DIVA模型中运动感觉系统传输延迟问题的研究

高丽琴，张少白
（南京邮电大学 计算机学院，江苏 南京 210003）

摘 要：语音生成与获取的控制问题，是机器人发声系统急需解决的问题。早期的DIVA（Directions Into Velocities of Articulators）模型并不完全具备神经生理学意义上的控制功能。为使机器人模拟声道发出类似人类语言的声音，解决好运动感觉系统中感官反馈时间延迟的问题，文中通过采用前馈控制机制补充反馈控制的策略，来处理语音运动控制方面的延迟问题，并且最后针对模型神经网络学习过程中延迟参数变化引起的网络状态扰动问题，进行了颚扰动补偿仿真实验。实验结果表明，其给出的方法行之有效且鲁棒性良好。

关键词：DIVA模型；反馈控制；前馈控制；扰动；神经传输延迟
中图分类号：TP31 **文献标识码**：A **文章编号**：1673-629X(2012)03-0117-04

Study of DIVA Model Movement Sensory Systems Latency Problems

GAO Li-qin, ZHANG Shao-bai
(Computer College, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)

Abstract: The problem of the speech acquisition and production control needs to be solved for the robot voice system. Earlier versions of the DIVA(Directions Into Velocities of Articulators) model does not have the nerve physiology control function. To make the robot simulate the human sound and solve the problem of the feedback realistic delays in the sensory system, it adopts the strategy that feedforward control mechanism supplements feedback control mechanism. Aiming at the network state disturbance caused by parameters changes during the neural network learning process, it does related emulation experiments. Experimental results show that this method is effective and gives good robustness.

Key words: DIVA model; feedback control; feedforward control; perturbation; neural transmission delays

0 引 言

DIVA模型是一种关于语音生成与获取后描述相关处理过程的数学模型，同时也是一种为了生成单词、音节或者音素，被用来控制模拟声道运动的自适应网络模型[1]。自Guenther1994年首次提出DIVA模型以来[2]，涌现了不少新的版本。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。该模型相对简单且被广泛应用于对各种语音现象的解释，包括等价运动、协同发音、速度/距离关系、说话速率效应和语音技能的掌握[3]。但是目前该模型还并不完全具备神经生理学意义上的控制功能，尽管做了各方面的改进，但是有关运动感觉系统中感官时间延迟的问题还是没有得到很好的处理[4]。目前研究普遍认为，语音生成与获取的神经系统主要是一个皮层到皮层的网络。以大脑皮层中反馈处理机制为基础的控制系统主要是利用一次行动的感官期望值与实际感官反馈之间的误差，但基于反馈的控制对于声道的迅速运动反映显然太慢，即一个纯粹以反馈为基础的控制系统存在着以上提及的延迟问题并会产生不稳定的语音速率。因此，文中采用反馈和前馈控制子系统一体化方法来处理语音控制运动中的传输延迟，同时在正常扰动情况下对颚进行计算机仿真实验。

收稿日期：2011-05-15；**修回日期**：2011-08-21
基金项目：国家自然科学基金(61073115)
作者简介：高丽琴(1986-)，女，山西晋中人，硕士研究生，研究方向为模式识别与智能系统；张少白，硕士研究生导师，研究方向为人工智能与认知科学，信息获取、处理与识别等。

1 DIVA模型的运动感觉系统传输延迟及实时处理方法

1.1 传输延迟

早些时候DIVA模型版本是假设瞬间传送神经信

号的，即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，并且系统使用的是瞬时反馈控制，即假定没有神经传输延迟。因此，新版本 DIVA 模型中，纳入了神经实时延迟的处理。然而，神经系统必须应付潜在的不稳定延误以控制发音器官的运动。例如，初级运动皮层的运动命令通常需要 40ms 或更长时间才能将它作用于相关语音器官控制运动，顶部沟到感觉刺激延迟大约为 10ms，而一些体觉皮层细胞还存在大约 50 ms 的更长延迟。同样地，来自发音器官和耳蜗的感知信息到达初级感觉皮层延迟将达到 10ms。大多数成年人对单词“dilapidated”的发音需要时间不到 1s，但这个单词需要 10 次音素的转换，每次转换需要大约 100ms 才能完成。

