基于 MMSE 准则的基频模型

刘浩杰 杜利民
(中国科学院声学研究所语音交互技术研究中心 北京 100080)

摘 要: 在声调与语调相互作用理论的基础上，该文利用最小均方误差准则有效地提取了连续语流基频曲线的高音线及低音线，从量化的角度证实了高音线及低音线对连续语流基频曲线的作用及其区别。该文还对声调与语调相互作用的数学模型做了初步探讨，建立了基频曲线的双线调节及调中值模型，为合成系统基频灵活有效地调整提供了新的手段，提高了语音合成系统的自然度。

关键词：语音信号处理， 基频模型， 高音线， 低音线， 最小均方误差准则
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2005)12-1932-05

Study on Fundamental Frequency Model Based on MMSE Principle

Liu Hao-Jie Du Li-min
(*SITR, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract Based on the interaction of tone and intonation, a new method is proposed to get the top line and the bottom line for continuous speech of mandarin, using the Minimum Mean Square Error (MMSE) principle. The research results demonstrate that the top line and the bottom line have different function to the intonation of continuous speech on the part of quantum. The paper also proposes a primary mathematical model to figure the interaction between the tone and intonation. On the basis of the extraction of two lines and the primary model, the two-lines model and the median fundamental frequency model are established separately, which provides a flexible and effective method to modify the intonation and improves the naturalness of speech synthesis system of mandarin greatly.

Key words Speech signal processing, Fundamental frequency model, Top line, Bottom line, MMSE principle

1 引言

为了得到更高自然度的合成语音，基频曲线无疑起很重要的作用。人类自然言语的基频曲线融入了语法、语音、情感及发音实体等多方面的因素，因而发出的语音自然流畅。要使电脑发出同样高质量的合成语音，同样要使基频曲线融入更多的信息[1]。但我们对这多种综合因素相互影响作用的机理了解的不是很清楚，限制了其在合成系统中的应用。

对不同的语音单元，基频曲线有不同的表现形式[2]。通常对单音节，称为声调，对连续语流，称为语调。就当前的研究而言，着重探讨的是声调与语调的关系及相互作用机制[3-6]。在对人类发音器官的生理机制研究的基础上，研究人员提出了很多的解释言语基频现象的理论[2,7]。赵元任的“大波浪”与“小波浪”说以及音域的概念[3]是对声调及连续语流语调研究的奠基学说，它首次阐明了连续语流基频曲线的本质。一些重要的基频生成模型都是建立在该理论的基础之上[2,7]。沈炯[4]主要从“大波浪”的角度对语流的基频曲线进行了研究，对音域在连续语流语调中的作用机制做了较深入的探讨。王安红等[6]、王蓓等等[7]从大规模语料库中统计的结果都直接或间接地证实了音域在基频曲线中的作用机制。从以上学者的研究结果，我们可以得到连续语流基频曲线的一般认识，即音域的变化是连续语流中基频曲线变化的主要控制因素，高音线主要与重音的强度相关，低音线主要与韵律单元节奏的完整性有关；单音节的基频主要与音节本身的特殊性质有关，是连续语流基频变化的基础；连续语流中基频有下倾趋势，但由于受到重音或感情色彩的影响而发生局部的跳跃性变化，这些跳跃性变化虽然干扰了我们对基频曲线的研究，但也为我们更具有丰富表现力的合成系统的开发提供了研究的基础。

2004-07-24 收到, 2005-05-10 改回

由以上的认识，可以得到对语调研究的两个需待解决的问题：(1)连续语流中高、低音线的求取方法；(2)基于高音线、低音线的基频调节模型的建立。赵元任[3]的大小波浪比喻为进一步的研究指明了方向；沈炯的“音域聚合”[4]实验验证并扩充了这种思想，提出了高低音线的概念；王安红等[6]提出了利用“低音点”和“次低音点”求取低音线的方法。本文在基于以上认识的基础上，利用最小均方误差准则(Minimum Mean Square Error， MMSE)，求取高音线及低音线，并建立了灵活有效的基频控制模型，提高了语音合成系统的质量。

