第 28 卷第 5 期
2003 年 9 月

声 学 学 报
ACTA ACUSTICA

Vol. 28, No. 5
Sept., 2003

基于音节韵律特征分类的汉语语音合成中韵律模型的研究*

陶建华 蔡莲红

(清华大学计算机科学与技术系 北京 100084)

2001 年 12 月 10 日收到
2002 年 8 月 13 日定稿

摘要 论述了采用基于统计模型进行韵律建模的思路。在此基础上，提出了基于音节韵律特征分类的韵律建模思路，并采用韵律模板和韵律代价函数实现了韵律的自动预测。对该模型的自动训练算法进行了详细的阐述。根据统计的韵律建模方法，还分析了韵律特征间相互关联对音节韵律模板选取的影响。最后，进一步分析了统计韵律模型的进行韵律预测的误差分布情况，表明了该模型能够使语音合成系统具有较高自然度和高灵活性的特性。

PACS 数： 43.70

Study of prosody model on Chinese speech synthesis based on the classification of syllabic prosody features

TAO Jianhua CAI Lianhong

(*Department of Computer Science and Technology Tsinghua University* Beijing 100084)

Received Dev. 10, 2001
Revised Aug. 13, 2002

Abstract A prosody modeling method based on statistic model is described. Based on this, a Chinese prosody model based on the classification of syllabic prosody features is presented, which makes automatic prosody prediction with prosody templates and prosody cost function. And the automatic training algorithm of the model in detail is described. Further more, according to statistic prosody modeling method, the influence to prosody template selection with the help of the analysis of the prosody interaction among prosody elements is analyzed. Finally, the error distribution of the statistic method based prosody prediction is given. The results show good naturalness and much flexible in application.

引言

语音合成技术自上世纪 60 年代 Klatt 语音合成器诞生以来，进入了一个语音数字计算的时代，伴随着计算机技术的飞速发展和大量普及，语音的整体研究水平获得了不断的进步。尤其是进入 20 世纪 90 年代以来，自然语言理解、信号处理、随机过程、模式识别等技术在语音处理中得到了非常成功的应用，导致了语音技术在多项关键性技术上的突破。在这种背景下，语音合成中韵律建模的研究也同样融入了许多新的概念。近 10 年来，由于数据驱动技术的发展，使语音合成的研究由过去的侧重声学模型和韵律规则的研究转变为侧重采用连续语料的数据驱动方法的研究。因而，语音合成的研究更多地超出了语音信号处理本身，而需要更多地对大容量数据进行分析和特征自动提取技术。语音合成技术的研究也正朝着一个系统工程的方向发展。

连续语流中的韵律特征，实质上可以认为是一种韵律特征向量的随机过程，在此基础上，建立韵律模型，有利于统计方法在其中更好的体现。汉语是

* 863 资助项目 (2001AA114072)

一种有调语言，相较西方语言，其韵律特征的表现有其非常大的独特性。正如我们所知，汉语音节共有 5 种声调：阴平 (1)、阳平 (2)、上声 (3)、去声 (4) 和轻声。除轻声外，其余声调的基频都有着基本的调型表现。由于汉语语音节奏性强，单个音节除具有自身韵律特征外，在连续语流中由于音节所处的语境不同，则会发生音高、音强、音长变化等协同发音现象。如：受句中的位置和不同字组声调组合的影响，汉语音节或词的调域、音长均有不同的变化模式，而表现出规律性的差异。而由基频调域、停顿，以及音长的变换，则产生不同的轻重模式，用以体现不同语句的语气和情感。汉语音节基频曲线的特殊性，使其非常适用于特征分类的方法，对其不同的韵律特征表现加以归类、分析。因而，本文提出了一种基于韵律特征分类的汉语语音合成中韵律建模的新思路，它与统计模型相结合，较好地解决了汉语语音发音的平稳性和连贯性的问题。

文章在第 2 节中论述了基于统计的韵律建模思路，在第 3 节中进一步阐述了音节韵律模板的生成，以及采取基于统计模型的韵律代价函数来进行韵律模板的选取算法。论文还仔细分析了韵律特征间相互关联对音节韵律模板选取的影响，并在此基础上构筑了模板选取的矩阵，并采用路径搜索的方式来确定韵律参数。通过进一步的合成语音的韵律分析结果和测听结果的分析 (第 4 节)，表明采用本文提出的韵律建模的思路，对汉语语音合成系统的构筑，达到了明显的自然度提高的效果。同时，本文还分析了基于统计的韵律模型的误差分布情况，指出这种方法对于韵律预测可计算性研究的重要性。

