文章编号：1003-0077(2007)03-0122-07

支持重音合成的汉语语音合成系统

朱维彬

(北京交通大学 信息科学研究所，北京 100044)

摘 要：针对基于单元挑选的汉语语音合成系统中重音预测及实现，本文采用了知识指导下的数据驱动建模策略。首先，采用经过感知结果优化的重音检测器，实现了语音数据库的自动标注；其次，利用重音标注数据库，训练得到支持重音预测的韵律预测模型；用重音韵律预测模型替代原语音合成系统中的相应模型，从而构成了支持重音合成的语音合成系统。实验结果分析表明，基于感知结果优化的重音检测器的标注结果是可靠的；支持重音的韵律声学预测模型是合理的；新的合成系统能够合成出带有轻重变化的语音。

关键词：计算机应用；中文信息处理；重音；韵律模型；语音合成

中图分类号：TP391　　**文献标识码**：A

A Chinese Speech Synthesis System with Capability of Accent Realizing

ZHU Wei-bin

(Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China)

Abstract: To aim to predict and realize Chinese accent in a unit-selection based speech synthesis system, a data-driven method was used to build an accent-supported prosody module. First, with the help of Accent-Index detector which had been optimized with perceptual annotations, a speech corpus had been auto-annotated with Accent-Index. Then, a prosody predictive module supporting accent had been trained with the corpus. Replaced with the new prosody predictive module, the speech synthesis system could synthesize speech with various levels of accent. The results on the experiments had proved the accuracy of auto-detected accents, and the validity of the prosody predictor, and also the capability of accent realizing of the speech synthesis system.

Key words: computer application; Chinese information processing; accent; prosodic model; speech synthesis

1 引言

重音是实现语义上强调和聚焦的一种重要手段，对于语义传递的准确性具有不可或缺、有时甚至是决定性的作用。当前，语音合成系统技术路线的主流，是采用基于单元挑选的波形拼接方案，多利用层级结构来描述韵律结构特征。韵律层级结构配合声调信息，在相当大的程度上涵盖了汉语韵律特征的变化，合成语音的质量也达到了相当水平[1~6]。但其缺陷也是明显的：声音清晰但音色单调，语调平缓但一成不变，当然也没有为传递特定语义而有意为之的轻重变化。究其原因，是因为在现行系统的韵律模型中，没有关于重音的刻画。

实现支持重音的语音合成系统，其核心任务是建立支持重音的韵律模型。文献[7]中采用了 Fujisaki 模型，研究了宽/窄焦点对重音指令(Accent Command)的影响，而重音指令参数可以通过语音数据分析获得，但文中未能建立重音指令参数与焦点间的量化映射关系，也未涉及语音合成时重音指令如何预测。文献[8]中介绍了基于 Stem_ML (Soft TEMplate Mark-up Language)实现了语音波形中音节的韵律强度(Prosody Strength)的检测，并通过调整 Stem_ML 模型的韵律强度，使之可以产生逼近原始语音的音高曲线。两种方式中，无论是重音指令还是韵律强度都来自语音波形模型/参数

收稿日期：2006-12-25　**定稿日期**：2007-03-14

作者简介：朱维彬(1966—)，男，博士，研究方向为语音声学模型，韵律建模以及言语数据库构建等。

分析，和听觉重音级别没有直接的关联。且只解决了由韵律声学参数的预测，如何由待合成文本预测重音指令或韵律强度，没有涉及。

为了解决重音标注的感知关联性，及重音韵律预测模型的建模，我们采用的基本技术思路是，知识指导下的数据驱动韵律建模策略。首先，构建相当规模的带有韵律信息—包括重音信息—标注的语音数据库，采用经感知加权的重音检测器自动标注重音。其次，采用数据驱动技术建立韵律预测模型，模型由两部分构成：文本到韵律事件预测模块，和韵律事件到韵律声学参数预测模块。已经完成的工作包括：汉语重音标注系统定义；重音标注的实施；支持重音的韵律声学参数预测模块的实现。本文是对这部分工作的介绍，尤其侧重技术实现。

