ISSN 1000-0054 清华大学学报（自然科学版） 2013 年第 53 卷第 7 期
CN 11-2223/N J Tsinghuauniv (Sci & Tech), 2013, Vol.53, No.7 2036-2038

基于 HMM 语音合成的语调控制

王永鑫，贾珈，张雨辰，蔡莲红
（清华大学计算机科学与技术系，清华信息科学与技术国家实验室（筹），普适计算教育部重点实验室，北京 100084）

摘 要：语调是语音分析和合成领域关注的重要课题，可计算的语调模型是实现语调控制的关键。本文分析了大规模语句的音节音高在句子中的变化，归纳了语调模式。陈述语调主要表现为基调的升降和音高下倾；疑问语调主要特点是疑问焦点的音高提升和调型变化。本文提出了一种陈述语调归一化描述方法，以及疑问语调调型差异模型。利用基于隐马尔可夫模型的语音合成系统的控制机制，实现了对语调的控制。试验表明，基于陈述句语调归一化描述方法模拟了陈述句语调的变化，基于疑问语调调型差异模型实现了陈述到疑问语调的转换。语调控制使合成语音的表现力得到了提高。

关键词：语调控制；语调；基于隐马尔可夫模型(Hidden Markov Model, HMM)语音合成；语调模型
中图分类号：TP 3

Control of Intonation in HMM based TTS System

WANG Yongxin, JIA Jia, ZHANG Yuchen and CAI Lianhong
(Key Laboratory of Pervasive Computing, Ministry of Education; Tsinghua National Laboratory for Information Science and Technology (TNList); Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Intonation is an important research topic in speech analysis and synthesis, and a computable intonation model is the key to achieve the control of intonation in synthesis. In this paper, we analyze the pitch movement of syllables in a sentence, and reveal the intonation patterns in Chinese sentences. The downtrend of the pitch is the main pattern of declarative sentences; pattern of interrogative sentences include the raise of pitch and more elaborated pitch contours. We proposed a unified intonation representation, and intonation pattern difference model for interrogative and declarative sentences. Within the framework of HMM speech synthesis, we realize the control strategies for sentence intonation. The experiment shows that the proposed unified intonation representation can analog the different declarative intonations, and with the intonation pattern difference model the conversion of interrogative intonation from declarative intonation can be realized. More expressive speech can be generated with the control of intonation.

Key words: control of intonation; intonation; HMM speech synthesis; intonation model

狭义语调指语句音高的变化。在汉语这一声调语言中，音高承载了声调和语调信息。赵元任提出了“代数和”理论表示汉语中声调与语调的相互叠加的关系[1]。Fujisaki提出重音和短语以对数形式相加的语调模型[2]，许毅提出平行叠加模型(Parallel Encoding and Target Approximation, PENTA)[3]。Shih研究了陈述句的音高下降，指出影响语句音高曲线的因素很多，包括下倾、降阶、边界下降、重音、声调、语调类型等，且下倾、降阶等对音高曲线的影响是全局的，但未给出定量描述[4]。黄贤军认为陈述句低音线呈现以韵律短语为基本单元的下倾，且声调组合、音节在韵律词中的位置不同，低音线下倾的斜率也不同[5,6]。这是因为使用音节的音高直接描述语调。音节的音高承载了语调信息，但不等于语调。在语言运用中，会根据表达需求选用不同的语调模式，从而使语言更富有表现力，如较高的语调可以用于应答，较低的语调可以用于求证等[7]。

在基于隐马尔可夫模型(Hidden Markov Model, HMM)语音合成中，语调调型被隐含在了基频模型中。句子的基频曲线根据当前音节的拼音、声调与上下文环境，由语料库中的基频统计模型得到[8]。合成语音的语调可以认为是语料库中语调的平均。这种基频建模方式，生成的语调趋于平淡，不能生成指定的语调模式。为改善 HMM 语音合成中参数过平滑问题，有研究使用全局方差的方式增加合成语音的变化，改善了合成语音的音质，但对语调的影响不大[9]。另外有研究对合成语料库进行层次建模，使合成语音中的语调过平滑的现象得到改善[10]。但这一方法未能实现对合成语音语调的有效控制。