1 韵律建模的概率描述

人在不同的语境下会有不同的韵律特征，语境与韵律特征之间具有很强的相关性。谈到语音和句法以及语义之间有密切关系的时候，林焘先生强调“绝对不能把语言的这三方面割裂开来孤立地进行研究。”语境与韵律特征之间具有很强的相关性。“语调构造由语势重音配合而形成。它是一种语音形式，它通过信息聚焦来实施超语法的功能语义。”[5] Katherine Morton[2] 在他的对话系统的语音合成模块中，根据几个基本的上下文模式，加入情感的变化，使合成出的语音变得有些生动。

汉语中的语境信息，根据的文本的上下文信息，可以按照其对汉语韵律特征不同层次的影响，将其沿着语句 (Sentence) - 韵律短语 (Prosody Phrase) - 韵律词 (Prosody Word) - 音节 (Syllable) 的思路划分开。共分为 5 组：当前音节信息 (声母类型、韵母类型、声调类型、在词中位置、与前音节耦合度和与后音节耦合度)；相邻前音节信息 (韵母类型和声调类型)；相邻后音节信息 (声母类型和声调类型)；音节所在韵律短语信息 (音节数、在句中位置、重音类型、距前一个重音距离和距后一个重音距离) 以及语句信息 (语句类型和韵律短语个数)，共 17 个参数，与前后音节耦合度指的是相邻音节的边界特性。

语句信息、短语信息反映了整个句子的语气变化和重音的情况。所有这些标注信息共同决定着音节基频、音长等韵律参数的基本特性。

连续语流中的韵律特征，是一种韵律特征向量的随机过程，韵律特征参数的分布受着语境信息的影响，这种影响又满足一定的概率关联关系，而不是一个简单的函数映射。正如在汉语轻声的研究工作中，沈炯先生在“从轻音现象看语音与语法研究的关系”[7] 一文中阐述：“当我们说一种语音现象很明显的时候，主要是指它的离散成分很容易把握，一般并不指它的音理是否容易认知。离散成分是平常人都能把握的形式 —— 在言语使用过程和语言教学中分别把握它。只要它能被把握，语音研究就可以在离散基础上进行，可以用文字或其他表音手段来列举语言样品，讨论问题就方便了。”

从概率的角度出发，对于一个已知的语句，其韵律单元的语境序列可以表示为 $(\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_N)$，与每一个韵律单元相对应的韵律特征参数，为其所有可能的韵律特征参数中出现概率最大的一组，即为：

$$\boldsymbol{Y}_n = \arg\max_m P(\boldsymbol{Y}_{n,m}|\boldsymbol{A}_n) \tag{1}$$

其中，$\boldsymbol{A}_n$ 表示语句中第 n 个韵律单元的语境信息，$\boldsymbol{Y}_{n,m}$ 表示第 n 个韵律单元第 m 个可能的韵律特征参数，$\boldsymbol{Y}_n$ 表示第 n 个韵律单元最有可能出现的韵律特征参数。

由 Bayesian 公式可以得到：

$$\boldsymbol{Y}_n = \arg\max_m P(\boldsymbol{Y}_{n,m}|\boldsymbol{A}_n) = \arg\max_m \frac{P(\boldsymbol{A}_n|\boldsymbol{Y}_{n,m})P(\boldsymbol{Y}_{n,m})}{P(\boldsymbol{A}_n)}. \tag{2}$$

公式 (2) 中，$P(\boldsymbol{A}_n|\boldsymbol{Y}_{n,m})$ 表示不同韵律特征下的语境参数出现概率，它是一个可以通过语料统计或分析得到的先验概率模型；$P(\boldsymbol{Y}_{n,m})$ 表示不同韵律特征在真实语言中的统计分布，由于它受某一特定韵律单元的影响，因而它的分布不仅和韵律单元的分布有关，同时也受相邻韵律单元的影响，而这种影响又通过韵律特征间的出现概率和相互关联来体现。$P(\boldsymbol{A}_n)$ 则表示语境信息的统计分布，由于真实

语言丰富多彩，在超大规模的语料中，语境信息往往表现出均匀分布的特性，因而在处理中往往可视其为常数，将其忽略。公式 (2) 可以进一步转换为：

$$\boldsymbol{Y}_n = \arg\max_m P(\boldsymbol{Y}_{n,m}|\boldsymbol{A}_n) = \arg\max_m P(\boldsymbol{A}_n|\boldsymbol{Y}_{n,m})\,P(\boldsymbol{Y}_{n,m}) \tag{3}$$

公式 (3) 表明，为求 $P(\boldsymbol{Y}_{n,m}|\boldsymbol{A}_n)$ 这样的后验概率，可以将其转换为求 $P(\boldsymbol{A}_n|\boldsymbol{Y}_{n,m})$ 这个先验概率。若对公式 (3) 进行不同的条件约束，则可以得到其不同的体现形式。

假设 1: 对于 $\boldsymbol{A}_n$ 存在 $\boldsymbol{Y}_{nk}$，使得

$$\begin{cases} P(\boldsymbol{Y}_{n,k=l}|\boldsymbol{A}_n) = 1 \\ P(\boldsymbol{Y}_{n,k\neq l}|\boldsymbol{A}_n) = 0 \end{cases} \quad (k,l \in I,\ n \in N)$$