2　汉语重音标注系统定义及人工标注

从功能方面考虑，重音是对语流中特定成分的着重和强调，其范围可以跨越音节、韵律词、韵律短语等多个层级，甚至有特别突显声、韵、调的情形。对于汉语来讲，发音和感知的基本单元是音节，即便是突出声、韵、调，也是通过音节实现的；处于突显地位的韵律词或韵律短语，也并非每个音节都是重的[9]，而是取决于其中的重的音节。所以我们以音节为基本单元来考察各个层级的重音变化，同时也和现有数据库、韵律模型保持一致[1,10]。

无论是从发音、感知，还是其声学实现角度观察，从轻到重，重音的变化都是连续的。但为了和符号描述相对应，必须将重音级别离散化。综合考虑了实际语料及其应用，以及重音级别的韵律意义，在已有的韵律层级的基础上，我们设计了一套离散化的汉语重音指数（AI）用以描述实际语流中重音的变化。

2.1　重音标注系统定义

我们在两个层级上设定重音指数：一个是语调短语，另一个为韵律词。在短语层级上，重音指数是由韵律词来承载的，分为三个级别。在韵律词层级上，重音指数是由音节来承担的，也分为三个级别。具体的定义如下：

在短语层级上，重音指数是由韵律词来承载的，在我们的定义中被分为三个级别。

- 重，是重音的最高级别。对应着语义的焦点，感知为短语中被强调或突显的部分。
- 中，是重音的正常级别。对应着语义的一般表达成分，感知为短语中正常发音部分。
- 轻，是重音的最低级别。通常对应着短语中的联接或附属成分，感知为短语中轻/弱发音部分。

在韵律词层级上，重音指数是由音节来承担的，在我们的定义中也被分为三个级别。

- 重，是重音的最高级别。通常对应着重的韵律词中被重读的音节。
- 中，是重音的正常级别。可以是任意重音级别韵律词中的正常读出的音节。
- 轻，是重音的最低级别。可以是任意重音级别韵律词中的轻/弱读音节。另外，轻声音节暂时归为轻。

这样，对于数据库的韵律标注将包括以下内容：

- 韵律层级：语调短语、韵律短语和韵律词；
- 语调短语层面韵律词的重音级别：重、中、轻；
- 韵律词层面音节的重音级别：重、中、轻。

2.2　重音级别的人工标注

人工标注的目的，一方面是通过标注的过程修正并完善关于重音指数的定义；另一方面完成一定规模的数据库重音标注，用于重音自动检测系统性能的评测和系统参数的优化。

在实施重音级别标注时，综合利用了听觉、视觉信息。听觉上是语音录音的回放，同时还辅助以语音波形、基频曲线、语图等视觉信息。在判定重音级别时，须综合考察语音单元的发音饱和度以及突显度。

饱和度，是指发音到位、声调完整的程度。可以从音色是否饱满、音高是否达到声调的目标值加以判断。同时饱和度的变化，还会引起时长、音强的相应变化。

突显度，是指被考察语音单元与相邻单元存在差别的程度。主要体现在音域、音阶的差别，以及节奏的变化。有时突显度的增强还会引入附加的停顿。

在语调短语层级上，韵律词重音指数主要考察词一级的突显度和词内最强音节的发音饱和度，按如下约定：

- 重，饱和度和突显度两者至少有一超常，另一不低于正常水平。
- 中，饱和度和突显度两者都保持正常水平。
- 轻，饱和度和突显度两者至少有一低于正常

水平。

在韵律词层级上，音节的重音指数主要考察其发音饱和度，按如下约定：

- 重，发音饱和度超常。
- 中，正常发音。
- 轻，发音饱和度偏低，表现为音色央化和/或音高不到位。

基于以上关于重音指数的定义和标注约定，经过多次重复性实验，作者完成了60个句子的重音指数的手工标注。将最后完成的两次标注结果比较，韵律词和音节的重音指数相关系数分别达到了0.78和0.81，一致性令人满意，可以用于自动标注结果的评判和参数优化。

图1为一段语音的韵律标注例子。图中韵律层级(Break Index)标注符号BP2表示语调短语边界，BP1表示韵律短语边界，字符间的空格表示韵律词边界，BP0表示带有显著且单纯停顿的韵律词边界。语调短语层面韵律词的重音标记(Phrase AI)中，及韵律词层面音节的重音标记(Word AI)中，2、1、0分别对应重、中、轻。