本文在分析的基础上，设计了一种陈述语调归一化描述方法，建立了一种疑问语调调型差异模型，利用 HMM 语音合成系统的声学参数控制机制，实现对合成语音语调的控制，从而可以让 HMM 语音合成系统生成的语调富于变化，语音表现力更强。

1 陈述语调模式

陈述语句用于表事。通常认为陈述句的音高曲线存在下倾。而下倾的定义，下倾参数的描述却各不相同，如采用语句中各上声或阳平音节的低音点的连线表示语调。但当语句中这些音节所处韵律位置不同，或未出现该声调音节时，对语调的分析和描述会遇到困难。本文利用统计的方法，在大语料库中对陈述句中不同声调音节的音高变化进行分析，并在此基础上分析了不同的声调音节音高之间的相互关系，提出了一种陈述语调的归一化描述方法。

本文所使用的语料库包含由一名女性专业播音员录制的两万句新闻风格录音语料，句子平均长度为 16 个音节，每句包含 2～4 个韵律短语。语料库标注了音节边界、基频、韵律词与韵律短语边界。

本文抽取基频均值、低音点等音高特征，并选用阴平音节的音高均值，阳平、上声、去声音节的音高曲线最低点为音高特征值，并转换为半音值：

*收稿日期：2013-04-27
基金项目：国家“九七三”重点基础研究项目(2013CB329304)，国家“八六三”高技术项目(2012AA011602),国家自然科学基金(61003094)
作者简介：王永鑫（1982－）,男（满族）,河北,博士研究生。
通信作者：蔡莲红，教授，E-mail: clh-dcs@tsinghua.edu.cn

$$P = 12\log_2(F0/F0_{\text{ref}}) \quad (1)$$

半音值P是基频$F0$的一种对数表示。$F0_{\text{ref}}$为计算半音值时选取的参考基频，本文选取$F0_{\text{ref}}$=100 Hz。

本文计算了相同韵律位置、相同声调音节的音高特征值的均值，并比较了音节音高特征值在韵律词、韵律短语、语句中的表现。

分析结果表明，语调单元（韵律词、韵语律短语、语句）中，末音节的音高特征值低于首音节的音高特征值；在韵律短语中，韵律词末音节的音高特征值随韵律词在韵律短语中的位置后移而降低；在语句中，韵律短语末音节的音高特征值随韵律短语在语句中的位置后移而降低；这表明韵律结构末音节承担着韵律结构分界的作用。韵律短语中韵律词末音节音高变化模式如图 1 所示。图中将韵律短语的长度进行了归一化。图中以不同线型表示韵律词末音节的不同声调，而以不同的标记表示不同的韵律短语长度。

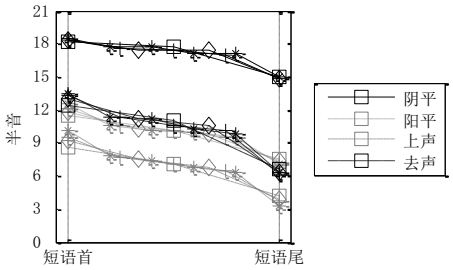

图 1 韵律词末音节在韵律短语中音高变化

从图 1 中可以看出，韵律短语调型的中部呈下降趋势，且总体趋势一致。次首韵律词有额外的下降，表明首韵律词承载调首功能。末韵律词有额外的下降，凸显陈述语调下倾功能，并起到标志韵律短语边界的作用。不同声调所承载的调高不等。根据以上的分析，可以定义陈述韵律短语调描述如下：

$$P_I = \alpha_t P_{I0} - \beta x \quad (2)$$

式中P_{I0}表示基调，α_t表示与声调相关的基调修正系数，t表示不同的声调。β表示韵律短语中部的降阶指数，x表示韵律词在韵律短语中的归一化位置。P_I则表示以音节的音高特征值描述的韵律短语调型。

基调P_{I0}与发音人、情感、强调等因素有关。本文选取首韵律词的去声末音节的音高特征值作为基调的调高，此时$\alpha_4 = 1$。如首韵律词的末音节为其它声调，由语料库分析可得其修正系数α_t。

降阶指数β，即韵律短语内韵律词音高的下降速率。本文分析表明，韵律短语平均下倾约为 6 个半音(0.5 倍频程)。音节声调调型的音高范围约为 1 倍频程，整个韵律短语音高分布在 1.5 倍频程范围内。图 1 中声调不同的音节所代表的语调下倾有所不同。