则公式 (3) 可以转换为函数表示，即：

$$\boldsymbol{Y}_n = \varphi(\boldsymbol{A}_n). \tag{4}$$

由此，韵律模型被简化成了简单的函数映射，韵律模型则变成了规则模型，公式 (3) 中的先验概率模型则可以通过规则统计来实现。

假设 2：　$P(\boldsymbol{Y}_{n,m}) =$ 常数C

则公式 (3) 变为：

$$\boldsymbol{Y}_n = \arg\max_m P(\boldsymbol{Y}_{n,m}|\boldsymbol{A}_n) = \arg\max_m P(\boldsymbol{A}_n|\boldsymbol{Y}_{n,m}). \tag{5}$$

公式 (5) 表明在基于 $P(\boldsymbol{Y}_{n,m}) =$ 常数 C 假设的下，即假设韵律特征本身不产生相互关联的情况下，语境信息对韵律特征的影响和韵律特征受语境的制约是等价的，从而便于采用简化的统计模型实现韵律特征的预测。

2　汉语音节基频特征分类及韵律模板生成

2.1　汉语音节基频聚类分析

针对汉语音节的不同韵律表现，韵律特征分类是将不同声调中所有音节基频曲线进行聚类处理，将不同的音节基频曲线归类并量化成不同的韵律模板。

聚类算法表述如下：

在每一个音节的基音曲线上均匀地选取 10 个基频值：$f(t_0), f(t_1), \cdots, f(t_i), \cdots, f(t_9)$。其中，$t_i$ 表示音节基频曲线中的第 i 个基频值。基频曲线聚类的算法基于不同的声调，具体描述为：

第 1 步：将所有音节的基频曲线的基频值归一化到 0 和 1 之间。

第 2 步：计算每一个音节基频的平均值：

$$\overline{f_j(t_i)} = \frac{1}{10}\sum_i f_j(t_i). \tag{6}$$

其中，j 表示音节的序号。

则得到：

$$f'_{ji} = f_j(t_i) - \overline{f_j(t_i)}. \tag{7}$$

第 3 步：在同一个音节声调值中，计算两两音节基频曲线的相关系数：

$$R_{jk} = \frac{\sum_{i=0}^{9}(f'_{ji}f'_{ki})}{\sum_{i=0}^{9}(f'_{ji}f'_{ji}) + \sum_{i=0}^{9}(f'_{ki}f'_{ki})} \tag{8}$$

第 4 步：用这些相关系数构成 5 个不同声调的 $(M \times M)$ 相似矩阵。

$$\boldsymbol{A} = \begin{bmatrix} R_{00} & R_{01} & \cdots & R_{0M} \\ R_{10} & R_{11} & \cdots & R_{1M} \\ \cdots & \cdots & \cdots & \cdots \\ R_{M0} & R_{M1} & \cdots & R_{MM} \end{bmatrix}. \tag{9}$$

这里，M 表示相同声调中所有音节的个数。进而，可以采用 Max-Tree 方法，利用这个相似矩阵，将其中的所有的元素分类，如图 1 所示。

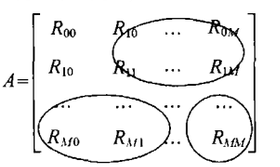

图 1　音节基频曲线聚类示意图

第 5 步：对于每一类，计算该类中出现的所有音节的平均基频曲线，同时得到基频曲线的均方误差：

$$\overline{f_i} = \frac{1}{M}\sum_j f_j(t_i)$$

$$\alpha = \frac{1}{N}\sum_{j=0}^{N-1}\sqrt{\frac{1}{10}\sum_{i=0}^{9}\left\{\left[f_j(t_i) - \overline{f_i}\right]^2\right\}}$$

最后：重新计算每一类中处于均方误差范围内的音节平均基频曲线，把这个平均基频曲线作为属于该类中所有音节的归一化基频模板：

$$FR_i = \frac{1}{M}\sum_j f_j(t_i). \tag{10}$$

这里，$f_j(t_i)$ 与该类均值的差小于该类的均方误差，即 $f_j(t_i) - \overline{f_i} \le \alpha$。

在实际工作中，根据分类结果，每一种声调 (包括轻声) 各取 20 种类型，共得到 $20 \times 5 = 100$ 个典型的基频曲线分类，在此基础上提取 SPiS 参数[9]，构筑成汉语音节基频曲线模板。

3 音节韵律模板的选取

汉语的韵律特征相对语境信息的变化，通过韵律模板的方式体现为韵律模板特征与语境的关系。因而，韵律模型中的韵律预测转化为汉语音节韵律模板的选取与韵律声学参数的生成两个方面。