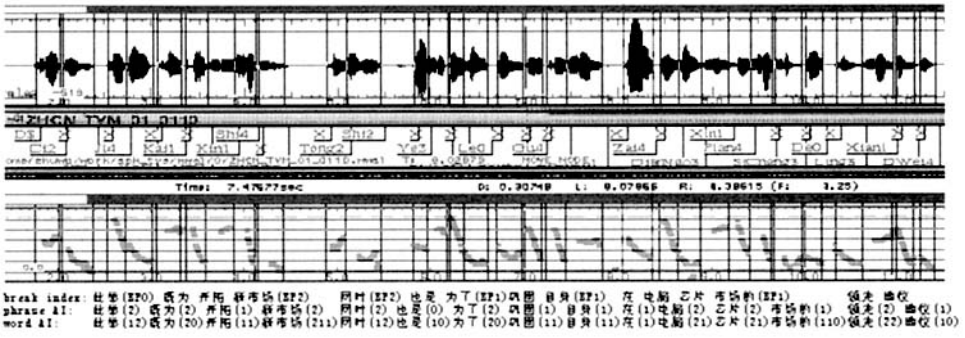

图1 发音内容为“此举既为开拓新市场同时也是为了自身在电脑芯片市场的领先地位”的韵律标注

3 重音自动标注

3.1 自动标注技术思路

为了解决重音标注的效率和一致性问题，我们提出了一套汉语重音自动判决方案[11]。该方案的基本思想是：首先，利用已有的韵律模型[1]，以韵律层级信息作为输入，韵律模型输出预测的韵律声学统计参数，其均值则可看作是轻重等同的“中性语调”的韵律声学实现。其次，将实际语音的韵律声学参数，和预测得到的“中性语调”韵律声学参数进行比较，其差值可以看作主要由轻重变化引起，进而可由差值推导出重音的级别。

各个韵律层级单位的韵律声学参数的差值对于音节/韵律词的重音指数的贡献显然会有不同。对于这些差值施以不同的权重并求和，并根据人工重音标注结果调整/优化这些权重系数，使得差值加权求和的数值与人工重音标注结果相匹配，如此我们就得到了经过听觉优化的重音自动检测器[12]，而差值加权求和的数值就是自动检测得到的重音指数。这样一种直接的模型结构，可以方便地加载经验性的知识，更为重要的是参数优化所需的训练数据量极少。

在重音自动检测器中，韵律声学参数差值对于重音指数的贡献，分解为突显度、发音饱和度两个方面。其中，突显度设定为韵律词整体性音阶和时长的变化，用韵律词平均音高和平均时长的实际值与预测值之差值加以度量，实际值愈大，突显度愈大。饱和度设定为对声调目标值的逼近程度，用音节的韵律声学参数，包括音节音阶、音高斜率和音节时长的差值来度量。由于目标值不同，在计算饱和度时，对于不同声调采用了不同的权重系数。重音指数是突显度和饱和度的加权和。

3.2 韵律声学参数

在韵律建模过程中，我们是以音节为基本单元来刻画韵律的声学变化的。在重音自动检测器中，每一个音节采用了三个韵律声学参数，包括音节音阶、音高斜率和音节时长。

音节音阶 $F0$。为了获得更为合理的音阶代表值，在计算时，我们采用音节能量质点处的音高表征该音节的总体性音高，即音阶的数值。音节的能量质点 k 采用以下公式计算：

$$k = \sum_{n=1}^{N} n \cdot s^2(n) \Big/ \sum_{n=1}^{N} s^2(n) \qquad (1)$$

其中 n 为按采样点计数，N 为音节的采样点，$s(n)$为语音波形幅度数值。音高采用对数频率。

音高斜率 Slope。因为在音节的后半段才能在最大程度上逼近音高的目标值[13]，所以我们取音节后半段的音高差与时长的比率作为该参数的数值。这里的音高也采用对数频率。

