. DIVA 模型是一种用于语音生成和获取后描述相关处理过程的数学模型, 也是一种为了生成单词、音节或者音素, 被用于控制模拟声道运动的自适应网络模型.

但 DIVA 模型是以英语发音的 29 个基本音素作为研究背景的, 由于不同语种发音过程中音素所对应的脑区激活程度和范围都有可能不同, DIVA 模型中音素 — 脑区的对应关系以及处理方式也会随之不同, 汉语语音生成与获取的相关问题也就需要予以专门解决. 针对这些差异, 本文详细分析了 DIVA 模型在处理汉语语言时可能遇到的问题, 并提出了可能的解决方法和途径.

至于 CNS 的发展方向, 我们认为是以 DIVA 模型为基础, 在大数据分析环境下, 构造符合汉语语音发声规律且具有真正生物学意义的神经计算模型, 从而为进一步构造具有中国人思维特征的汉语神经分析系统 CNS 奠定理论和实践.

参考文献

1 Guenther F H, Brumberg J S, Wright E J, et al. A wireless brain-machine interface for real-time speech synthesis. PLoS one, 2009, 4: e8218

2 Brumberg J S, Nieto-Castanon A, Kennedy P R, et al. Brain-computer interfaces for speech communication. Speech Commun, 2010, 52: 367–379

3 Guenther F H. A neural network model of speech acquisition and motor equivalent speech production. Biol Cyber, 1994, 72: 43–53

4 Guenther F H, Ghosh S S, Nieto-Castanon A, et al. A neural model of speech production. In: Proceedings of the 6th International Seminar on Speech Production, Sydney, 2003. 85–90

5 Tourville J A, Guenther F H. The DIVA model: a neural theory of speech acquisition and production. Lang Cogn Process, 2011, 26: 952–981