3.1 韵律模版选取中的韵律代价函数

韵律模板的选取是基于统计模型的韵律模型的核心思想，通过式 (5) 来进行，在实现过程中可以采用韵律代价函数来替代，设定韵律代价函数为：

$$S_{n,m}=\sum_i \gamma_i V(a_{n,m,i}), \tag{11}$$

其中 $\gamma_i = f(\omega_i)$, $a_{n,m,i}$ 表示待合成语句中第 n 个音节中第 m 个候选样本中第 i 个语境参数值，它是语境信息的数值化表示，通常取非负整数。函数 $V(a_{n,m,i})$ 表示候选韵律模板的语境参数 $a_{n,m,i}$ 与目标语境参数的逼近度，它归一化到 0~1 之间的值。本文将语境参数按其数学特性分为不分级量化和分级量化两类。不分级量化类包括：词性、声韵母类型等，不同参数不代表参数之间的层次关系，只反映了参数的类型；分级量化类包括：重音级别、边界特性、所有位置信息和距离信息等，这些参数的值具有量化可比性。针对不分级量化类，函数 $V(a_{n,m,i})$ 可以表示为：$\begin{cases} 0 \text{ 当} a_{n,m,i}=\hat{a}_{n,i} \\ 1 \text{ 当} a_{n,m,i}\neq\hat{a}_{n,i} \end{cases}$，针对分级量化类，函数 $V(a_{n,m,i})$ 可以表示为：$1-\dfrac{|a_{n,m,i}-\hat{a}_{n,i}|}{\max\limits_i(a_{n,m,i})}$。其中，$\hat{a}_{n,i}$ 表示待合成语句中第n个音节第i个语境参数值。

韵律代价函数的结果为一组语境信息的加权统计值。通过权值的记忆和训练达到适应不同语料库的目的。由公式 (5) 可以进一步得到：

$$\boldsymbol{Y}_n=\arg\max_m P(\boldsymbol{A}_n|\boldsymbol{Y}_{n,m})=\arg\max_m(S_{n,m})=\arg\max_m\left[\sum_i \gamma_i V(a_{n,m,i})\right]. \tag{12}$$

通常情况下可以认为：

$$\gamma_i=f(\omega_i)\approx\omega_i. \tag{13}$$

其中 ω_i 为不同语境参数产生贡献的影响因子或称权值。函数权值 ω_i 的初始值确定，对韵律处理的影响很大，虽然进一步的权值调整可以通过训练机制来实现。

3.2 初始值确定

韵律代价函数的权值参数是构成韵律代价函数的最重要组成部分，对选取的结果以及合成语音的自然度有着直接而关键的影响。权值的初始值确定实际上就是一个寻找语境参数各个分量对韵律特征变化的灵敏度问题。当语境参数的分量变化带来较大的韵律特征变化，且变化的特性比较接近时，该参数对应的权值相应也较大，本文将该语境参数定义为语境敏感参数；反之，当语境参数的分量变化不带来明显的韵律特征变化，或韵律特征的分布比较杂乱时，该参数对应的权值则较小，本文将其定义为语境非敏感参数。在每一类中，分别统计与之相应的语境各参数的分布情况。当每一类中语境参数的分布较为集中，且与其它类中语境参数的分布特性不同时，该语境参数属于语境敏感参数；反之，该语境参数属于语境非敏感参数。

试验结果表明，针对汉语，语境敏感参数包括：音节所在的位置信息、韵律短语边界、声调信息、重音、词性等。它们对韵律的影响较大，占据着较为重要的作用，因而与它们相对应的权值也较大。通过该结果，辅以人为经验，可以非常方便地确定韵律代价函数的权值初始值。

3.3 权值的训练

虽然，韵律代价函数的权值的最终确定可以依靠设计者的经验，并经过大量测试来人工确定。但依靠自动训练的机制或统计的方法，却能给模型带来很大的弹性，并能产生自动适应语料库的能力。

假设：韵律代价函数中初始权值向量为：$\boldsymbol{\omega}^0=\{\omega_1^0,\omega_2^0,\cdots\omega_p^0\}$，经过 $j-1$ 次训练后，权重系数为 $\boldsymbol{\omega}^j=\{\omega_1^j,\omega_2^j,\cdots\omega_p^j\}$，其中 p 为权值矢量的维数，j 为非负整数。它们满足约束条件：

$$\sum_{i=1}^{p}\omega_i^j=1. \tag{14}$$

约束条件的目的是使得训练尽量能够产生收敛，同时使得权值的调整在整个向量空间保持均衡。

训练所采用的样本集为：$\{\widehat{\boldsymbol{Y}_1},\widehat{\boldsymbol{Y}_2},\cdots,\widehat{\boldsymbol{Y}_N}\}$，通过韵律代价函数选择得到的样本集为：$\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_N\}$