另外，如果系统受到外部扰动的影响，例如主体的听觉反馈出现实时扭曲，那么所听到的即为错误的声音，此时系统就会激活听觉误差细胞并试图纠正扰动。但是，由于神经传输的延迟发生在以上提及的肌肉激活和由此产生的行为活动中，那么系统对这些纠正命令相对突然发生的扰动会延迟大约 75 ~ 150ms。由此可见，为了减少传输延迟带来的影响，必须控制语音发音器官的运动速度。

1.2 语音实时处理方法

由于延迟时间的变化可能引起控制效果的恶化，甚至使系统变得不稳定，所以过程控制界一直关注延迟系统问题的研究，并提出许多控制策略。例如，自适应大延迟对象控制策略，但它们都假定延迟时间为已知[5]；基于线性系统的时间延迟辨别方法，但计算复杂难以应用到实际过程中去[6]；基于时间延迟反馈控制混沌的方法，该方法的难点是参数的确定，如延迟时间等[7]；基于混沌优化方法确定出延迟反馈控制的延迟时间，从而得到延迟反馈控制方法的参数，进而解决延迟反馈控制中参数难以选取的问题[8]；基于神经网络辨识与基于模型补偿的延迟系统控制策略相结合，此方法可以解决模型误差问题，应用于具有变化参数或者不确定性延迟时间的大延迟系统的控制。

可见，神经网络补偿控制的最大优点是它能克服时间延迟的变化，这也是过程控制很重要的要求。由于一个纯粹以反馈为基础的控制系统存在着传输延迟问题，进而产生不稳定的语音速率。因此，为了切实解决问题，本模型采用了神经网络补偿控制，即语音生成系统采用了前馈控制机制补充反馈控制的策略来处理实际语音运动控制方面的延迟变动。

2 前馈控制与反馈控制

2.1 反馈控制子系统

DIVA 模型中的反馈控制子系统具有以下功能。第一，激活与运动皮层相对应的语音映射细胞并且输出已经学习到声音的听觉和体觉状态。第二，通过感官反馈，将当前的听觉、体觉状态与目标的高层听觉和感觉皮层进行比较。如果当前的感官状态在目标区域以外，高层感觉皮层中将会出现一个误差信号，然后通过从感知误差到运动皮层的投射的学习，大脑皮层细胞将这些误差信号矫正为适当的运动命令。

接下来将详细描述上述过程，包括听觉、体觉状态的映射，语音的听觉和体觉误差映射以及将感觉误差转化为正确的运动行为。

2.1.1 听觉状态映射

模型中，听觉状态映射对应于听觉皮层区域（ BA 41，42，22）。这些细胞的活动由以下等式表示：

$$\mathrm{Acoust}(t) = f_{ArAc}(\mathrm{Artic}(t)) \quad (1)$$

（1）式中，Acoust（ t ）表示由当前发音器所形成的听觉信号；f_{ArAc} 函数是由发音合成软件转换得到的；Artic（t）表示语音映射细胞的运动位置。

$$Au(t) = f_{AcAu}(\mathrm{Acoust}(t - \tau_{AcAu})) \quad (2)$$

（2）式中，$Au(t)$ 表示语音听觉状态映射细胞活动的矢量，f_{AcAu} 函数是将听觉信号转变为相对应的皮层区域的反应，τ_{AcAu} 表示通过耳蜗转换的声音信号传到听觉皮层区域需要的时间。Schroeder 和 Foxe（2002）[9]指出从听觉刺激开始到高阶听觉皮层 A1 区及脑回沟 STP 区反应的时间延迟大约为 10ms。基于这些数据，在以下仿真试验中会用到一个估计量 τ_{AcAu} =20ms。

对于 f_{AcAu}，在模型中用到各种不同的听觉特性，包括共振峰频率、日志共振峰率和基于小波变换的声音信号转变。这些不同的听觉的空间模拟已经得到了类似的结果。在如下计算机仿真结果报告中，使用了共振峰频率特性三维的向量 $Au(t)$，即其元件符合第 3 共振峰频率声波的信号。