6 Guenther F H, Vladusich T. A neural theory of speech acquisition and production. J Neurolinguist, 2012, 25: 408–422
7 Qi Z Q, Peng D L. Brain mechanism research of speech processing: current situation, puzzles, and prospect. J Beijing Norm Univ Soc Sci, 2010, 220: 40–47 [祁志强, 彭聃龄. 语音加工的脑机制研究: 现状、困惑及展望. 北京师范大学学报 (社会科学版), 2010, 220: 40–47]
8 Qi Z Q, Peng D L. The brain mechanism of Chinese speakers—a comparative study of phoneme and syllable processing. Dissertation for Ph.D. Degree. Beijing: Beijing Normal University, 2007 [祁志强, 彭聃龄. 汉语者语音加工的脑机制 —— 音素与音节加工的比较研究. 博士论文. 北京: 北京师范大学, 2007]
9 Mazziotta J, Toga A, Evans A, et al. A four-dimensional probabilistic atlas of the human brain. J Am Med Inform Assoc, 2001, 8: 401–430
10 Sondhi M M. Models of speech production for speech analysis and synthesis. J Acoust Soc, 1990, 87: 14–24
11 Kröger B J, Birkholz P, Lowit A, et al. Phonemic, Sensory, and Motor Representations in an Action-Based Neurocomputational Model of Speech Production. Speech Motor Control. Oxford: Oxford University Press, 2010. 23–36
12 Ghosh S S. Understanding cortical contributions to speech production through modeling and functional imaging. Dissertation for Ph.D. Degree. Boston: Boston University, 2005
13 Tourville J A, Reilly K J, Guenther F H. Neural mechanisms underlying auditory feedback control of speech. NeuroImage, 2008, 39: 1429–1443
14 Max L, Guenther F H, Ghosh S S. Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: a theoretical model of stuttering. Contemp Issues Commun Sci Disord, 2004, 31: 105–122
15 Civier O, Guenther F H. Simulations of feedback and feedforward control in stuttering. In: Proceedings of the 7th Oxford Dysfluency Conference, Oxford, 2005. 1–7
16 Nieto-Castanon A, Guenther F H, Perkell J S, et al. A modeling investigation of articulatory variability and acoustic stability during American English/r/ production. J Acoust Soc Am, 2005, 117: 3196–3212
17 Tourville J A, Guenther F H. The DIVA model: a neural theory of speech acquisition and production. Lang Cogn Process, 2011, 26: 952–981
18 Zhao J, Guo J, Zhou F, et al. Time course of Chinese monosyllabic spoken word recognition: evidence from ERP analysis. Neuropsychologia, 2011, 49: 1761–1770
19 Chee M W L, Hon N H H, Caplan D, et al. Frequency of concrete words modulates prefrontal activation during semantic judgments. Neuroimage, 2002, 16: 259–263
20 Xi J, Zhang L, Shu H, et al. Categorical perception of lexical tones in Chinese revealed by mismatch negativity. Neurosci, 2010, 170: 223–231
21 Zhang L, Shu H, Zhou F, et al. Common and distinct neural substrates for the perception of speech rhythm and intonation. Hum Brain Mapp, 2010, 31: 1106–1116
22 Liu Y, Shu H, Wei J. Spoken word recognition in context: evidence from Chinese ERP analyses. Brain Lang, 2006, 96: 37–48
23 Bohland J W, Guenther F H. An fMRI investigation of syllable sequence production. NeuroImage, 2006, 32: 821–841
24 Katz G, Giesbrecht E. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In: Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Stroudsburg, 2006. 12–19
25 Church K W, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguist, 1990, 16: 22–29
26 Schutze H. Automatic word sense discrimination. Comput Linguist, 1998, 24: 97–123
27 Zhou Q. Base chunk scheme for the Chinese language. J Chin Inform Process, 2007, 21: 21–27 [周强. 汉语基本块描述体系. 中文信息学报, 2007, 21: 21–27]
28 Yu H, Zhou Q. Intra-chunk relationship analyse for Chinese base chunk recognition systems. J Tsinghua Univ Sci Tech 2009, 49: 136–140 [宇航, 周强. 汉语基本块的内部关系分析. 清华大学学报 (自然科学版), 2009, 49: 136–140]
29 Attia M, Toral A, Tounsi L, et al. Automatic extraction of Arabic multiword expression. In: Proceedings of Multiword Expressions: from Theory to Applications (MWE), Beijing, 2010. 18–26
30 Cruys T V, Moirón B V. Lexico-semantic multiword expression extraction. In: Proceedings of Computational Linguistics, Netherlands, 2007. 175–190

31 Yang W, Lee J, Chang Y, et al. Hidden markov model for mandarin lexical tone recognition. IEEE Trans ASSP, 1988, 36: 988–992

32 Guan C T, Chen Y B. Speaker-independent tone recognition for Chinese speech. Acta Acust, 1993, 18: 379–385 [关存太, 陈永彬. 非特定人四声识别. 声学学报, 1993, 18: 379–385]

33 Laures J, Weismer G. The effects of flattened fundamental frequency on intelligibility at the sentence level. Speech Lang Hear Res, 1999, 42: 1148–1156

34 Wu C M, Wang T W. Study of neural correlates of mandarin tonal production with neural network model. J Med Biol Eng, 2011, 32: 169–174

35 Chao Y R. A Grammar of Spoken Chinese. Beijing: The Commercial Press, 2011. 203–211

36 Liu L, Peng G L, Ding G S, et al. Dissociation in the neural basis underlying Chinese tone and vowel production. Neuroimage, 2005, 29: 515–523

37 Ladefoged P. Some physiological parameters in speech. Lang Speech, 1993, 6: 109–110

38 Denise K, Robert J Z, Brenda M, et al. A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. NeruoImage, 2001, 13: 646–653

39 Zhang S B, Ji Y C. Research on the pronunciation mechanism of syllable accent in English based on DIVA model. Neurocomput, 2015, 152: 11–18