其输出误差为：

$$E(\boldsymbol{\omega}^j)=E\left(\boldsymbol{\omega}^j,\boldsymbol{Y}\right)=\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{\boldsymbol{Y}_n}-\boldsymbol{Y}_n\right)^2 \tag{15}$$

其中，$\boldsymbol{Y}$ 由样本的声学参数构成向量空间，它由样本的基频数据、音长组成，即：$\boldsymbol{Y}=(\boldsymbol{P},D)$。$\boldsymbol{P}$表示基频数据，为文献9中对音节基频描述的 SPiS 参数；D 为音节的音长。

当训练进入第 j 步时，函数的权重调节通过下式进行：

$$\omega_i^{j+1} = \omega_i^j + \eta^j d_i^j, \tag{16}$$

其中 η^j 为第 j 步权重调节的步长， d^j 则表示第 j 步权值调节的方向。

对于 ω_i^j，曲面变换最陡的方向即为 $\nabla E(\omega_i^j)$ 所指的方向。由此，权值调节方向 d^j 应为 $\nabla E(\omega_i^j)$ 的函数。即：

$$d = f\left[\frac{\partial E(\omega_i^j)}{\partial \omega_i}\right] \tag{17}$$

本文取：

$$d = -\frac{\partial E(\omega_i^j)}{\partial \omega_i}. \tag{18}$$

在实现过程中，式 (18) 可以用如下步骤来实现：

定义：$\Delta \boldsymbol{Y}_n$ 为第 n 个音节，韵律模板选择的结果与目标音节之间的韵律特征误差，它可以表示为：

$$\Delta \boldsymbol{Y}_n = \left| \arg\max_m \left[\sum_i \gamma_i V(a_{n,m,i}) \right] - \widehat{\boldsymbol{Y}_n} \right|,$$

与之相对应的语境参数为：$A_n^s = (a_{n,1}^s, a_{n,2}^s, \cdots, a_{n,p}^s)$，相应的语境参数各个分量与目标语音的语境参数各分量之间的误差为：

$$\Delta(V(a_{n,i}^s)) = [V(a_{n,i}^s) - V(\widehat{a}_{n,i})]$$

定义： 第 n 个音节中，所有候选样本与目标音节韵律特征误差最小的样本，其语境参数为 $A_n^{\min} = (a_{n,1}^{\min}, a_{n,2}^{\min}, \cdots, a_{n,p}^{\min})$，最小韵律特征误差为：$(\Delta \boldsymbol{Y}_n)_{\min}$，相应的语境参数各个分量与目标语音的语境参数各个分量之间的误差为：$\Delta(V(a_{n,i}^{\min})) = [V(a_{n,i}^{\min}) - V(\widehat{a}_{n,i})]$

实际运算过程中，语境参数已经被归一化到 0 和 1 之间，因而可以计算得到：

$$d_i^j = \left[1 - \frac{(\Delta \boldsymbol{Y}_n)_{\min}}{\Delta \boldsymbol{Y}_n}\right] \left[\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s))\right] + C. \tag{19}$$

由约束条件 (14)，并结合式 (16) 可以得到：

$$\sum_{i=1}^P \omega_i^{j+1} = \sum_{i=1}^P \omega_i^j + \sum_{i=1}^P \eta^j d_i^j = 1.$$

进而可以求出：

$$\sum_{i=1}^P \eta^j d_i^j = 0.$$

通常为了简化运算过程，η^j 被假定为 0～1 之间的一个常数 η。则可以得到：

$$\sum_{i=1}^P d_i^j = 0.$$

从而：

$$C = \frac{1}{P}\left[\frac{(\Delta \boldsymbol{Y}_n)_{\min}}{\Delta \boldsymbol{Y}_n} - 1\right] \sum_{i=1}^P \left[\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s))\right]. \tag{20}$$

利用公式 (16)，(19) 和 (20) 可以实现训练的整个过程。

4　韵律特征的转移特性

在通过权值确定的韵律模板选择模型中，并没有考虑韵律特征本身的相互关联。由于韵律参数预测除了要有较好的语境信息到韵律参数之间的预测模型外，必须将整个模型放在整句甚至是篇章的整体环境中考虑。而在连续的语句中，韵律特征间的耦合效应则不可避免地出现，故在韵律建模中，引入反映韵律特征转移的函数，对从整体上完善韵律模型并提高韵律参数预测质量，将会起到一定的作用。

