中 文 信 息 学 报
第 18 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 2

文章编号:1003 - 0077(2004)02 - 0044 - 07

语音合成中的韵律关联模型*

吴志勇,蔡莲红

(清华大学 智能技术与系统国家重点实验室,北京 100084)

摘要:基于大规模语音数据库的文语转换系统(Text-to-Speech, TTS)中,如何选取合适的语音基元是提高合成语音自然度的重要因素。本文研究了连续语流中的韵律关联现象,提出了包含韵律关联参数的汉语韵律特征参数集,基于数据挖掘中的关联规则模型(Association Rules Model)建立韵律关联模型,并将该模型应用于基元选取。实验表明,该方法有效地利用了语音基元的韵律及关联信息,符合人耳的知觉感受,使得合成语音自然度的主观评测 MOS(Mean Opinion Score)得分与不考虑韵律关联时的结果相比提高了 12.22 %(3.49/3.11)。

关键词:计算机应用;中文信息处理;文语转换;基元选取;韵律关联

中图分类号:TP391 **文献标识码**:A

Prosodic Correlation Model in Text-to-Speech Synthesis

WU Zhi-yong, CAI Lian-hong

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: In this paper, a new unit selection approach for concatenative Text-to-Speech (TTS) synthesis based on prosodic correlation model is proposed. Firstly, prosodic correlations in continuous speech are studied. Then, some prosodic parameters, including prosodic correlation parameters, are concluded. Thirdly, a prosodic correlation model (association rules model from data mining) is put into use in unit selection. The experiments show that the unit selection method described in this paper can improve the naturalness of the synthesized speech: the MOS score can achieve 12.22 % higher than before (3.49/3.11).

Key words: computer application; Chinese information processing; Text-to-Speech (TTS), unit selection; prosodic correlation

1 引言

基于大规模语音数据库的文语转换系统中,语音基元选自包含大量语句的语音数据库,由于这些基元来源于自然语句,体现了所在上下文的音段和韵律特性,基于这些基元拼接出来的语流具有较高的自然度。而在此类系统中,语音基元选取的参数和算法是其关键。

关于基元选取的参数和算法,国内外学者作了大量的研究。ATR 实验室提出了基于匹配代价和拼接代价的基元选取算法[1]。李琳山(Lin-Shan Lee)等利用决策树进行基元选取,将汉语的音节特征参数归结为音节序号、词中位置、声调特征等 13 个语境参数[2]。陶建华等分

* 收稿日期:2003 - 11 - 21
基金项目:国家自然科学基金项目(60275014);863 资助项目(2002AA117010 - 05,2001AA114072)
作者简介:吴志勇(1977—),男,博士生,主要研究方向为语音合成、生物特征识别.

析归纳了影响汉语韵律的 17 个语境参数,并提出一种神经网络的韵律训练模型[3]。初敏等用 7 维的环境特征矢量描述每个音节所处的语境,利用平均拼接代价作为合成语音自然度的度量方法以指导基元选取[4,5]。上述文献将基元选取的特征参数主要归结为语境信息,试图以语境匹配来体现韵律匹配,然而,当局部语境相同而韵律表现不同时,仅考虑语境参数的基元选取难以给出最优的决策。G. Fant 曾指出:"寻找特征要素是发现以语境为基础的关系对比的体现"[6]。汉语 TTS 的研究表明韵律特征在语音合成中有较大作用。王玮等基于数据挖掘算法研究了汉语相邻音节韵律参数的预测[7],为语音合成中韵律特征的应用作了很好的尝试。本文重点研究了连续语流中的韵律关联关系,基于数据挖掘中的关联规则建立韵律关联模型,并应用于基元选取。实验表明,该模型优化了语音基元的选取,改善了语音合成的自然度。本文第二节介绍了韵律关联的概念,分析了在语音合成中引入韵律关联的必要性和意义,并基于韵律关联提出了汉语韵律特征参数;在第三节中分析了连续语流中韵律关联建模的问题,建立基于关联规则的韵律关联模型;第四节将韵律特征参数和韵律关联模型用于韵律代价函数的计算,提出了基于韵律关联模型的基元选取策略;最后给出了实验及分析。

2 韵律关联

基频、时长、幅度等声学参数体现了韵律特征,而且相邻基元的韵律特征的变化是影响语音自然度的重要因素。本文研究了语音基元韵律特征的匹配及前后关联的情况,并在语音基元的选取策略中引入韵律关联信息。