2.1.2 体觉状态映射

该模型也包括一个体觉状态映射，它对应于发音器的体觉皮层区域（ BA 1，2，3，40，43）。这些细胞的活动由以下等式表示：

$$S(t) = f_{Ars}(\mathrm{Artic}(t - \tau_{Ars})) \quad (3)$$

（3）式中，$S(t)$ 表示体觉状态映射细胞活动的 22 维矢量，f_{Ars} 函数是将发音器的当前状态转化为相应的体觉皮层区域的反应，τ_{Ars} 表示由神经末梢到高阶体觉皮层区的体觉反馈时间。报告指出延迟大约为（ 5 ~ 20） ms，其中一些体觉皮层细胞还存在近似 50 ms 的更长延迟。Schroeder and Foxe （2002）[10] 指出顶部沟到感觉刺激延迟大约为 10 ms （手部神经的电子刺激）。基于这些结果，在以下仿真实验中用了一个估计量 τ_{Ars} =15ms。

此外,f_{Ars}是将发音状态转换为22维的体觉映射表示$S(t)$。其中前16个维度的$S(t)$是与发音矫正器中本体反馈的当前位置相对应的,每一维度是由对应体对的细胞表示的。换句话说,这个部分决定了前16个维度的$S(t)$。而其余的6维对应触觉反馈,包括腭和唇触觉。

2.1.3　听觉和体觉误差映射

首先,将当前声音目标区域的感觉与模型的高阶感觉皮层感觉信息作比较。如果当前感觉状态存在目标区域外的误差,那么将这些误差信号映射为正确的运动指令。其次,DIVA模型的语音误差映射将对语音的听觉目标与目前听觉状态之间的差别进行编码。这些听觉误差映射细胞的活动由以下等式表示:

$$\Delta Au(t) = Au(t) - P(t - \tau_{pAu}) z_{pAu}(t) \quad (4)$$

(4)式中,τ_{pAu}表示运动前区到听觉皮层的信号传输延迟(假设为3ms),$z_{pAu}(t)$表示对声音的听觉期望进行编码的突触权值。在语音生成时,如果说话者的听觉反馈偏离了语音的听觉目标区域,那么听觉误差细胞将会被激活。

(3)式表示的运动前区皮层的投射抑制了听觉误差映射细胞。Houde, Nagarajan, Sekihara, and Merzenich (2002)[11]指出由MEG产生的听觉反应小于本体生成语音的听觉反应,并且在相同的环境条件下产生的噪音刺激反应是相同的,但研究还没有发现在语音生成时,缘上回能够起到抑制作用。

该模型的体觉误差映射是对语音的体觉目标和当前目标状态的误差进行编码的:

$$\Delta S(t) = S(t) - P(t - \tau_{PS}) z_{PS}(t) \quad (5)$$

(5)式中,τ_{PS}表示从运动前区皮层到体觉皮层的传输延迟(仿真中为3ms),$z_{PS}(t)$权值表示声音的体觉期望,$P(t - \tau_{PS})$表示运动前区皮层的状态。当说话者声带的体觉反馈偏离声音生成的体觉目标区时,体觉误差细胞将被激活。

2.1.4　将感觉误差转化为正确的运动行为

在模型中,通过感觉皮层区投射到运动皮层的学习途径,将听觉或体觉误差映射细胞产生的误差映射为正确的运动指令。这些反馈控制信号的投射由以下等式表示:

$$\dot{M}_{Feedback}(t) = \Delta Au(t - \tau_{AuM}) Z_{AuM} + \Delta S(t - \tau_{SM}) Z_{SM} \quad (6)$$

(6)式中,Z_{AuM}和Z_{SM}是突触权值,将定向感觉误差转化为正确的运动速度;其中τ_{AuM}和τ_{SM}是皮质皮层的传输延迟(3ms)。在数学语言上,用权值和一个近似雅克比行列式的伪逆算子得出相应感觉状态的发音器官位置(M)。即在早期含糊不清的说话阶段,通过检测移动指令来调整这些权值。这些突触权值的有效实施,被称为"伪逆模型"。