由公式 (5) 是在假设 $P(\boldsymbol{Y}_{n,m})$ 为常数的前提下所得，即 $P(\boldsymbol{Y}_{n,m})$ 所起的作用被忽略。而很多情况下， $P(\boldsymbol{Y}_{n,m})$ 能够反映韵律特征本身的相互关联，且非常明显。例如：当一个音节本身被重读时，通常会影响后续音节的发音等。这一结论，在吴宗济先生的 “普通话三字组变调规律”[4]，以及林茂灿的 “北京话轻声的声学性质”[10] 、 “普通话轻声与轻重音”[8] 等论文中均得到了不同程度的论证。如：林茂灿提到的 “轻声音节 F_0 曲线的形成，是由于它跟前面重读音节发生声调协同发音所致”，就是这一现象的典型反映。又如：汉语中出现的词中，其音节基频曲线的形状，往往受着前后发音的基频曲线或发音轻重的影响。而这些因素却常常被语音合成系统的韵律模型所忽视。因而，一个好的韵律模型必须要能够反应韵律受语境的制约情况，而且还因能够体现韵律自身相互关联的现象。

更多的情况是 $P(\boldsymbol{Y}_{n,m})$ 通过韵律单元的韵律特征转移概率来反应，从整句上看可以用下式来替代：

$$P(\boldsymbol{Y}_{n,m}) = \sum_{g=1}^M P(\boldsymbol{Y}_{n,m}|\boldsymbol{Y}_{n-1,g}) P(\boldsymbol{Y}_{n-1,g}). \tag{21}$$

在采用韵律代价函数进行音节基元选取的韵律模型中，当只考虑相邻音节的韵律特征关系时，其整句评估因子可以表示为：

$$\boldsymbol{Y}_n = \underset{m}{\operatorname{argmax}} \left\{ \sum_g \sum_i \left[\omega_i V(a_{n,m,i}) P(\boldsymbol{Y}_{n,m}|\boldsymbol{Y}_{n-1,g}) P(\boldsymbol{Y}_{n-1,g}) \right] \right\} \tag{22}$$

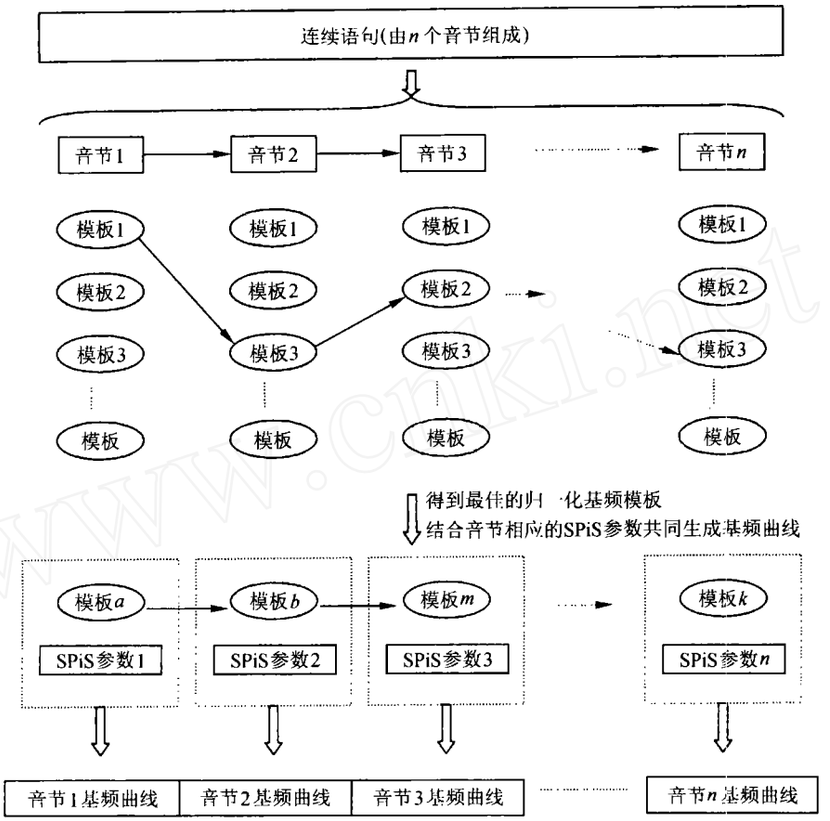

图 2 韵律模型整体示意图

具体实现过程中，通过在韵律模型中增加反馈的方式，或通过将整个韵律模型转换为根据韵律参数出现的同现概率来实现。运用韵律参数出现的同现概率，将韵律模板的选取转化为矩阵中路径搜索的数学问题，如图 2 所示。

5 结果分析

5.1 合成语音的韵律特征分析

基于本文的韵律建摸方法，在韵律预测中获得了较好的应用。由于公式 (12) 和 (22) 的差异，因而它们使得韵律预测结果同样存在一定的差异。相对来说，公式(22)由于考虑到了韵律单元韵律特征相互的关系，因而，获得的结果具有全局较优的特性，辅以 Viterbi 等路径搜索算法，可以较为有效的求取整句的韵律参数。下面是一个对比测试。测试的语料共有 1000 个句子，文本全部来自人民日报，由新闻播音员用正常语速朗读。韵律的标注全部由人工完成。该语料覆盖了汉语所有的有调音节和词性，总共有 16446 个汉字，5669 个语法词 (语法词长为 1~4 个汉字)。测试的方法为分别将基于公式 (12) 和公式 (22) 的基频预测结果与人工标注的结果进行对照误差分析。结果如图 3 所示。