2.1 韵律关联

连续语流中同一音节的发音由于受不同上下文的影响而出现多种变化,这种变化既有因"协同发音"引起的音段特征(如共振峰)的改变,也有因语句的韵律不同而引起的超音段特征(如基频、时长、幅度等)的变化。上述变化共同导致了人们对语音知觉效果的改变。研究表明,基频、时长、幅度等超音段的韵律特征对知觉效果的影响更为显著:"即使前后音节在音段方面达到了很好的匹配,如果它们在体现超音段内容的声学参数上有很大的差异,两者之间的知觉差异仍不可忽略;而如果体现超音段内容的声学参数本身十分接近,那么即使相邻音节不匹配,它们在知觉上的差异也是比较小的"[8]。因此,在语音合成中引入韵律相关信息,从韵律匹配及变化的角度出发指导基元选取,更符合人的知觉特性,有利于提高语音合成的自然度。

另一方面,基于语境信息的基元选取,通过语境参数的匹配来选取具有不同发音特性的语音基元:语音的韵律变化被认为隐含在上下文语境中,并可以通过上下文语境的匹配加以重建。然而,随着语音数据库规模的不断增大,众多语境相同而韵律表现不同的语音基元出现在语音库中,此时,由于缺乏韵律信息的直接指导,上述算法往往难以给出最优的决策。例如,在本文的实验语音库中,音节"guo2"共有 400 多个样本,经过统计及人工听辨分析,按语境参数可以归并为 56 类,而按韵律表现可以分为 103 类。这表明,语境与韵律不是一一对应的关系,仅利用语境信息来选取基元,可能造成不合适的韵律特征的搭配关系而出现非预期的停顿、跃变等现象,影响语音合成的效果。

本文注意到韵律特征与语境参数的非唯一性,利用数据挖掘技术发现相邻音节韵律特征间的相关性,并利用关联规则模型对其加以建模,从韵律匹配及变化的角度考虑语音基元的选取。本文将相邻韵律成分的韵律参数相关性称为"韵律关联",并为其选定描述参数。

2.2 汉语韵律特征参数

本文在对汉语韵律层级结构[9]分析的基础上,在基元选取的特征参数中引入 5 维的韵律

参数以及 12 维的韵律关联参数,形成新的汉语韵律特征参数集。

这些参数根据其功能性,归结为 4 组特征向量,共 32 个韵律特征参数:音段参数 $\vec{C}$(7 维)、位置参数 $\vec{P}$(8 维)、韵律参数 $\vec{S}$(5 维)和关联参数 $\vec{G}$(12 维)。其中音段和位置参数[9]着眼于上下文的语境分析,而韵律参数[9]和关联参数则是上下文韵律相关性的体现。

本文中韵律关联参数包括:当前音节与前后音节的耦合度 g_{prev} g_{next}、与前后音节的韵律关联参数集 $\vec{G}_{prev}\vec{G}_{next}$(各 5 维)等。其中,耦合度 g_{prev} g_{next}为当前音节与前后相邻音节的关联程度,与汉语的韵律层级结构对应,以 3～1 分三级度量,分别表示相邻音节位于同一韵律词、韵律短语以及语句内部,反映了整个句子不同层级的节奏紧凑程度,随着层级的升高耦合度逐渐降低。前后音节的韵律关联参数集 $\vec{G}_{prev}\vec{G}_{next}$包括了相应前后音节的时长、基频特征、幅度、与当前音节的间隔时长、基频重设(F0 Reset)等。其中,基频特征采用加入了基频均值的基频规格化模型 SPiS 参数[3]表示,幅度以归一化的平均幅度表示,而基频重设则通过后音节基频的起始值减去前音节基频的终止值计算得到。

上述关联参数中,耦合度 g_{prev} g_{next}可以直接根据上下文由文本分析得到;韵律关联参数集 $\vec{G}_{prev}\vec{G}_{next}$,在语音数据库中可以直接根据语音数据通过韵律标注计算得到,而合成时的目标参数通过韵律建模模块根据文本分析的结果预测得到。

3 韵律关联规则模型

目前,关于韵律特征的变化规律主要基于孤立字词进行语言学、语音学等学科的研究,并经由专家知识的规则方式为语音基元的选取提供指导。必须看到,连续语流中韵律特征的变化是复杂的、多样化的,因此需要使用新的方法对连续语流中韵律特征的变化规律加以发现、描述并在基元选取中加以应用。

王玮等利用关联规则模型研究了汉语合成中相邻音节基频中值参数的预测[7],对关联规则在基元选取中的应用作了可行性的分析和描述。

本文在此基础上进一步研究了基频均值、时长和归一化的平均幅度等韵律特征的关联规则模型,并对其在基元选取中的应用作了进一步的探讨:将韵律关联规则模型和基元选取中拼接代价的计算相结合,提出了基于关联规则模型的韵律拼接代价的计算方法(4.2 节)。