40 Zhang X X, Fang Z, Du Y C, et al. The centro-parietal N200: an event-related potential component specific to Chinese visual word recognition. Chin Sci Bull, 2012, 51: 1–16 [张学新, 方卓, 杜英春, 等. 顶中区 N200: 一个中文视觉词汇识别特有的脑电反应. 科学通报, 2012, 51: 1–16]

41 Kröger B J, Kannampuzha J, Eckers C, et al. The neurophonetic model of speech processing ACT: structure, knowledge acquisition, and function modes. Cogn Behav Syst, 2012, 7403: 398–404

42 Kröger B J, Kopp S. A model for production, perception, and acquisition of actions in face-to-face communication. Cogn Process, 2010, 11: 187–205

43 Kröger B J, Kannampuzha J, Neuschaefer-Rube C. Towards a neurocomputational model of speech production and perception. Speech Commun, 2009, 51: 793–809

44 Kröger B J, Birkholz P, Kannampuzha J, et al. Categorical perception of consonants and vowels: evidence from a neurophonetic model of speech production and perception. In: Proceedings of 3rd COST International Training School, Caserta, 2011. 354–361

45 Kröger B J, Birkholz P, Kannampuzha J, et al. Towards the Acquisition of a Sensorimotor Vocal Tract Action Repository Within a Neural Model of Speech Processing. Communication and Enactment 2010. Berlin: Springer, 2011. 287–293

46 Traunmüller H. Conventional, biological, and environmental factors in speech communication: a modulation theory. Phonetica, 1994, 51: 170–183

47 Lacerda F, Klintfors E, Gustavsson L. Multisensory information as an improvement for communication systems efficiency. In: Proceedings of Fonetik, Gothenburg, 2005. 83–86

48 Kuhl P K. Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. Percept Psychophys, 1991, 50: 93–107

49 Hickok G, Poeppel D. The cortical organization of speech processing. Nat Rev Neurosci, 2007, 8: 393–402

50 Kuhl P K, Williams K A, Lacerda F, et al. Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 1992, 255: 606–608

51 Caviness V S, Meyer J, Makris N, et al. MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. J Cogn Neurosci, 1996, 8: 566–587

52 Hao D M. Reserch on EEG Classification and EEG Models for Brain-Computer Interface. Dissertation for Ph.D. Degree. Beijing: Beijing University of Technology, 2005 [郝冬梅. "脑 — 计算机" 系统中脑电信号分类与脑电信号模型研究. 博士论文. 北京: 北京工业大学, 2005]

53 Du X Y, Li Y J, Zhu Y S, et al. Removal of artifacts from EEG signal. J Biomed Eng, 2008, 25: 464–471 [杜晓燕, 李颖洁, 朱贻盛, 等. 脑电信号伪迹去除的研究进展. 生物医学工程学杂志, 2008, 25: 464–471]

54 Poornachandra S. Wavelet-based denoising using subband dependent threshold for ECG signals. Digit Signal Process, 2008, 18: 49–55