该模型预测的听觉和体觉误差,将会通过基于反馈的控制机制所纠正,如果一直遇到这些误差,这些纠正最终会编码到前馈控制器。这种情况如听觉扰动应用(如一个或更多实时移位共振峰频率)或体扰动应用(如颚扰动)。Tremblay, Shiller, 和 Ostry (2003)[12]做了一个简单音节生成颚运动实验,通过对颚施力,其生成音节的结果不会受到影响。尽管受试者的下巴受到作用力,其声音也没有发生变化,这就表明他们使用了如DIVA模型中的$Z_{PS}(t)$权值。

2.2　前馈控制子系统

根据模型,小脑补充了运动前区到初级运动皮层的投射,这些投射出现于音节生成时前馈命令的学习过程中,从而形成了前馈运动指令。这些前馈命令假设既可以直接也可以经小脑由腹外侧的运动前区皮层投射到初级运动皮层。同时,高层听觉和体觉皮层区可以为选择运动命令提供重要的状态信息。

模型中,在一个婴儿没有任何的发音练习前,前馈控制信号运动命令的作用是相当小的,因为它还没有被调整过。因此,在最初新的声音产生时,主要控制方式是基于反馈的控制。在早期的语音生成时,前馈控制系统通过检测由反馈控制系统产生的运动命令来进行自我调整。随着时间的推移,前馈控制会变得越来越好,几乎消除了对基于反馈控制的需要,除非发音器受到外部约束或是听觉反馈被人为干扰。因此随着语音器的扩充,其基于反馈控制的系统可以纳入前馈控制命令。这也就允许尽管在使用过程中存在尺寸和形状动态的变化,前馈控制器也能被持续正确地调整。

通过以下等式生成前馈运动控制命令:

$$\dot{M}_{Feedforward}(t) = P(t) Z_{PM}(t) - M(t) \quad (7)$$

(7)式中,$Z_{PM}(t)$权值是对语音生成的前馈命令进行编码。这个命令是从反馈命令控制子系统尝试加入新的前馈命令相结合的过程中学习到的。

如上所述,当模型学习到一个适当的前馈控制指令序列时,这个序列将会生成语音。另外,在语音生成时如果在发音器或者听觉信号没有出现意外干扰的时候,由于没有感觉误差,反馈子系统将不会受到约束,此时反馈子系统则会发挥较大的作用。

2.3　前馈控制子系统与反馈控制子系统结合

前馈与反馈复合控制系统既能发挥前馈调节控制及时的优点,又能保持反馈控制对各种扰动因素都有抑制作用的长处。根据神经生理学,人类的语言发声是一种运动,其涉及到运动平衡的问题。而人或动物的小脑具有运动控制和非运动控制方面的功能,利用小脑提供的被很好同步的前馈命令来补充基于反馈的皮层控制命令是一种很好的解决方案。因此,利用小