表 1 给出了基于本文的实验语音库所发现的关于基频均值和时长特征的部分关联规则的举例。需要注意的是,由于音节的基频特征和其声调类型有很大的关系,而时长则和韵母类型相关,因此本文将基频特征按照不同的声调组合、将时长特征根据不同的韵母类型分别进行关联规则的发现。表 1 中给出的数据分别为当前音节为 1 声(阴平)、后音节为 4 声(去声)的基频关联规则,以及当前音节和后音节韵母均为"ang"的时长关联规则。

表 1　相邻音节基频均值及时长特征的关联规则举例(基频单位:Hz,时长单位:ms)

序号	当前音节基频均值范围	相邻后音节基频均值范围	规则支持度(%)	序号	当前音节时长范围	相邻后音节时长范围	规则支持度(%)
1	364 - 374	207 - 211	15	1	308 - 316	210 - 219	4
2	364 - 374	239 - 242	10	2	308 - 316	220 - 226	5
3	364 - 374	243 - 246	12	3	308 - 316	301 - 307	5
4	375 - 422	277 - 280	12	4	308 - 316	321 - 328	9
……	……	……	……	……	……	……	……

表 1 中的每一行分别表示一条基频均值和时长的关联规则。假设 A、B 分别为关联规则模型所产生的某些数据项的集合(如表 1 中所示的基频均值范围或时长范围),则 $A \rightarrow B$ 的关

联规则表示为:

$$A \rightarrow B \mid S \tag{1}$$

其中,S 为 $A \rightarrow B$ 的规则支持度,说明了语音数据库中满足 A 且满足 B 的数据出现的频率,对于大规模语音数据库而言,可认为其接近于当前规则的分布概率。在行文中,$A \rightarrow B$ 的规则支持度通常表示为:

$$S(A \rightarrow B) \tag{2}$$

韵律关联规则模型隐含了不同语境及韵律上下文情况下的韵律关联关系。在公式的关联规则中,A、B 分别反映了当前音节与后音节的韵律特征,为当前韵律特征上下文的体现,而规则支持度 $S(A \rightarrow B)$ 则反映了该韵律特征搭配关系出现的可能性。韵律关联规则的支持度越大,说明当前韵律特征的过渡变化情况在自然语音中出现得越频繁。考虑到基元选取中的应用,满足支持度大的规则的基元选取结果,其拼接合成的效果也更自然。以表 1 中基频均值的关联规则为例,如果当前音节候选基元的基频均值在 364 - 374Hz 之间,而后音节候选基元的基频均值有的落在 207 - 211Hz 之间、有的落在 239 - 242Hz 之间,则基元选取时,从基频均值变化的角度应尽可能选择落在 207 - 211Hz 之间的基元。

4 基于韵律关联模型的基元选取

4.1 韵律代价函数

基于韵律关联信息的基元选取,既要考虑当前音节上下文的韵律特征与目标特征间的匹配情况,也要考虑当前候选基元与其前后基元拼接时韵律特征的平滑过渡情况。假设待合成语句中音节数为 n,定义韵律代价函数为:

$$C = \sum_{j=1}^{n} V(j) + \sum_{j=2}^{n} C(j) \tag{3}$$

其中,$V(j)$ 为韵律匹配代价,描述了当前候选语音基元的韵律特征参数与目标韵律特征参数间的匹配程度,韵律匹配代价越小候选基元越符合相应的待合成文本单元的韵律特征要求;而 $C(j)$ 为韵律拼接代价,其意义为当前候选语音基元与相邻前音节基元拼接时在拼接处韵律特征参数过渡的平滑性,韵律拼接代价越小韵律特征过渡越自然。

韵律匹配代价 $V(j)$ 被定义为:

$$V(j) = \sum_{i=1}^{p} w_i V_i(j) \tag{4}$$

其中 p 为韵律匹配时所考虑的韵律特征参数的数目,$V_i(j)$ 表示候选基元的第 i 个韵律特征参数和目标韵律特征参数间的逼近程度,而 w_i 是反映第 i 个韵律特征参数对整体韵律匹配的影响因子的权重。在本文中,使用 2.2 节中介绍的 32 维的汉语韵律特征参数;对于 $V_i(j)$,声韵母类别、声调类别等音段参数 $\vec{C}$,由于不同参数具有不同的类型,因此采用分类量化的方法计算;而位置参数 $\vec{P}$、韵律参数 $\vec{S}$ 和关联参数 $\vec{G}$ 等,这些参数本身具有量化可比性,因此直接采用欧氏距离度量的方法对其加以计算。

韵律拼接代价 $C(j)$ 的定义为:

$$C(j) = \sum_{i=1}^{q} w_{ci} C_i(j) \tag{5}$$

其中 q 为韵律拼接时所考虑的韵律参数的个数,$C_i(j)$ 为当前候选基元与相邻前音节拼接时第 i 个韵律参数的拼接平滑过渡效果,w_{ci} 是该韵律参数拼接代价的权重因子。本文中考