2　高低音线的求取

2.1　理论探讨

为了能够有效地描述连续语流中基频变化的规律，赵元任引入了音域的概念[3]。他定义音域为“不同声调之间以及声调升降极限内的音高范围，是一个由发音力量和声带振动力量所决定的变量”。沈炯[4]利用声调音域将声调与语调分解成两个相对的音高体系，声调音域是指连续语流中特定位置上声调音高特征分布的范围，它存在“高”和“低”相对比的声调特征，即高音线和低音线。声调音域在语流中不断改变它的高低宽窄，语调就是以句子为单位的声调音域系列。在连续语流中，音域的变化，对应着高音线或低音线在绝对音高坐标系中的移动。音域的加宽和压缩导致音域内部声调曲拱发生量变，但曲拱所反映的声调特征并不改变。一旦确定了高音线及低音线在绝对音高坐标系中的对应音高值，音域内部各声调的绝对音高曲线也就相应确定下来。连续语流中特定音节的基频曲线是在该音节单独发音时的基频曲线的基础上经过这种音域的调整得到的。也就是说，单独发音时的音节的基频曲线是在连续语流的音域变化的控制之下，成为连续语流中的基频曲线。

设单独发音时某音节基频曲线为 f_s，其在连续语流中的基频曲线为 f_c，连续发音时作用于该音节的音域为 f_h 和 f_l。参考基频曲线归一化的相关理论[9]，可得如下方程：$f_c = a \cdot f_s + b$，其中 $a = f_h - f_l, b = (f_h + f_l)/2$。由 MMSE 准则，可得方程 $\mathrm{Err} = \sum_i [f_c(i) - (a \cdot f_s(i) + b)]^2$，使 Err 最小化的过程就是求得该音节的相应最优高低音线的过程。当单独发音时音节的基频曲线与连续语流中相应音节的基频曲线具有很高相关性时，说明发该音节时声道的发声状态是一致的，而两者的实现上的差别就是连续语流中由于声门下压力变化引起的音域的变化造成的。 高相似度也排除了相邻音节声调间的相互作用。所以，当单独发音时的音节与连续语流中相应音节的基频曲线相似度很高时由 MMSE 准则求得的高低音线，可以认为是语调变化的控制因素。

经过单音节基频曲线模板提取、连续语流基频曲线提取、基频归整及 MMSE 准则处理后，就可得到连续语流的高音线及低音线变化的规律。

2.2 实验考察

为了验证以上的假设，本文从单音节、短语和连续语句 3 个层面上对基频曲线进行了研究。

本文首先对单音节的音域进行了考察。对所有具备“声调聚合”[4]的单音节进行音域的研究发现，单音节的高音线和低音线都在正态分布的范围内随机变化，其均值和方差分别为 [152.9,10],[72.6,6] 。这说明人们在发单音节时的音域在误差允许的范围内是稳定的。而且，在对所有音节 4 种声调组合的去声的最高点(高音线)、上声的谷值(低音线)、阴平的平均值、阳平的平均值变化规律的研究中发现，低音线、阴平的平均值、阳平的平均值都随高音线的增大而增大，这表明各个音节单独发音时的基频曲线主要受声道特殊性质的影响，也就是说每个音节的基频曲线有其自身的本质规律。因而，可针对各个单独发声的音节建立该音节的基频曲线模板。

在短语层面上，图 1 为短语“断断续续duan4duan4xu4-xu4 ”经 MMSE 准则提取后得到高音线、低音线变化示意图。从图 1 可以看出，高音线逐渐下降，而低音线随韵律节奏下降(“断断”为一韵律词组，“续续”为一韵律词组)。同样的韵律现象可以在众多的韵律短语中得到。这表明低音线主要与节奏单元的完整性有关，高音线与强度有关[4,6,8]，从而也说明了利用 MMSE 准则提取高音线及低音线的可靠性。

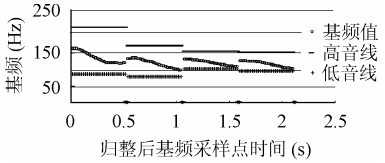

图 1　短语中高音线及低音线变化示意图

本文对连续语流利用 MMSE 准则也成功地提取了高音线及低音线随时间变化的规律。由于音段间协同发音等因素的影响，在单音节基频曲线模板与连续语流中相应音节的基频曲线相关系数不高时有可能出现异常音域。如果我们只考虑两者相关系数很大时(如大于 0.8)音节的参数，将得到很合乎自然言语规律的高音线低音线分布,而且这样做也并不影响我们对总体音域变化趋势的把握。图 2 为对句子“会议还