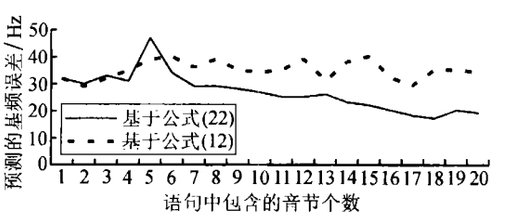

图 3 预测的基频误差与语句中音节个数的关系图

从图 3 中可以看出，从整体上看，韵律预测的误差较低，同时基于公式 (22) 的韵律预测结果误差较基于公式 (12) 的结果为低。图中还发现，在基于公式 (22) 的韵律预测结果中，其误差随着整句中音节的个数增加有下降的趋势。这同时验证了韵律的关联具有提高韵律结果的精度的作用的结论。相对来说，基于公式 (12) 的结果，则表现为与音节个数基本无关。相应的听感测试也同样能够验证这一结果。

图 4 和图 5 则通过对一句具有 17 个音节的句子 (“周进功的这一义举已引起海内外的关注”) 的韵律预测结果，进一步说明公式(12)和公式(22)的差异。

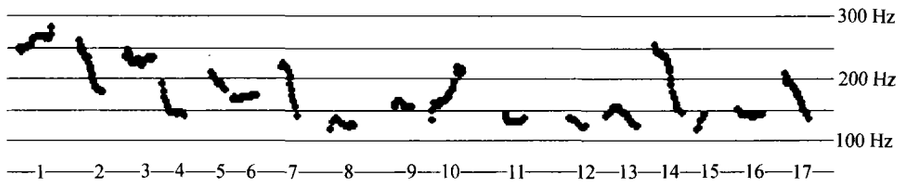

图 4　基于公式 (11) 的韵律预测结果示意图

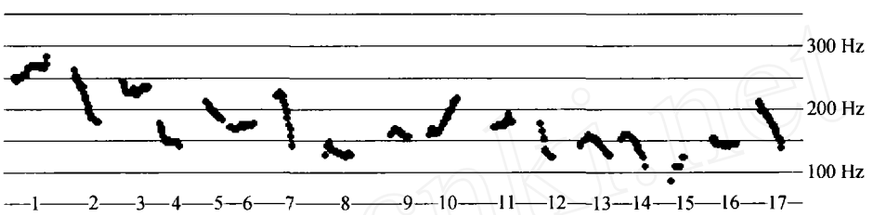

图 5　基于公式 (22) 的韵律预测结果示意图

由图 5 中第 10, 11, 14, 15 音节的基频曲线的变化可以看出，图 5 较好地显示了基频的预测结果，它使得整个句子在第 8 和第 9 音节处，即音节 “举” 和 “已” 的地方呈现出韵律节奏的边界特性，相较图 4 的结果，体现了更好的节奏感，使整句的自然度有所提高。

5.2　基于概率的韵律模型的误差分析

用类条件概率或联合分布表示韵律建模的方法实际上是内插和外插数据的一种方法，也是基于最大后验概率的分类过程。最佳的判决是基于最大后验概率的判断。但最佳的判决并不意味着没有误差，只能认为总的系统误差是最小的。

在模型中，问题不在于是否判断出错，而是要尽可能减少错误。为了评价一个判决规则的好坏，必须计算误判概率，即将一个韵律量化样本错分到其它类别中去的概率。

通过连续联合概率分布可以非常简便地显示出系统误差的基本性质。对于韵律特征量化参数可以将其划分为与语境 A 对应的一类 $\boldsymbol{Y}_1$，和与语境 A 非对应的一类 $\boldsymbol{Y}_2$，如果是基于最大后验概率原则进行判决，则对于语境参数 A，可以得到判决：

$$\text{当且仅当}\quad P(\boldsymbol{Y}_1|A) > P(\boldsymbol{Y}_2|A). \tag{23}$$

即，在 $P(\boldsymbol{Y}_1|A) > P(\boldsymbol{Y}_2|A)$ 的情况下，可以由语境参数 A 得到韵律量化参数 $\boldsymbol{Y}_1$ 。

对于任意 A，误差概率密度由下式给出：

$$\text{误差概率密度} = P_l(A) = \min\{P(\boldsymbol{Y}_1|A), P(\boldsymbol{Y}_2|A)\}. \tag{24}$$

可以用两联合概率分布 $P(\boldsymbol{Y}, A)$ 的光滑曲线来描述，如图 6 所示。

令 A_{ik} 表示两个联合概率分布相交点的 A 值，则根据表达式 (24)，当且仅当 $A > A_{ik}$ 时，判决语境 A 与 $\boldsymbol{Y}_1$ 对应，当且仅当 $A < A_{ik}$ 时，判决语境 A 与 $\boldsymbol{Y}_2$ 对应；对于 $A = A_{ik}$ 的特殊情况，可以判属任意一类。