基于式(2)的语调归一化表示可用来分析或标注语调。首先标注韵律词的去声末音节的音高特征值，然后根据基调修正系数，将其它声调的韵律词末音节音高转化为相当于去声的归一化音高，最后连接这些归一化音高就得到了语调基调曲线。图 2 给出了一个归一化音高在韵律短语中变化的实例。

本文的研究表明，陈述语句的调型可以由基调P_{I0}、韵律短语下倾β、句子下倾等参数描述。再辅以基调修正系数α_t，给出了陈述语调的归一化描述方法。本方法可以对陈述语调进行分析，也可以通过这些参数的调整实现对语调的控制，以实现汉语中特定模式的语调的生成，提高合成语音的表现力。

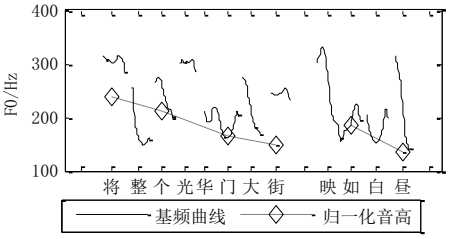

图 2 基频曲线与韵律词归一化音高曲线

2 疑问语调

疑问语句用于提问。通常认为疑问句中疑问焦点的音高上升。本节对比了疑问焦点在疑问句与陈述句中的调型变化，基于疑问语调与陈述语调的调型的差异，建立了疑问语调调型差异模型。基于疑问语调调型差异模型，利用 HMM 语音合成中的基频控制机制，可以在陈述句合成系统的基础上实现疑问语句的合成。差异模型可以通过少量的训练语料生成，并可以应用到不同发音人训练的合成系统中。

本文自行设计并录制了同文本的陈述与疑问短句 256 句。短句包含完整的主谓宾结构，无疑问标志，每句三至四个音节，疑问焦点位于句末。标注了语句的音节边界与基频曲线。

本文比较了疑问句与陈述句中对应音节的调形。疑问焦点处音节的对比如图 3 所示。从图 3 中可以看到，疑问焦点处的音节与陈述句中的相比，基频有所提升，调型也表现得更加充分。这主要是由于疑问焦点音节同时承载疑问与焦点的两重信息。疑问信息的表达引起了音节基频的提高，而焦点信息的表达使音节的调形表达得更加充分。两种信息的叠加使得疑问焦点处的音节相对陈述句音节调形表现出一种非线性的提高。

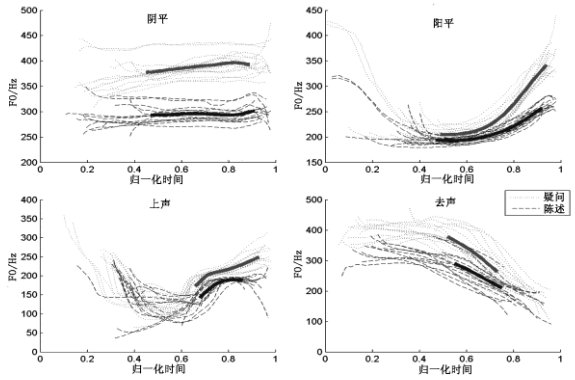

图 3 不同声调调形在疑问焦点处与陈述句中的对比。图中粗实线为调形核心段的平均值

语句中各音节的声调曲线级联形成语调。对比陈述语调与疑问语调的调型，可以发现疑问语调的整句基调有所提升。在句末疑问焦点处，音高提升和能量增强最大，距离疑问焦点越远变化越小，与英语焦点分析结果相似[11]。

本文建立了疑问语调调型差异模型，以疑问句

与陈述句中的对应音节音高曲线的差异表示调型的差异，并使用 HMM 模型对其进行建模。

在计算差异时，首先将疑问句与陈述句中的对应音节的时长均归一化为 N 个点。疑问句与陈述句对应音节的调型曲线分别为 $P_{\text{ques}}(n)$ 与 $P_{\text{decl}}(n)$。疑问句与陈述句对应音节的差异 $P_{\text{diff}}(n)$ 可表示为：

$$P_{\text{diff}}(n) = P_{\text{ques}}(n) - P_{\text{decl}}(n), \quad 1 \le n \le N \qquad (3)$$