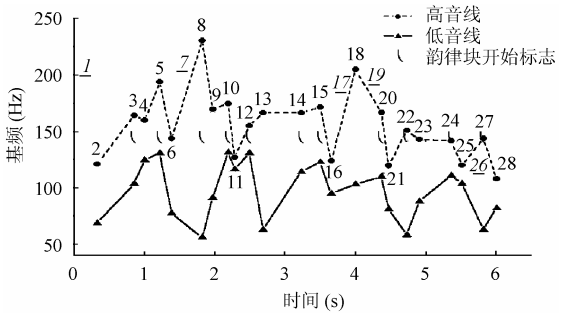

图 2 高音线低音线在例句中的变化

就遏制柬国内政治暴力并创造中立的政治环境进行了讨论。hui4yi4 hai2 jiu4 e4zhi4 jian3 guo2nei4 zheng4zhi4 bao4li4 bing4 chuang4zao4 zhong1li4de5 zheng4zhi4 huan2jing4 jin4xing2liao5 tao3lun4”所提取的高音线及低音线在相关系数大于 0.8 时的变化示意图。图中数字分别表示音节在例句中的序号，带下划线的斜体数字表示该音节的高低音线提取不正确，故没有显示。

从图 2 可以看出，高音线及低音线整体是下倾的[4,6,8](从各个韵律块的首音节的音域可以看出)，且在每个韵律块内部也呈现同样的下倾趋势(从各个韵律短语可以看出)，只是下倾的幅度随短语的不同而不同。这与自然言语的真实基频状态是一致的。连续语流中，由于受生理条件的限制，音域有逐渐下倾的趋势，同时人们也可以灵活地根据环境以及需要强调的内容在不同的韵律单元实现不同的音域调节。

由单音节声调聚合、短语高音线低音线提取及句子音域提取的结果可以得出结论，利用 MMSE 准则提取的音域是可靠的。由于受音节间协同发音及发声机制灵活性的影响，单个音节与连续语流中相应音节的基频曲线的相关性有可能很差，导致并不是每个句子都能用该方法得到相应的高音域及低音域，或者说理想的高音域及低音域。去声音节在连续语流中的变化很稳定，其余的音节类型很容易受前后音节变化的影响，所以去声音节的相似度很高。为了尽可能去除其余因素的影响，只对相关系数大于 0.8 的音节进行研究。

3 双线模型的建立

高低音线的提取只是手段，其目的就是要利用高低音线建立完善的基频调节模型。由于人类在自然言语中有很大的随意性和灵活性，即便同一个句子也可以以不同的韵律发出而得到听感上自然的语音。“能从少量语言现象中得到的规律并不意味着可以从众多的语言现象中得到”[1]。所以，从单个的句子中就有可能得到符合人类发音的一般规律，再经过适当的调整，就可应用于合成系统的韵律调整。下面以上例所提取的高音线及低音线建立相应的基频调整模型。

为了能更直观地研究“大波浪”及“小波浪”[3]的变化，本文首先提取各个韵律块的首音节对“大波浪”进行研究，然后把韵律块进行时间归整后组合在一起对“小波浪”进行研究。为进一步排除音段间协同因素的影响，从整体上把握音域的变化，该项研究剔除了韵律块中被中性声调化的音节。

3.1 理论模型的建立

由弹性膜的振动定律[2]可得

$$F_0 = c_0 \cdot \sqrt{T/\sigma} \tag{1}$$

其中，c_0 振动膜的性质参数，σ 为薄膜单位面积上所承受的强度，T 为薄膜所承受的压力，F_0 为薄膜的振动频率。声门下的气流压力随呼出气流的增加有减小的趋势。我们假定声门下气流压力在肺体积大致不变的条件下，基本成线性率减。相应公式为

$$T = a \cdot P_v \cdot T_m \tag{2}$$

$$P_v = b \cdot (S - v \cdot t) \tag{3}$$

$$\begin{aligned} T &= a \cdot b \cdot (S - v \cdot t) \cdot T_m \\ &= a \cdot b \cdot (S - v \cdot t) \cdot T_m \end{aligned} \tag{4}$$

其中，α 是与发音人性质有关的常量，P_v 是由于声门下气流的变化所施加于声带的作用因子，T_m 为声带本身所受的张力，b 为常量；S 为肺的总体积，v 为气流呼出的速度，t 为时间。

结合式(1)及式(4)并两边取对数，可得式(5)：

$$\begin{aligned} \ln(F_0) &= \ln\left(c_0 \cdot \sqrt{\frac{a \cdot b}{\sigma}}\right) + \frac{1}{2} \cdot \ln(S - v \cdot t) + \frac{1}{2} \cdot \ln(T_m) \\ &= \ln(F_b) + \frac{1}{2} \cdot \ln(S - v \cdot t) + \frac{1}{2} \cdot \ln(T_m) \end{aligned} \tag{5}$$