脑机制，语音生成系统采用前馈控制机制补充反馈控制的策略，即采用反馈和前馈控制子系统一体化方法，处理语音运动控制中的传输延迟。

3 仿真实例

针对模型神经网络学习过程中延迟参数变化引起的网络状态扰动问题，模型通过前馈控制对扰动信号进行补偿，以下是应用 MATLAB 和 C++进行颚扰动补偿的仿真实验。

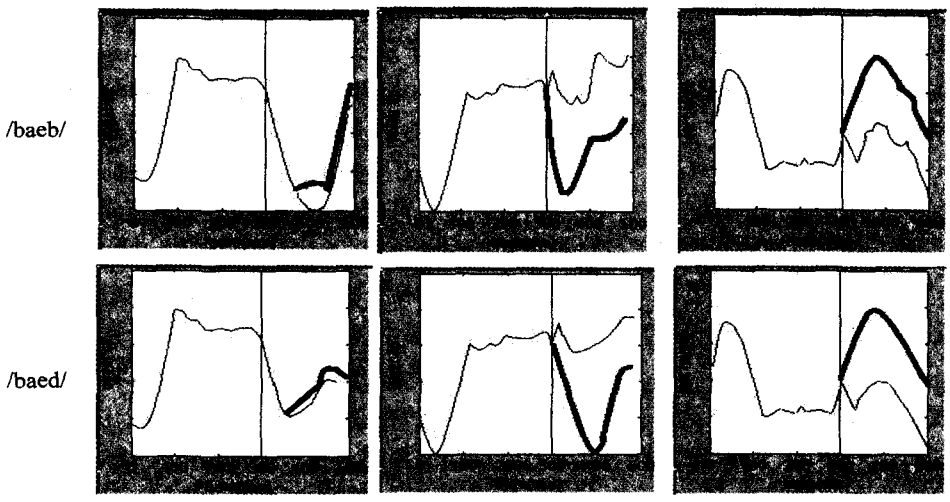

图 1 DIVA 模型仿真

图 1 为 DIVA 模型仿真结果，通过对/baeb/和/baed/两个音节的生成证明了颚扰动的影响。即在颚向上移动试验中，扰动随着负荷的增加也会变得稳定。另外在声带模型的发音器仿真实验时给下颚高度一个常量，可以看出扰动一直作用于整个发声期间。垂线表示扰动的开始，并且扰动是由颚移动的位置和速度决定的。粗线表示在正常情况下的实验发音器官的位置（未扰动），细线表示在扰动情况下发音器官的位置。如在实验中，当生成双唇音/baeb/扰动时，上唇通过进一步向下移动起补偿作用，齿槽音/baed/扰动则没有。

4 结束语

文中采用前馈补充反馈的策略，研究了 DIVA 模型中神经传输延迟的问题，从而使有关感官反馈中时间延迟的同步问题以及各种神经系统中引起传输延迟的传导速率问题得到某种程度的改进。仿真结果表明，文中运用的方法提高了控制系统的鲁棒性。

参考文献：

[1] Zhang Shaobai, Ruan Xiaogang, Cheng Xiefeng. A new constructing method of cerebellum model applying to DIVA model [C]//*Control and Decision Conference*, CCDC 2009. Guilin, Chinese: [s. n.], 2009.

[2] Guenther F H. A neural network model of speech acquisition and motor equivalent speech production[J]. Biological Cybernetics, 1994, 72(1): 43-53.

[3] Guenther F H, Ghosh S S, Niet-o-Castanon A. A neural model of speech production[C]//Proceedings of the 6th International Seminar on Speech Production. Sydney, Australia: [s. n.], 2003: 85-90.

[4] Maeda S. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model [M]// Speech production and speech modeling. Boston: Kluwer Academic Publisher, 1990: 131-149.

[5] 程 莹，刘文波. 基于自适应遗传算法的细胞神经网络模板设计[J]. 计算机技术与发展，2008，18(5)：54-56.

[6] 邱勤伟，焦贤发，朱良燕. 对称二值噪声下线性系统的随机共振[J]. 计算机技术与发展，2010，20(8)：239-242.

[7] 张晓明，黄德云，彭建华. 延迟反馈法控制混沌的解析研究[J]. 深圳大学学报(理工版)，2004(1)：43-48.

[8] 丁华福，宋宇航，唐远新，等. 小波混沌神经网络的研究与应用[J]. 计算机技术与发展，2011，21(8)：93-96.

[9] Guenther F H, Ghosh S S, Tourville J A. Neural modeling and imaging of the cortical interactions underlying syllable production[J]. Brain and Language, 2006, 96(3): 280-301.

[10] Houde J F, Nagarajan S S, Sekihara K, et al. Modulation of the auditory cortex during speech: an MEG study[J]. Journal of Cognitive Neuroscience, 2002, 14(8): 1125-1138.

[11] Houde J F, Jordan M I. Sensorimotor adaptation in speech production[J]. Science, 1998, 279(5354): 1213-1216.

[12] Tremblay S, Shiller D M, Ostry D J. Somatosensory basis of speech production[J]. Nature, 2003, 423(6942): 866-869.