音节长度 Duration。为一个音节的长度，取对数。

最后，还是以音节为单位，基于数据库统计三个参数的均值与方差，对数据库中每个音节的韵律声学参数分别取 z-Score，进行归一化处理。计算公式如下：

$$y = (x - \mu)/\sigma \qquad (2)$$

其中 μ、σ 为参数 x 的均值与方差，y 为 x 的 z-Score。三个参数本身的声学意义较为直观；对参数取对数是为了参数数值与人的听觉特性更加匹配；取 z-Score 后，参数在数值分布上得到归一。

3.3　重音检测器的优化

为了解决重音自动检测器的权重参数的优化，同时对自动标注结果进行评估，我们引入了自动检测结果与手工标注结果的相关系数作为客观指标。

在实验中，我们检测了 60 个句子中音节和韵律词的重音指数，并计算其与手工标注结果的相关系数。将相关系数乘以-1，这样就构成了一个以权重系数为变量、负相关系数为输出函数。利用 Matlab 优化工具箱内提供的工具 Fminsearch 来搜索该函数最小值所对应的变量取值，就可以得到相关系数最大时的一套权重系数，即重音检测器优化的权重系数。从而实现了利用听觉感知结果对重音自动检测器的优化。

表 1 为优化前后，自动检测结果与手工标注结果相似度的比较。韵律词重音指数相关系数 CC_word 由 0.62 提高到 0.77，音节重音指数相关系数 CC_syllable 由 0.66 提高到 0.80，优化效果显著。

表 1　优化前后相关系数的变化

	初始值	优化后
CC_Word	0.62	0.77
CC_Syllable	0.66	0.80

图 2 展示了发音内容为“不少顾客专门借买糖来看看这位一团火的传人”语音实施重音自动检测的一个实例。图中三个韵律声学参数，音节音阶 F0、音高斜率 Slope、音节长度 Duration 分别显示。

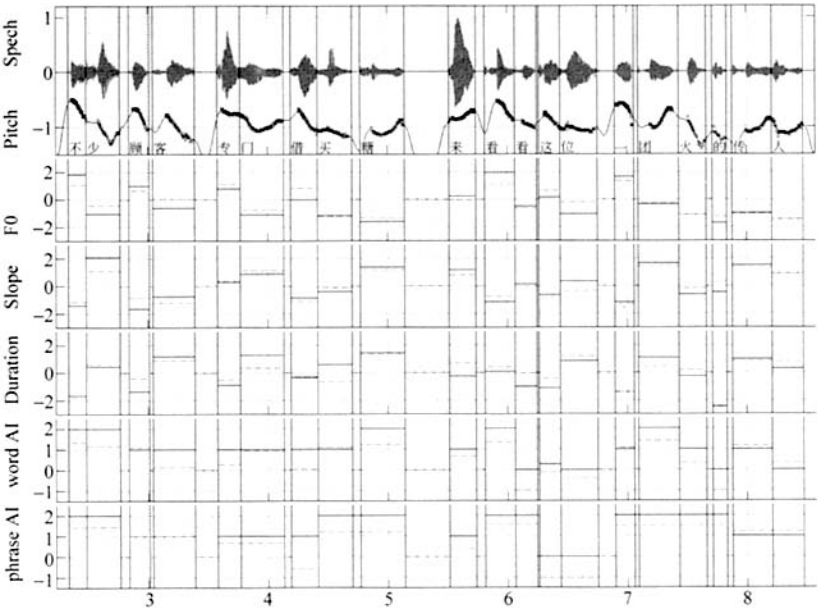

图 2　基于 F0、Slope、Duration 的 AI 自动检测

其中，实线为原始语音信号中的提取值，虚线为利用层级信息预测的韵律声学统计参数中的均值。图中Word AI项为韵律词内音节重音指数，Phrase AI项为语调短语内韵律词的重音指数，其中实线为人工标注结果，虚线为自动检测数值(数值离散化之前)。结果显示了自动检测结果与人工标注结果相当吻合。