55 Chen R X, Tang B P, Lü Z L. De-noising method based on correlation coefficient for EEMD rotor vibration signal. J

Vib Meas Diagn, 2012, 32: 542–546 [陈仁乡, 汤宝平, 吕中亮. 基于相关系数的 EEMD 转子震动信号降噪方法. 震动、测试与诊断, 2012, 32: 542–546]
56 Mallat S, Zhang Z. Matching pursuit with time-frequency dictionariesdictionaries. IEEE Trans Signal Process, 1993, 41: 3397–3415
57 Zhang S B, Wang Y. Research on the Method of EEG Signal De-noising Based on the DIVA Model. Acta Electron Sin, 2015, 43: 700–707 [张少白, 王勇. 基于 DIVA 模型的脑电信号去噪方法研究. 电子学报, 2015, 43: 700–707]
58 Chen S, Donoho D, Saunders M. Atomic decomposition by basis pursuit. SIAM J Sci Comput, 1999, 20: 33–61
59 Wu W, Chen Z, Gao S K, et al. A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG. NeuroImage, 2011, 56: 1929–1945
60 Zhang S B, Ruan X G, Cheng X F. A new constructing method of cerebellum model applying to DIVA model. In: Proceedings of Chinese Control and Decision Conference, Guilin, 2009. 954–959 [张少白, 阮晓钢, 成谢锋. 一种新的适用于 DIVA 模型的小脑模型构建方法. 中国控制与决策会议, 桂林, 2009. 954–959]
61 Zhang S B, Ruan X G, Cheng X F. A new cerebellar control scheme and simulation based on Kalman Estimator. Chin J Electron, 2009, 18: 297–301
62 Zhang S B, Zhou N N, Feng Z Q. Cerebellar control model design for the temporal coordination of arm transport and hand preshape. J Pattern Intell, 2012, 2: 8–14
63 Zhang S B, Cheng W Q, Cheng X F. An application of cerebellar control model for prehension movements. Neural Comput Appl, 2014, 24: 1059–1066
64 Zhang S B, Zhou N N. Development of general cerebellar cognitive module used for robot motor. J Nanjing Univ Post Telecommun Nat Sci, 2012, 32: 69–74 [张少白, 周宁宁. 用于机器人运动控制的通用小脑认知模块的构建. 南京邮电大学学报 (自然科学版), 2012, 32: 69–74]

Research status and prospect of Chinese Neuralynx System

ZHANG ShaoBai*, WANG Yong, HE LiWen & CHENG XieFeng

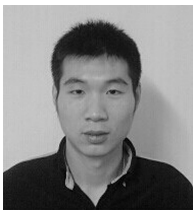
College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046*, China*
*E-mail: adzsb@163.com

Abstract Simulating and describing the functions of brain regions involved in speech acquisition and production based on neurophysiology and neuroanatomy is an important research topic in artificial speech synthesis systems. To facilitate this research, a team led by Professor Frank Guenther at the Boston University Speech Lab developed the Neuralynx System, an instrument that allows users to accurately express their ideas using a speech synthesis system based on 29 basic English phonemes. However, without modifications, it cannot decipher the thought processes of a Chinese speaker. The manners of articulation and the processing methods of brain mechanisms differ greatly between Chinese and English speakers. For example, the Chinese language has more than 70 basic phonemes. Therefore, the design established by Professor Guenther and his team will need to be supplemented and modified in order to construct a Chinese Neuralynx System (CNS) suitable for Chinese thought processes. To achieve this goal, we analyze the development trends, important features, and difficult points related to CNS. These are the main issues described and discussed in this article. The main content of this article is as follows: (1) the Neuralynx System's research status; (2) the international and domestic research status related to CNS, and its existing problems; and (3) CNS development trends. Through the information in this article, we expect that researchers will gain inspiration and learn from other researchers involved in the areas of speech acquisition and production, and Chinese brain mechanisms.

Keywords neuralynx system, Chinese, phonemes, DIVA model, speech acquisition and production

ZHANG ShaoBai was born in 1953. He graduated from Beijing University of Aeronautics & Astronautics in 1977. Afterwards he received the master's degree from Huazhong University of Science and Technology and doctor's degree from Beijing University of Technology. Now he is Professor of Nanjing University of Posts and Telecommunications. His current research fields include artificial intelligence and pattern recognition, and intelligent information processing. He has published more than 100 papers and takes charge of many projects supported by the National Natural Science Foundation of China.

HE LiWen graduated from the University of Sheffield, UK, with a Ph.D. in computer science. In 1999, he joined BT labs as a research engineer; he subsequently became a senior researcher. In 2005, he worked as Principle researcher for BT Group CTO Office. In 2009, he attained the position of Security CTO for Huawei Technologies. In 2012, he joined Nanjing University of Posts and Telecommunications as a Distinguished Professor and director, to establish an information engineering and cloud computing research institute. His major research interests are network/information security, cloud computing/big data technologies and applications, and intelligent information processing.



WANG Yong was born in 1990. He is currently a student in the Computer Department at Nanjing University of Posts and Telecommunications. His research interests include pattern recognition and intelligent systems.