式(5)的第 1 部分可以认为是基准值；第 2 部分为“大波浪”；第 3 部分为“小波浪”。大波浪为对数率减形式；大波浪与小波浪成对数迭加的形式。

3.2 模型参数的求取

本文在 MMSE 准则得到高音线和低音线的基础上，利用上面所描述的理论模型，根据合成分析方法，得到了相应的未知参数，从而建立了基频控制的“大波浪”“小波浪”模型。

根据式(5)，可首先假定“大波浪”数学模型为

$$\mathrm{In}\, f_p = \mathrm{base} + \mathrm{In}(b - c \cdot x) \tag{6}$$

其中 f_p 为产生的基频，base 为基准值，a,b 分别为发音的个性特征参数。

只对各个韵律块的第 1 个音节进行考察，便得到大波浪的形状及模型参数如图 3 所示。从图上可以看出，高音线及低音线的模型的相关系数都很高，表明在误差允许的范围内，“大波浪”符合对数率减的规律。

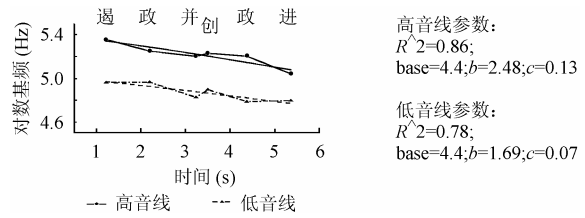

图 3　大波浪及其参数模型

以各个音节在韵律块中的时间为X轴，便可得到“小波浪”波形。根据“小波浪”波形状，我们结合Fujisaki模型[2]构建了“小波浪”模型。即

$$\ln f_p = \text{base} + a \cdot \ln(b - c \cdot x_g) + a_c \cdot \text{alpha} \cdot x_c \cdot e^{(-\text{alpha} \cdot x_c)} \qquad (7)$$

其中 x_g 为全局时间， x_c 为韵律块时间，其余为与发音人和发音状态有关的参数。在图 3 所示“大波浪”模型参数的基础上，得到“小波浪”模型参数(如图 4 所示)。

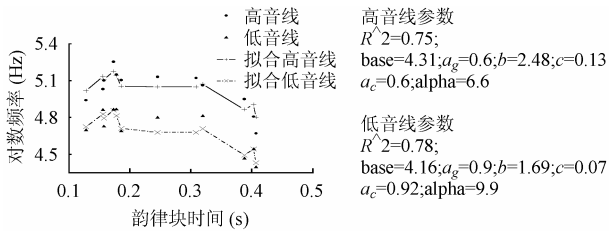

图 4　小波浪及其参数模型

从模型的高音线和低音线的变化来看，低音线的变化明显地大于高音线，说明低音线主要与节奏的完整性有关，这与沈炯等人的相关理论[4,6,8]一致。“大波浪”模型基本在10s左右达到极小值，这与人们发音时的极限状态相对应，在不吸气的条件下，10s 以后很难发出正常的声音，所以，在一定的时间间隔后应设立“大波浪”的重置；模型的“小波浪”基本上在0.8s 时率减到零，这与韵律块单元的长度相对应。另外，从所提取的高音域及低音域值可以看出，正常的发音多以两字组音节为一韵律单元，这与自然语气时的规律也是一致的。

4　调中值及调域模型的建立

调中值和调域[8]作为音节基频的两个特征，广泛应用于合成系统的基频调节。本文利用所求取的高音线及低音线，建立了基频控制的调中值模型。相应的模型参数及示意图如图 5 所示。

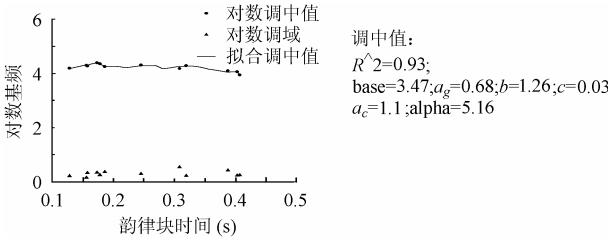

图 5　调中值及调域模型示意图

与“大波浪”“小波浪”模型相似，调中值的“大波浪”模型用对数率减模型来表示；“小波浪”模型用 Fujisaki 模型来表示；调中值的相关系数高达 0.93 以上，而调域却小于 0.3。说明，调中值与韵律单元节奏的完整性高度一致，是一种基本的趋势,而调域主要受发音随机性干扰所导致的强度或故意强调的程度有关。