利用优化后的汉语重音检测器，我们完成了对一个规模为5 000句发音数据库的重音自动标注。

4 重音在基于单元挑选技术的语音合成系统中的实现

4.1 实验系统平台

图3为一语音合成系统的原理框图，该系统由两个主模块组成：前端与后端；前端完成由输入文本到发音符号描述的分析，后端完成由发音符号描述到语音波形的输出；其间的接口—语境信息包括了注音符号、韵律事件符号等对发音的符号描述。这是一个基于单元挑选的、以韵律为选择导向的实现方案，其中的韵律事件预测模块、韵律声学预测模块是系统的核心，都是采用数据驱动的原理，利用标注数据库进行训练得到的统计模型[1,14]。

基于重音标注数据库，我们仍然采用决策树-GMM(Gaussian Mixed Model)方案[1]，训练了一个支持重音的韵律声学参数预测概率模型。将原系统中的韵律声学参数预测概率模型用新的重音模型替代，从而构成了一个基于数据库的波形拼接语音合成系统，可用以检验重音标注和重音韵律声学参数预测概率模型的有效性。

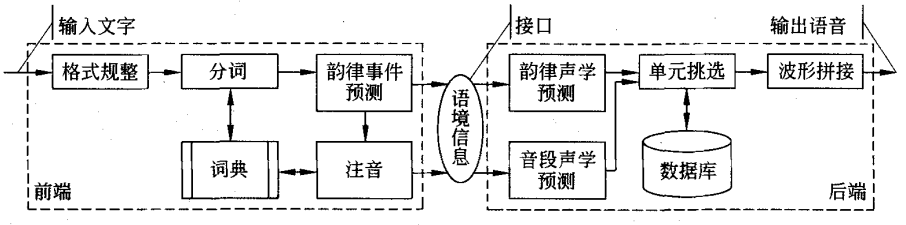

图3 基于单元挑选的TTS系统结构

4.2 重音韵律声学预测模型构建

在构建韵律声学预测模型时，模型的输入是合成基本单元的语境信息，输出的是以概率形式表示的合成单元韵律声学参数预测。与重音检测器不同，在合成系统中，设置了两类韵律声学预测模型。除了重音检测器采用的，针对每个基本单元也采用的目标模型外，还设置了反映相邻基本单元相对变化的转移模型。韵律声学参数的设置也有所改变，采用了4个音高值加1个音节时长值，构成目标模型声学参数；转移模型中的声学参数则由2个前后音节音高差值加1个前后音节时长差值构成。而每个音节的音高值是按照等间距，在音节归一化时长0.125，0.375，0.625，0.875处采样得到的。2个音高差值分别是当前音节0.125处与之前音节0.875处，当前音节0.375与之前音节0.625处的音高差值。

与原有韵律声学参数预测概率模型[1]的差别在于：输入信息中关于语境描述除了注音、声调、韵律层级结构之外，新的模型中还增加了AI信息。相应的，在决策树的问题集中，也增加了关于AI的内容。修正后支持重音的决策树问题集中，包括了以下的内容：

- 声调类：当前音节声调，左音节声调，右音节声调；

音位：当前音节音位类型(声/韵母类型)，左语境音位类型(韵母类型)，右语境音位类型(声母类型)；

- 边界：当前音节左边界类型，当前音节右边界类型，韵律词中的音节长度及位置，韵律短语中的韵律词个数及位置；

- 重音：当前音节韵律词内重音级别，当前韵律词短语内的重音级别。

我们利用自动重音标注数据库，分别训练出了韵律声学参数预测目标概率模型，及转移概率模型。

由文本到韵律事件符号描述预测，是韵律模型中的另一关键任务。重音级别轻重的变化，反映了语义的变化。而语义的预测，即便是字面意义，仅靠句子层面的分词结果、词性是无法充分解决的，我们

的尝试性实验，即，仍只利用分词及词性训练重音级别预测模型，其预测结果召回率在 50% 以下。这也说明了，要完全解决重音事件预测，须更深入的文本分析，须结合句法分析及句子层面之上的语义分析的结果。

由于文本分析模块暂时还不具备输出重音标记的能力，为了合成具有重音变化的语音，所需的重音信息还须人工设置。

4.3 支持重音的韵律模型预测误差分析及合成实验

实验一，预测误差分析

我们首先比较了重音信息对韵律声学参数模型预测数值的影响。因为需用到直接的预测数值，而合成系统中既有目标模型，又有转移模型，输出的都是韵律参数的概率分布，且相互约束，无法方便地得到直接表示的最优的预测值，所以这里仍采用音节音阶、音高斜率和音节时长三个声学参数的韵律预测模型。模型采用决策树＋单高斯结构，决策树的每个叶子关联一个单高斯函数，其均值就作为预测值。我们利用了 5 000 句的数据库，在决策树训练时，分别加载或屏蔽重音信息，得到了两个预测模型。