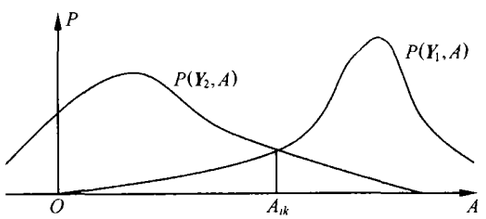

图 6　阈值判断图

由概率密度的特性，误差概率为在所有 A 取值上的平均。

$$\begin{aligned}\text{系统误差} = &\int_{-\infty}^{A_{ik}} P(\boldsymbol{Y}_1|A)\,P(A)\mathrm{d}A + \\ &\int_{A_{ik}}^{+\infty} P(\boldsymbol{Y}_2|A)\,P(A)\mathrm{d}A = \\ &\int_{-\infty}^{A_{ik}} \frac{P(A|\boldsymbol{Y}_1)P(\boldsymbol{Y}_1)}{P(A)} P(A)\mathrm{d}A + \\ &\int_{A_{ik}}^{+\infty} \frac{P(A|\boldsymbol{Y}_2)P(\boldsymbol{Y}_2)}{P(A)} P(A)\mathrm{d}A = \\ &\int_{-\infty}^{A_{ik}} P(A, \boldsymbol{Y}_1)\mathrm{d}A + \int_{A\infty}^{+\infty} P(A, \boldsymbol{Y}_2)\mathrm{d}A.\end{aligned} \tag{25}$$

由式 (25) 可以得出，系统误差即为图 5 两个曲线交集部分。这种情况下音节韵律特征的分类和选取最终不需要知道概率分布的详细情况，所要求的只是一个数 A_{ik}，将其作为一个决策函数。

6　结论

本文采用统计中概率的方法对韵律建模思路进行了阐述，构筑了汉语音节的韵律模板，并建立了模板的选取算法和训练算法。同时本文利用该概率

模型，指出了韵律特征间的相互关联在韵律建模中的重要作用，以及对提高整句合成语音自然度的影响。采取韵律模板的思路，不仅较好地实现了汉语韵律的预测，在语音合成系统得到了成功的应用，也使语音合成系统的音库减小，使系统变得更灵活，运算效率加大，并使其在网络等需要高速运算的场合得到非常成功的应用。

新一代的语音合成系统正向着概念到语音或意念到语音的方向发展，情感语音合成的研究正变得越来越迫切，针对这些研究，其语料库的组成也越来越庞大，信息含量也变得更为丰富。语音合成需要逐步摆脱过渡依赖设计人经验，变得规范化，在这种发展趋势中，有关整个语音的研究可计算化，尤其是语音合成各组成核心模组的数值模型化，将会变得越来越重要。

参 考 文 献

1 Selkirk E. Phonology and syntax: the relation between sound and structure. Cambridge, MA: MIT Press, 1984

2 Katherine Morton. Adding emotion to synthetic to synthetic speech dialogue systems. ICSLP98, 1998: 675—678

3 Achim Mueller, Jianhua Tao, Ruediger Hoffmann. Data-driven importance analysis of linguistic and phonetic information. ICSLP2000

4 吴宗济. 普通话三字组变调规律. 中国语言学报, 1985(2)

5 沈 炯. 北京话上声连读的调型组合和节奏形式. 中国语文, 1994(4)

6 沈 炯. 汉语语调模型议. 语文研究, 1992; **4**: 16—24

7 沈 炯. 从轻音现象看语音与语法研究的关系. 吕叔湘等著，马庆株编《语法研究入门》，商务印书馆, 1999: 158

8 林茂灿, 颜景助. 普通话轻声与轻重音. 语言教学与研究, 1990(3)

9 陶建华, 蔡莲红等. 汉语 TTS 系统中可训练韵律模型的研究. 声学学报, 2001; **26**(1): 67--72

10 陶建华, 蔡莲红等. 基于统计模型的韵律建模方法. 第六届全国人机语音通讯学术会议论文集, 2001:212—216

11 林茂灿, 颜景助. 北京话轻声的声学性质. 方言, 1980(3)

12 Andrew J Hunt, Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. ICASSP 96

13 Bulyko I, Ostendorf M. Joint prosody prediction and unit selection for concatenative speech synthesis. ICASSP2001, 2001:781—784

14 Toda, Tomoki, Kawai, Hisashi *et al.* Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit. ICASSP2000, 2000: I/465—I/468

15 Chu M, Peng H *et al.* Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. ICASSP2001, 2001: 785—788

16 Wang Wern Jun, Campbell W N *et al.* Tree-based unit selection for English speech synthesis. ICASSP1993, 1993: 191—194