5　模型比较及应用效果分析

本文在理论及数据统计分析的基础上，基于 MMSE 准则建立了高音线低音线双线调节模型及调中值模型。两模型的本质一致，调节的形式及适用的范围不同。不同的语调类型可以通过调节模型的重置来实现(在对该语调模型分析的基础上)。所建立的控制模型的基本机制都与相关的语音学理论一致，这说明了该方法的可靠性及适用性。同时，基频控制模型的建立也为合成系统基频的灵活有效调整提供了基础，从而能提高语音合成系统的自然度。

本文所建立的模型是在对语调理论整体把握的基础上单句分析的结果。由于人类发音的灵活性，每个语句都有其特殊的自然性，所以把很多语句综合起来建立相应的模型是很困难的。通过特殊的例子推广到更广泛的应用，并在实践中检验结果的可靠，是一种有效的方法。

图 6 为根据双线模型对基频曲线模板调节前后的比较示意图。从图 6 可以看出，调整后的基频曲线与连续语流的基频曲线十分的接近，表明了双线模型的效果是适用的。

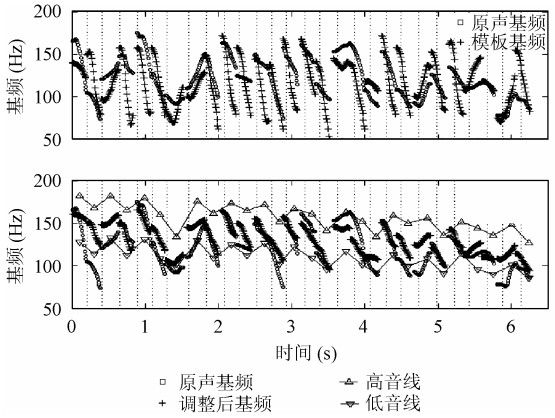

图 6　双线模型效果分析及音域示意图

6　结束语

在对所有单音节音域进行实验考察的基础上，本文提出了基于最小均方误差准则(MMSE)的高音线、低音线提取方法。该方法突破了以往在语调研究中主要靠感性认识来研究复杂的语音学现象的限制，从理性的角度验证了语调的音域调控相关理论。本文对语调短语单元及连续语流的音域提取

的大量实验证明，汉语中存在音高下倾现象，高音线、低音线在语调的调控机制中承担不同的语言学功能，音域调控机制是不同语言学功能语调的主要作用因素。而且，该方法也为进一步表情语调的研究提供了基础。

本文对语调下倾的现象从生理机制的角度做了初步探讨，提出了相应的数学模型。在所提取的高音线、低音线的基础上，建立了反映语调特征的调节模型(双线调节和调中值模型)。模型的参数都有明确的物理意义，分别反映了不同的语音学特征，所以可以针对具体的语境做灵活的调整，从而实现更高自然度的合成效果。本项研究成功地在本实验室的合成系统实现了韵律基频灵活有效的调整，合成语句的自然度也得到相当的提高。

参 考 文 献

[1] Kochansaki Greg, Shih Chilin. Prosody modeling with soft templates. *Speech Communication*, 2003, 39(3-4)：311 – 352.

[2] Fujisaki Hiroya. The fundamental frequency contour of speech—Its modeling, underlying mechanisms, and application to multilingual speech synthesis. In Proceedings of ICSP’99，Seoul Korea，1999：19 – 26.

[3] 赵元任. 汉语的字调跟语调. 赵元任语言学论文集, 上海: 商务印书馆，2002：734 – 749.

[4] 沈炯. 北京话声调的音域和语调. 北京语音实验录, 北京: 北京大学出版社，1985：73 – 130.

[5] Xu Y, Wang Q E. What can tone studies tell us about intonation? In A. Botinis, *et al.* (Eds.) Intonation: Theory, Models and Applications, In Proceedings of an ESCA Workshop, European Speech Communication Association, Athens, Greece, 1997：337 – 340.

[6] 王安红, 陈明, 吕士楠. 基于言语数据库的汉语音高下倾现象研究. 声学学报，2004，29(4)：353 – 358.

[7] 王蓓，吕士楠，杨玉芳. 汉语语句中重读音节音高变化模式研究. 声学学报，2002，27(3)：234 – 240.

[8] 杨顺安. 浊声源动态特性对合成音质的影响. 中国语文，1986, 3：173 – 181.

刘浩杰： 男，1976 年生，博士生，从事语音信号处理、语音合成、声波测井信号处理研究。联系方式 Liuhj@iis.ac.cn.

杜利民： 男，1957 年生，博士，研究员，博士生导师，IEEE 高级会员，中国电子学会理事，研究方向为语音信号处理、自然语言理解、语音交互信息技术.