

文章编号: 1003-0077(2009)06-0086-05

基于 RFC 模型的基频曲线导数域编码方法研究

王磊, 刘加

(清华信息科学与技术国家实验室(筹) 清华大学电子工程系, 北京 100084)

摘要: 基频是发浊音时声带振动频率, 通常用 F_0 表示。在一个音节或连续的语音段中, F_0 是随时间变化的, 这种变化的轨迹形成了基频曲线。基频曲线的走势可以反映出语句的重音、语调等韵律信息, 所以对基频曲线的描述和研究就显得尤为重要。该文首先提出了一种基频曲线描述方法, 即导数域编码方法, 同时探讨了该编码方法在语音发音质量评价中对韵律的作用。实验结果表明基于该描述方法能够提高英语发音语调质量评价的性能, 主观和客观评价的相关性由原来的基于基音极值差的 0.38 提高到 0.49。

关键词: 人工智能; 模式识别; 基音频率; 导数; 编码; 应用

中图分类号: TN912.34 **文献标识码:** A

The Derivative Domain Codes of Pitch Curve and Applications

WANG Lei, LIU Jia

(Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering,
Tsinghua University, Beijing 100084, China)

Abstract: Fundamental frequency (or pitch), usually named as F_0 , is the vibration frequency of vocal cord during the production of voiced sounds. In a syllable or continuous voice paragraph, F_0 changes with time and yields the fundamental frequency (or pitch) curves. It is particularly important to descript and investigate the F_0 curve because it usually reflects the rhythm information, such as tone and stress. This paper first proposes a new method to describe F_0 curve—the derivative domain codes, and then it discusses the role of the coding method on the rhythm in the evaluation of speech pronunciation. Experimental results show that the method can be used to evaluate the English prosody. The correlation coefficient between the subjective and objective scores of pitch extreme difference improves from 0.38 to 0.49.

Keyword: artificial intelligence; pattern recognition; pitch; derivative; codes; application

1 引言

基音是指发浊音时声带振动引起的周期性, 而基音频率 F_0 是浊音声带振动的频率。在语音信号处理应用中, 如语音压缩编码、语音识别、语音分析与合成的韵律预测^[1]、语音发音质量评价等, 都需要获取基音频率信息。特别是在语音发音的音韵研究中, 基音与重音、语调等密切相关。因此, 在发音评

价中从音韵学角度评价时, 基音起到了重要的作用。利用基音进行研究, 首先就要对基音变化过程进行很好的描述。而基音频率 F_0 的大小与性别、年龄等有关, 甚至与发音时人的情绪和生理状况有关, 例如老年男性基频偏低, 小孩和青年女性偏高^[2]。在韵律分析研究中, 相对基频绝对高低, 一般来讲更关心基频曲线的走势, 因为基频曲线的走势通常反映了语句的韵律信息。因此, 如何对基频曲线的走势进行描述就显得尤为重要。目前比较成熟的描述基

收稿日期: 2009-02-03 定稿日期: 2009-04-21
基金资助: 国家自然科学基金委员会与微软亚洲研究院联合资助项目(60776800); 国家 863 高技术研究发展计划资助项目(2006AA010101, 2007AA04Z223, 2008AA02Z414)
作者简介: 王磊(1982—), 男, 硕士生, 研究方向为英语发音评价算法的研究和实现; 刘加(1954—), 男, 教授, 研究方向为语音识别、语音合成、语音编码、多媒体数字通信等。

频曲线的方法,例如 Fujisaki 模型^[3],虽然能够很好地拟合基频曲线、描述基频曲线的走势,但是涉及大量的参数,需要进行复杂的计算。本文根据语音合成中用来描述语调的 RFC 模型^[4],利用导数具有反映瞬时变化率的特点,提出采用导数域编码的方法描述基频曲线的走势过程,实验表明该方法具有简单、高效、实用的特点。