我们进行了数值分析，并有以下结论：

• 重音模型决策树的参数分布更加紧致，说明重音韵律声学参数预测模型较非重音模型更加精细。在其他训练条件保持一致的前提下，重音模型的叶子数目为 542，而非重音模型只有 418。

• 重音韵律声学参数预测模型预测更加准确。利用 5 000 句数据库的韵律标注信息作为输入，分别利用重音模型与非重音模型得到了预测值，然后分别统计两组预测值相对于语音实际值之间的误差。表 2 显示了三个参数的误差方差，重音模型较非重音模型相对减少了 10～16%。

表 2　重音/非重音模型预测误差方差

	F0	Slope	Duration
非重音模型	0.54	0.51	0.60
重音模型	0.49	0.43	0.54

实验二，重新合成测听

我们采用重音韵律声学参数预测目标概率模型及转移概率模型，替代语音合成系统中的对应模块，从而构成了支持重音合成的语音合成系统。我们选取了数据库中的部分语句，将其从波形样本候选集中剔除，另行构成一测试集。利用测试集中的语句标注信息作为合成系统的输入，由合成系统重新合成语音。我们选取了其中的 12 个句子，对原始录音、重音模型合成语音、非重音模型合成语音进行 5 分制自然度 MOS 测试。MOS 结果为见表 3。

表 3　原始录音与合成语音 MOS 得分

	原始录音	重音合成	非重音合成
MOS	4.9	4.5	4.2

表中结果显示，采用重音模型较非重音模重新合成语音的自然度要高。

为了反映重音模型的重音体现的准确性，作者对这 12 句由重音模型合成的语音进行了人工重音标注。仍然采用多次重复标注，最后两次韵律词和音节的重音指数相关系数分别达到了 0.86 和0.91，然后用最后一次人工标注的结果与合成时输入的自动标注重音指数进行相关比较，结果见表 4。

表 4　合成输入重音与合成语音感知重音的相关系数

	输入/感知	两次感知
CC_Word	0.71	0.86
CC_Syllable	0.76	0.91

结果显示，合成语音人工感知的重音指数与合成输入的重音指数有着较强的相关性，韵律词间的重音指数相关系数为 0.71，音节间的为 0.76。结果说明，重音模型能够较为准确地合成指定的重音级别；也就说明了，模型建模方法也是合理的，对于数据库实施的自动重音标注是有效的。

实验三，重音调整试验

通过人为设定重音的位置及级别，我们合成了一定数量的例句，用以验证重音韵律模型的有效性。图 4 中显示的是四个内容为“催眠师 有 相当的 威望”合成语音的基频曲线图。在语调短语层级上，以韵律词为单位，从“催眠师”到“威望”依次设定为最高重音级别，其他为正常重音级别，从而得到四个合成样本的基频曲线。

图中被标注为重音的韵律词的基频变化范围较其他时刻有了明显的扩张，和我们事先所掌握的重音的声学特征是一致的[13]。样本测听同样表明了重音实现是显著的。这就更直接证明了新的合成系统可以有效地合成出带有重音变化的语音。