当某时间点陈述句或疑问句中基频不存在时，该点差异值也不存在。$P_{\text{diff}}(n)$ 中包含了疑问句音节与陈述句音节之间的调型差异与音高差异。

疑问语调调型与陈述语调调型的差异由各个音节调型的差异组成。本文使用了多空间分布 HMM（MSDHMM）模型[12]对语调调型差异曲线进行了建模。在 MSDHMM 模型中，输出概率密度函数包含多个不同维度的概率密度函数，每个概率密度函数描述一个空间，每个空间有自己的权重。在音高差异曲线模型中，输出概率密度函数包含两个空间，音高差异存在时的 1 维空间，与音高差异不存在时的 0 维空间。模型对 MSDHMM 中每个状态的概率密度函数进行了决策树聚类，以实现对音高差异参数的预测。决策树聚类所使用的特征主要为上下文语境特征，包括声调、前后音节声调、韵律位置、当前音节与疑问焦点的相对位置等。

疑问语调调型差异模型对语调的调型在疑问句与陈述句的差异进行建模。使用 MSDHMM 模型，可以在描述调型差异随时间的非线性变化的同时，减少清音段对音高差异建模带来的影响。

利用疑问语调调型差异模型，可以进行陈述句与疑问句的调型差异预测，并据此由陈述句的音高变化曲线与调型差异生成同文本的疑问句音高变化曲线，从而可以通过语调控制的方式生成疑问语调。相对于对疑问语调的音高曲线直接建模，差异模型可以从较少语料得到疑问语调与陈述语调的调型差异，并且这一差异可以应用于不同的发音人中。

3 HMM 语音合成中的语调控制

在 HMM 参数化语音合成系统中，语料库中的语音被分解为基频与频谱参数，然后通过 HMM 建模与决策树聚类的方式对基频与频谱参数分别建模。在合成过程中，根据上下文语境信息在决策树中选取合适的 HMM 状态参数，并以输出概率最大的准则生成基频与频谱参数，接着通过源-滤波器模型生成语音，如图 4 左侧所示[8]。本文利用这一系统框架，在生成语音之前，对预测生成的基频等声学参数进行修改，实现了对语调的控制。图 4 右侧给出了基于 HMM 语音合成的语调控制框架。

在图 4 所示的基于 HMM 的语音合成中的语调控制框架下，本文实现了下面两种语调控制模型。

1）基于基调与降阶指数对陈述句语调的控制。

语调在语音中具有丰富的表达功能，它通过基调的升降、调型的变化以及节奏重音等实现传情达意。例如提高陈述句的基调可以改变语势，调整韵律短语的调域可改变节奏感，调整语句的降阶指数可改变语气表达。本文基于陈述语调归一化描述方法，通过对基调与降阶指数的控制，可以使得生成的语调更加丰富，合成语音更富有表现力。对基调的控制可以通过对句子音高整体抬升或降低的方式实现。对基调的修改，可以将基调按一定的比例改变，或者直接指定基调的高度。本模型中的基调变化比例或基调指定值是指相对首韵律词的去声末音节的音高特征值而言，若遇其它声调可根据式(2)中的基调修正系数 α_t 进行换算。

对降阶指数的控制需要对句子中每个音节的基频曲线进行修改。在韵律短语调中，降阶指数表示韵律词音高在韵律短语中的下倾速率。对一个含有 N 个韵律词的韵律短语，当其降阶指数变化为 $\Delta\beta$ 时，第 $n(1 \le n \le N)$ 个韵律词的归一化音高变化量为：