2 基频曲线的导数域编码

在语音合成中,有一种用来描述英语语调的模型,RFC(Rise/Fall/Connection)模型^[5],利用一连串 R(上升)、F(下降)和 C(连接)元素来描述语调。通过下面的模型公式,可以合成基频轨迹。

$$\begin{aligned} F_0 &= A - 2 \times A \times \left(\frac{t}{D}\right)^\gamma \quad 0 < t < \frac{D}{2} \\ F_0 &= 2 \times A \times \left(\frac{1-t}{D}\right)^\gamma \quad \frac{D}{2} < t < D \end{aligned} \quad (1)$$

其中 F_0 表示基音频率, A 表示 RFC 元素的幅度, t 表示时间, D 表示 RFC 元素的段长, γ 是 RFC 元素的曲率。图 1 是用该模型描述的基频曲线轨迹。

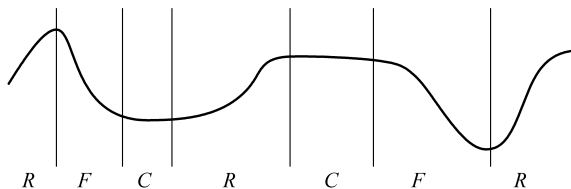


图 1 基于 RFC 模型的曲线升降描述示意图

RFC 模型是对曲线升降的描述,反映曲线的变化情况,而在数学中,导数的意义就是瞬时变化率,也就是函数在某一点上的变化率^[6]。所以可以在导数域中对基频曲线的升降情况进行描述。首先将基音曲线变换到导数域,然后设定门限值进行编码。与 RFC 模型相比,该方法非常简单,而且能很好的描述基频曲线的走势。如图 2 所示。

在图 2 中最上面的是原基频曲线,中间的是求导后的曲线(即导数域曲线)。在导数域中取一个较小的值 σ 作为阈值,导数值大于 $+\sigma$ 的部分标记为 1,导数值小于 $-\sigma$ 的部分标记为 -1,导数值介于 $+\sigma$ 和 $-\sigma$ 之间的标记为 0。得出的编码序列为 (0,1,1,1,1,0,0,0,1,1,1,-1,-1,-1,-1,-1,0,1,1,1,1,0,0,0,0)。然后在得到的编码序列中,将连续的 1 进行合并,得到 1,连续的一 1 进行合并得到 -1,连续的 0 进行合并,得到 (0,1,0,

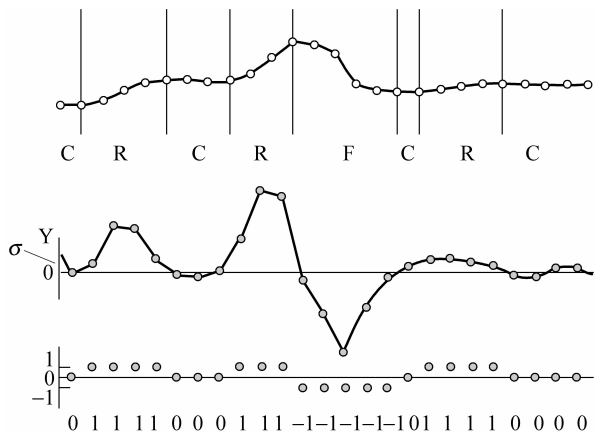


图 2 基于 RFC 模型的描述曲线升降的导数域编码示意图

1, -1, 0, 1, 0)。与 RFC 模型比照,此处 1 代表 R, -1 代表 F, 0 代表 C,则该编码与 RFC 模型完全吻合。

从以上的分析中可以发现导数域模型能够很好的描述原基频轨迹的升降过程。同时由于提取出的基频是离散的点,求导时可用相邻点的幅度差(频率差)得出,其计算量很小。只要采用 $(+1, 0, -1)$ 三种符号就可以对基频轨迹变换过程进行大致描述,如果设定更多的阈值和采用更多的符号(如 $+2, -2, +3, -3$ 等),可以更加详细地进行描述,应用时可根据实际需要进行调整。