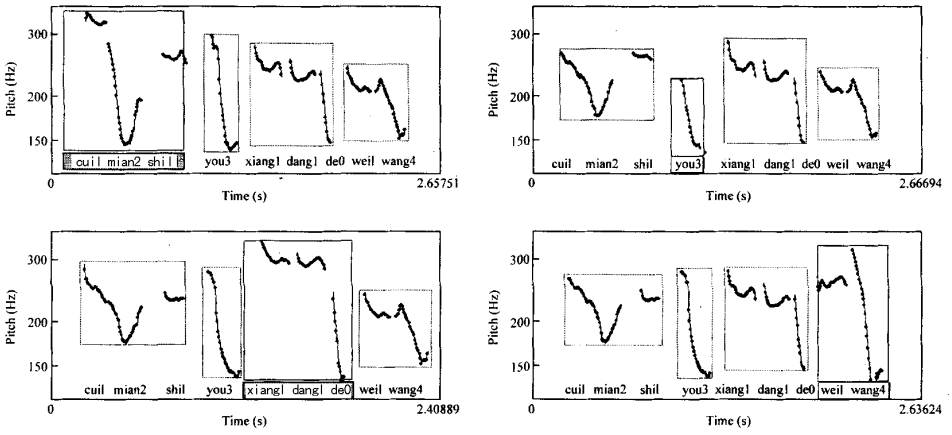

图4 重音标注调整得到的合成语音音高曲线，发音内容为“催眠师 有 相当的 威望”

5 总结

本文介绍了支持重音合成的汉语语音合成系统的技术路线。实现了对汉语语音数据库重音的自动标注，利用数据库构建支持重音信息的韵律预测模型，以替代原有语音合成系统中的非重音模型，从而构成了新的合成系统。实验结果表明：基于感知优化的重音检测器的标注结果是可靠的；采用决策树方案的重音韵律声学预测模型是合理的；新的合成系统能够合成出带有轻重变化的语音。

已完成的重音韵律声学参数预测模型只是整个重音韵律模型的一个部分，如何从文本中分析预测出重音事件，实现带有重音描述的韵律事件预测模型，将是我们所面临的更大挑战。

参考文献：

[1] Ma Xijun, Zhang Wei, Zhu Weibin, et al. Probability Prosody Model for Unit Selection [A]. In: Proc. of ICASSP 2004 [C]. Montreal, Canada, 2004. 649-652.

[2] Chu Min, Peng Hu, et al. Selecting Non-Uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer [A]. In: Proc. of ICASSP 2001 [C]. Sale Lake City, USA, 2001.

[3] 初敏. 自然言语的韵律组织中的不确定性及其在语音合成中的应用 [J]. 中文信息学报，2004，18(4)：66-71.

[4] 陶建华，赵晟，蔡莲红. 基于统计韵律模型的汉语语音合成系统的研究 [J]. 中文信息学报，2002，16(4)：1-6.

[5] 吴志勇，蔡莲红. 语音合成中的韵律关联模型 [J]. 中文信息学报，18(2)：44-50.

[6] 吴晓如，王仁华，刘庆峰. 基于韵律特征和语法信息的韵律边界检测模型 [J]. 中文信息学报，2003，17(5)：48-54.

[7] Chen Gaopeng, Hu Yu, Wang Renhua, et al. Quantitative Analysis and Synthesis of Focus in Mandarin [A]. In: Proc. of TAL 2004 [C]. Beijing, China, 2004. 25-28.

[8] Greg Kochanski, Chilin Shih, et al. Hierarchical Structure and Word Strength Prediction of Mandarin Prosody [J]. International Journal of Speech Technology, 2003,6(1): 33-43.

[9] 王韫佳，初敏，贺琳. 普通话语句重音在双音节韵律词中的分布 [J]. 语言科学，2004，3(5)：38-48.

[10] Zhu Weibin, Shi Qin, et al. Corpus Building for Data-Driven TTS Systems [A]. In: Proc. of IEEE TTS Workshop 2002 [C]. Santa Monica, USA, 2002. 199-202.

[11] Zhu Weibin, Zhang Wei, Shi Qin, et al. Automatic Detection of Chinese Accent-Index Based on Approximation-Ratio [A]. In: Proc. of ISCSLP 2004 [C]. Hong Kong, China, 2004. 85-88.

[12] Zhu Weibin. Perceptual Optimization of the Chinese Accent-Index Detector [A]. In: Proc. of Speech Prosody 2006 [C]. Dresden, German, 2006.

[13] Xu Yi. Effects of tone and focus on the formation and alignment of f0 contours [J]. Journal of Phonetics, 1999, 27: 55-105.

[14] Shi Qin, Ma Xijun, Zhu Weibin, et al. Statistic Prosody Structure Prediction [A]. In: Proc. of IEEE TTS Workshop 2002 [C]. Santa Monica, USA, 2002.