$$\Delta P(n) = (n-1)\Delta\beta/(N-1) \qquad (4)$$

若直接设置降阶指数为 β，则各个韵律词的音高特征值直接由式(2)给出。

在对降阶指数进行控制时，通过对韵律词的音高进行整体控制的方式控制韵律词归一化音高，不改变音节声调的调型。

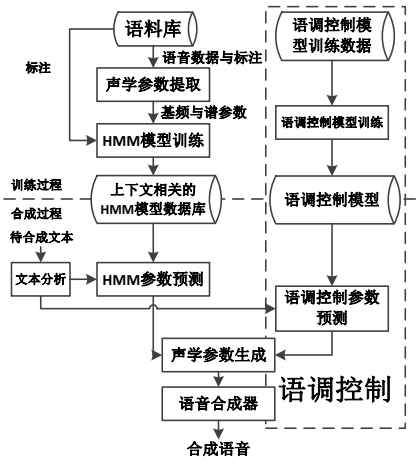

图 4 基于 HMM 的语音合成及语调控制流程图

2）基于疑问语调调型差异模型实现疑问句的合成。

疑问语调调型差异模型对疑问语调与陈述语调的调型差异进行建模，可以根据合成音节及上下文语境预测疑问语调与陈述语调的音高差异值。在此调型差异的基础上，利用语调控制机制，可以将陈述语调变换为疑问语调，实现疑问语调的合成。

应用差异值模型实现对语调的控制需要经过时长归一化、应用差异、后处理等步骤。

差异值模型预测得到的基频差异曲线的音节时长与合成语音的音节时长通常有所不同，因而需要进行时长归一化。归一化通常采用插值的方法。

应用差异值时，将合成音节音高曲线上的每个音高点分别应用预测得到的差异值。由于音高曲线与音高差异曲线在清音段无定义，因而在应用差异值模型时只处理两者都有定义的声调核心段部分，而对其它部分进行后处理。

后处理主要处理调型的过渡段。这一部分音高变化，可由前音节差异值曲线末尾的基频变化与当

前音节差异值曲线起始的基频变化经线性插值得到，同时要对相邻音节间的基频曲线进行平滑。

4 试验与结果

语音合成时使用 3 000 句陈述句作为训练语料，使用 HTK2.2 进行训练与合成。在 HMM 语音合成系统对声学参数控制机制的基础上，实现了对语调的控制，包括基于基调与降阶指数对陈述句语调的控制，以及基于疑问语调调型差异模型通过语调控制的方式实现疑问句合成。

基于基调与降阶指数可以对合成语音的语调进行控制，以生成某些特定的语调，使合成系统的语调更加丰富，合成语音的表现力更强。图 5 给出了加入陈述语调控制前后的合成语音的基频曲线与韵律词末音节归一化音高对比图，其中对语调控制包括基调的提高与降阶指数的增加。加入语调控制之后，可以使合成语音不再拘泥于模型中所生成的平均语调，而可以模拟自然语音中丰富的语调变化，从而提高合成语音的表现力。

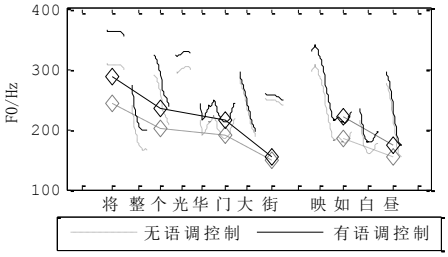

图 5 加入音高下倾补偿前后的音高变化对比

基于疑问语调调型差异模型，可以通过语调控制的方式，利用已有的陈述语句合成系统，生成疑问语调。本文在前述陈述句语音合成系统中，加入了基于疑问语调调型差异模型的语调控制，实现了疑问语调生成。试验中，训练疑问语调调型差异模型所使用的语料库的录音人与训练陈述句合成系统所使用的语料库的录音人为两名不同的女性录音人。合成疑问句时所使用的文本不含疑问标记。

本文通过主观试验的方式对合成结果进行了评价。6 句合成文本分别使用未加入与加入了疑问语调调型差异模型的系统进行合成，并以随机的顺序播放给 12 位听音人。听音人对所听到的句子为陈述句或疑问句进行判断。应用疑问语调差异模型之后的合成语音有 73.6%被判断为疑问句。可见疑问语调差异模型可以较好地描述汉语疑问句的语调，并可以实现陈述到疑问语调的变换。

5 结论

本文在陈述句与疑问句的语料库上，抽取了基频均值、低音点等音高特征，分析了陈述句中音节音高在句子中的变化，及疑问句疑问焦点音节的调型与陈述句的差异。

对大规模陈述句语料库的分析发现，不同声调的音节在韵律短语中表现为较为一致的线性下倾，这种语调的变化可以用基调与降阶指数等参数来描述。通过对不同声调基调的归一化，本文提出了一种陈述语调的归一化描述方法。该方法可以用于对含有不同声调音节的句子的语调变化进行分析。同时，可以通过基调与降阶指数等语调参数实现对语调的控制，模拟陈述句语调中的不同变化，使合成语音的语调更加丰富，表现力更强。