3 基频曲线的导数域编码在语音发音质量评价中的应用

3.1 基音轨迹的拟合

现有的基音轨迹拟合方法中,Fujisaki 模型^[3]是一种成熟并且广泛应用的模型,该模型是由日本的藤崎教授以人喉部的生理和机械运动机理为依据,推导出的基频轮廓模型,公式如下:

$$\begin{aligned} \ln F_0(t) &= \ln Fb + \sum_{i=1}^I A p_i G p(t - T_{oi}) \\ &+ \sum_{j=1}^J A a_j [Ga(t - T_{1j}) - Ga(t - T_{2j})] \end{aligned} \quad (2)$$

$$Gp(t) = \begin{cases} 0 & t < 0 \\ \alpha^2 t \exp(-\alpha t) & t \geq 0 \end{cases} \quad (3)$$

$$Ga(t) = \begin{cases} 0 & t < 0 \\ \min[1 - (1 + \beta t) \exp(-\beta t), \gamma] & t \geq 0 \end{cases} \quad (4)$$

可以看到该模型涉及很多参数,在拟合时需要

用到最小二乘法,不断地进行参数修正,计算量较大^[7]。考虑到导数域编码能很好的描述基频曲线的走势,反映了前后基音值的升降情况,利用此升降信息,可进行基频曲线的还原,如下面公式所示:

$$F_0(i+1) \approx F_0(i) + D[i] \tag{5}$$

其中 $F_0(i)$ 是第 i 点的基频值, $D[i]$ 是第 i 点的导数域编码值,二者之和近似等于第 $i+1$ 点的基频

值,如图 3 所示。在拟合时,首先确定第一个基音值,然后根据导数域编码序列的值,1 为升,0 为平,−1 为降,依次连接下去,即可拟合出一条与原基频曲线近似的曲线。如表 1 所示。与 Fujisaki 模型相比,利用导数域编码得到的序列值可以简便且快速地进行基音轨迹的拟合。

表 1 导数域编码和 Fujisaki 模型在基频曲线拟合上的比较





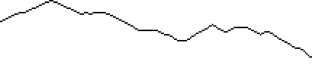







	导数域编码拟合	Fujisaki 模型拟合
原基频曲线和拟合得到的基频曲线		
原基频曲线		
拟合得到的基频曲线		
拟合所用时间	小于 1ms	15ms

表 1 中左列是用导数域 1、0、−1 编码进行的拟合,右列是用 Fujisaki 模型进行的拟合。比较二者的拟合基频曲线,可以发现,Fujisaki 模型的拟合基频曲线十分的平滑,只是描述了大的轮廓,而导数域编码的拟合基频曲线更好地反映了原基频曲线内部的细节走势。对于此条基频曲线的拟合,Fujisaki 模型需要 15ms 的时间,而导数域编码只需要不到 1ms,远远快于前者。这是由于 Fujisaki 模型涉及很多参数,需要大量的计算,而导数域编码方法简单,提取出的基频是离散的点,求导时用相邻点的频

率值相减并和阈值比较即可得出编码,所以计算量很小。

但是由于升降都使用了统一的编码,只是简单的描述基频曲线的走势,将前后基频值相比为升的都描述为 1,降的都描述为 −1,没有反映出升降值的多样性,如果想更加细致的描述基频曲线,可以设置更多的阈值,采用更多的编码,例如 2,−2 等,如表 2 所示。使用 0、1、−1、2、−2 编码更加细致的描述了基频曲线,而所用时间仍然小于 1ms。

表 2 采用更多编码的基频曲线拟合的比较

	0、1、−1、2、−2 编码	0、1、−1 编码
原基频曲线和拟合得到的基频曲线		
原基频曲线		
拟合得到的基频曲线		
拟合所用时间	<1ms	<1ms

考虑到基音的前后连续性,一般不应设置太高的阈值和太大的编码。当出现野点时,由于编码值的有限性,能消除野点的影响。如图 3 所示。

图 3 中上面的轨迹中两个圈的位置是两个野点,下面的轨迹是利用导数域编码拟合的轨迹,消除了这两个野点的影响。

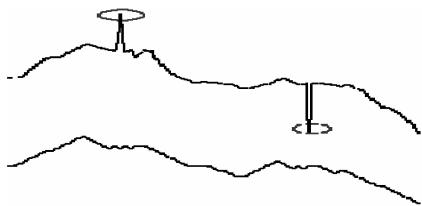


图3 轨迹中出现的野点示意图

3.2 基于导数域编码的语调的发音客观评价分数

对于像英语这样的非声调语言,从基频的变化可以确定语调高低升降变化的不同^[8]。在语调客观评价时,本文与参考文献[9]中采用的方法进行了对比,在参考文献[9]使用了平均基音差和基音极值差两种方法,对发音中的韵律进行评价。本文使用导数域编码进行建模,从基频曲线走势或语调上进行客观评价。模型及算法流程如下图所示:

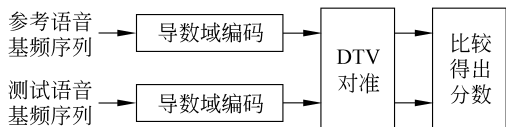


图4 基于导数域编码的语调客观评分流程

首先,将基音序列进行导数域编码,得到反映参考语音和测试语音语调走势的编码序列。此处的参考语音和测试语音基音序列,是经过前期语音强制对准和基音提取后的基音值序列。

由于得到的参考语音和测试语音的编码序列不一定等长,所以不能直接进行比较,此处采用 DTW 的方法,进行长度对准,然后比较得出分数。

3.2.1 导数域编码

设基频值序列为 $P = \{p(1), p(2), \dots, p(n)\}$, 其中 $p(i)$ 为第 i 个基频值, n 为总的基频值个数。

第一步: 计算 $k(i) = p(i+1) - p(i), i = 1, 2, \dots, n-1$ 得到 $K = \{k(1), k(2), \dots, k(n-1)\}$ 。

第二步: 设定域值 H 。

第三步: 量化各导数值 $K = \{k(1), k(2), \dots, k(n-1)\}$, 得到相应编码序列 $M = \{m(1), m(2), \dots, m(n-1)\}$ 。当 $-H < k(i) < +H$ 则 $m(i) = 0$; 当 $+H < k(i)$ 则 $m(i) = 1$; 当 $k(i) < -H$ 则 $m(i) = -1$ 。

3.2.2 DTW 对准

对测试语音和参考语音相应的基音轨迹分别进行编码后,得到编码序列 $M_{test} = \{m_t(1), m_t(2), \dots,$

$m_t(L)\}$, $M_{ref} = \{m_r(1), m_r(2), \dots, m_r(J)\}$ 。 L 和 J 分别为二者的编码个数,一般情况下不相同,可使用 DTW 的方法进行编码对准,使 M_{test} 和 M_{ref} 都为 N 个编码, $N = \max(L, J)$, 在 DTW 时只需把 $m_t(i)$ 和 $m_r(i)$ 看作是语音特征矢量即可,只不过 $m_t(i)$ 和 $m_r(i)$ 是一维的。得到对准后的编码序列 $M_{test_D} = \{m_{t_d}(1), m_{t_d}(2), \dots, m_{t_d}(N)\}$ 和 $M_{ref_D} = \{m_{r_d}(1), m_{r_d}(2), \dots, m_{r_d}(N)\}$ 。如下图所示:

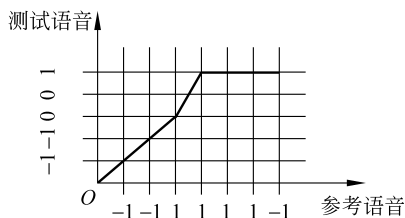


图5 导数域编码的 DTW 对准

当然,在 DTW 过程中, M_{ref} 和 M_{test} 的插入编码和删除编码也反映了二者的差距,故插入编码和删除编码的总数应该予以记录,然后影射到失真分数中。

3.2.3 客观评价分数

对 $M_{test_D} = \{m_{t_d}(1), m_{t_d}(2), \dots, m_{t_d}(N)\}$ 和 $M_{ref_D} = \{m_{r_d}(1), m_{r_d}(2), \dots, m_{r_d}(N)\}$ 进行比较,得出失真分数:

$$S = \sum_{i=1}^N (|m_{t_d}(i) - m_{r_d}(i)|) + D \quad (6)$$

其中, $m_{r_d}(i)$ 是 M_{ref_D} 序列中第 i 个元素值, $m_{t_d}(i)$ 是 M_{test_D} 序列中第 i 个元素值, D 是 DTW 对准时插入和删除的编码总数。 S 是失真分数,反映了测试语音和参考语音基音曲线相差的程度, S 越大,说明测试语音越不好。