对疑问句的分析发现疑问焦点音节的调型与陈述句相比，基频有所提高，调型实现得更加充分。这是由于疑问焦点音节同时携带了疑问与焦点两重信息。本文建立了疑问语调差异模型，对疑问语调与陈述语调的调型差异进行建模，语调的调型差异由语句中各个音节的调型差异组成。利用这一模型，可以实现陈述语调到疑问语调的变换。

利用 HMM 语音合成中的音高控制机制，本文实现了基于陈述语调的归一化描述方法与疑问语调调型差异模型的语调控制。实验表明，陈述语调的归一化描述方法可以模拟汉语陈述语调的不同变化，基于疑问语调调型差异模型可以实现陈述语调到疑问语调的变换。加入语调控制后，合成语音的表现力得到了提高。后续的工作将进一步优化语音合成系统中语调控制模型，提高合成语音的表现力。

6 参考文献

[1] 赵元任. 汉语的字调跟语调[G]//赵元任语言学论文集. 北京:商务印书馆, 2002.
CHAO Yuenren. Tone and intonation in Chinese[G]//Linguistic Publications of Chao Yuenren. Beijing: The Commercial Press, 2002. (in Chinese)

[2] Gu W, Hirose K, Fujisaki H. Modeling the effects of emphasis and question on fundamental frequency contours of Cantonese utterances[J]. *IEEE Transaction on Audio, Speech, and Language Processing*, 2006, **14**(4): 1155–1170.

[3] Prom-on S, Xu Y, Thipakorn B. Modeling tone and intonation in Mandarin and English as a process of target approximation[J]. *The Journal of the Acoustical Society of America*, 2009, **125**(1): 405–424.

[4] Shih C. Declination in Mandarin[C]//ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications. Athens, Greece: 1997: 293–296.

[5] 黄贤军, 杨玉芳, 吕士楠. 韵律短语的音高下倾实验研究[C]//第八届全国人机语音通信学术会议论文集. 北京: 2005: 360–364.
HUANG Xianjun, YANG Yufang, LV Shinan. Experimental study on the downtrend of prosodic phrase[C]//National Conference on Man-Machine Speech Communication. Beijing, China: 2005: 360-364. (in Chinese)

[6] 黄贤军, 高路, 杨玉芳, 等. 汉语语调音高下倾的实验研究[J]. 声学学报(中文版), **2009**(02): 158-166.
HUANG Xianjun, GAO Lu, YANG Yufang, et al. Experimental study on declination in Chinese intonation[J]. *Acta Acustica, Chinese Version*, **2009**(2): 158-166. (in Chinese)

[7] 劲松. 北京话的语气和语调[J]. 中国语文, 1992, **2**: 113–123.
JIN Song. Mood and intonation of Beijing dialect[J]. *Zhongguo Yuwen*, 1992, **2**: 113-123 (in Chinese)

[8] Tokuda K, Nankaku Y, Toda T, et al. Speech synthesis based on Hidden Markov Models[J]. *Proceedings of the IEEE*, 2013, **101**(5): 1234–1252.

[9] Toda T, Tokuda K. A Speech parameter generation algorithm considering global variance for HMM-Based speech synthesis[J]. *IEICE Transactions on Information and Systems*, 2007, **E90-D**(5): 816–824.

[10] Qian Y, Wu Z, Gao B, et al. Improved prosody generation by maximizing joint probability of state and longer Units[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(6): 1702–1710.

[11] Meng F, Wu Z, Meng H, et al. Generating emphasis from neutral speech using hierarchical perturbation model by decision tree and support vector machine[C]//2012 International Conference on Audio, Language and Image Processing (ICALIP). Shanghai, China: 2012: 442–448.

[12] Tokuda K, Masuko T, Miyazaki N, et al. Multi-space probability distribution HMM.[J]. *IEICE Transactions on Information and Systems*, 2002, **E85-D**(3): 455–464.