3.2.4 实验结果

实验中采用本实验室采集的 ESC 库(专家评分语音库),该库在实验室环境下采集,周围没有明显背景噪声。语音数据的采样率为 16 千赫,采样精度为 16 比特。

专家分数(主观分数)采用 MOS 主观评价分数对采集语音的发音质量进行评价^[10],即以平均意见分来衡量语音质量,用五个等级来表示语音的质量等级:优(5 分)、良(4 分)、一般(3 分)、差(2 分)、坏(1 分)。

将机器分数(客观分数)与专家分数(主观分数)进行相关性计算,相关系数越高则客观分数与主观分数越接近,该分数模型的性能越好。用平均基音差模型、基音极值差模型和导数域模型分别计算客

观分数,然后与主观分数比较,得出各相关系数如下表所示:

表 3 语调评分实验结果

模 型		主客观相关系数
平均基音差	句子级	0.13
	单词级	0.28
基音极值差	单词基音极值差	0.20
	元音基音极值差	0.38
导数域编码		0.49

其中元音基音极值差,是利用基音极值差的方法,只对语音的元音部分进行计算。从实验结果可以看出,采用导数域编码的相关性结果相对前面的两种方法有了明显的提高。由于平均基音差或者基音极值差的方法只是对一段语音的基音在宏观进行了比较,而导数域编码的方法描述了一段语音内部的语调走势即基频曲线走势,故得出的主客观相关系数高于前面两种方法。目前基于 DTW 对准时只是针对一维特征,对准的准确性不如多维的准确,故此处的算法还有待改进。

4 结 论

本文提出的导数域编码方法能较好的描述基频走势,拟合基频曲线变换过程,并消除野点的影响。如果采用更多的阈值和编码,则能够更细致的描述基频曲线。在语调发音客观评价中,使用导数域编码的方法比平均基音差和基音极值差的方法与主观评价更接近,其主客观相关性提高了 11 个百分点。

该编码可广泛应用于许多语音相关领域,如发音评价、语音通信等。与原始的基音值相比,使用导数域编码数据量小,例如仅使用 1、0、-1 这三个编码值就可以拟合基频走势。该方法可以进一步推广,在语音韵律分析、语音合成等领域中得到应用。

参考文献:

[1] 陈高鹏,胡郁,王仁华. 考虑语速和前后环境的基频 Target 模型及实现[J]. 中文信息学报,2004,18(3): 81-85.

[2] 韩纪庆,张磊,郑铁然. 语音信号处理[M]. 北京: 清华大学出版社,2004.

[3] H. Fujisaki, S. Ohno, O. Tom Ita. Automatic Parameter Extraction of Fundamental Frequency Contours of Speech Based on a Generative Model[J]. Proceedings of ICSP96,1996,1: 729-732.

[4] Paul A. Taylor. The Rise/Fall/Connection Model of Intonation[J]. Speech Communication,1995,15: 169-186.

[5] 朱芸. 计算机辅助英语学习系统中的韵律分析与建模方法研究[D]. 北京: 清华大学,2004.

[6] 高等数学(第五版)[M]. 同济大学应用数学系,北京: 高等教育出版社,2005.

[7] 王文剑,王长富,戴蓓倩,等. 基于藤崎模型的汉语语音基频轮廓的参数提取[J]. 小型微型计算机系统,1999, 20(10): 756-759.

[8] 覃福森. 英语音高与英语语调关系研究[J]. 学术问题研究(综合版),2007,(1): 76-82.

[9] 李超雷. 交互式语言学习系统中的发音质量客观评价方法研究[D]. 北京: 中国科学院电子学研究所,2007.

[10] IEEE. IEEE recommended practice for speech quality measurements [J]. IEEE Trans. on Audio and Electroacoust Sep. 1969: 227-246.