虑了基频均值、时长、归一化的平均幅度等 3 个韵律参数,拼接代价 $C_i(j)$ 采用 4.2 节中介绍的基于韵律关联模型的方法加以计算。

公式(4)和(5)中需要解决权重设定的问题,本文采用神经网络的权值抑制和有指导的权重训练方法[9,10]来解决这个问题。

4.2 基于韵律关联模型的韵律拼接代价计算

公式(5)中,$C_i(j)$ 描述了相邻音节的韵律特征参数在拼接处的平滑过渡效果,其取值越小,说明相应的韵律特征参数在拼接处过渡越自然。本文利用第 3 节中介绍的韵律关联规则模型进行韵律拼接代价的计算。

考察韵律关联规则模型 $A \rightarrow B|S$,其规则支持度 $S(A \rightarrow B)$ 近似反映了语音数据库中当前音节的韵律特征和相邻前音节韵律特征搭配的概率,规则支持度越高说明这种搭配关系的在自然语音中出现的概率越大,使用这两个音节进行拼接的平滑度效果也越高,因此韵律关联规则模型的规则支持度是韵律拼接代价的最好近似。

具体到韵律拼接代价的计算,如果已知前一音节语音基元的第 i 个韵律特征 $g_i(j-1)$,当前候选基元相应的韵律特征 $g_i(j)$,相应的规则支持度为 $S[g_i(j-1) \rightarrow g_i(j)]$,则相应的韵律拼接代价为:

$$C_i(j) = 1 - S[g_i(j-1) \rightarrow g_i(j)], \quad i = 1,2,3 \tag{6}$$

其中 $g_i(j-1)$ $g_i(j)$ $(i=1,2,3)$ 分别为相应音节的基频均值、时长、归一化的平均幅度等韵律参数。

4.3 基于韵律关联模型的基元选取

基元选取时,通过韵律匹配代价为每个文本单元选择韵律特征最优匹配的语音基元,而利用韵律拼接代价使得相邻语音基元间的韵律拼接效果达到最佳。基元选取的过程如图 1 所示,为方便起见,图中将音节的序号转化为下标表示。

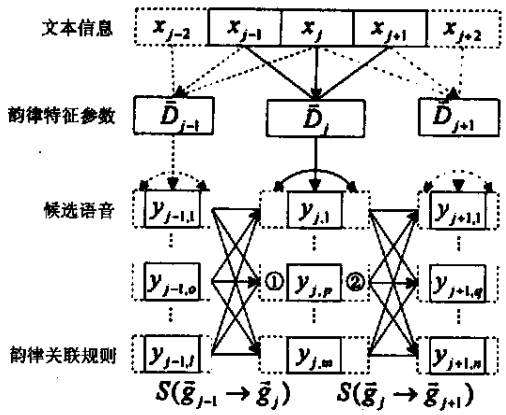

图 1 基于韵律关联模型的基元选取

进一步对公式(3)及基元选取的过程做一定说明。公式中 $V(j)$ 为韵律特征匹配的度量,如图 1 中由 $\vec{D_j}$ 指向 y_{jp} 的箭头所示。特别地,韵律特征参数中关联参数 $\vec{G}$ 的韵律关联信息 $\vec{G}_{prev}\vec{G}_{next}$ 和耦合度 g_{prev} g_{next} 决定了前后音节韵律关联关系的匹配,其中耦合度用于决定韵律关联信息的匹配程度,随着耦合度的降低,韵律关联匹配的要求也自然降低,如图 1 所示,$\vec{G}_{prev}$ g_{prev} 与当前候选语音 y_{jp} 的前音节 ①匹配,而 $\vec{G}_{next}$ g_{next} 与后音节 ②匹配。再者,公式中 $C(j)$ 为韵律拼接代价的度量,体现了相邻音节间的韵律关联关系,如图 1 中候选基元间互相连接的箭头所示,拼接代价由韵律关联规则支持度 $S(\vec{g}_{j-1} \rightarrow \vec{g}_j)$ 给出。

5 实验及分析

为了进行实验,我们收集了共 7000 多句的自然录音语句,由说标准普通话的女性播音员录制,语料包含近 85000 个音节,覆盖了汉语 417 种有调音节以及多种语境及韵律特征的搭配关系。基于上述数据抽取了一个中等规模的语音数据库作为研究对象。

实验时,利用该数据库通过数据挖掘算法进行韵律关联规则的获取,建立韵律关联规则模型,然后基于该模型及韵律代价函数进行基元选取并进行相应的实验。

5.1 选音对比实验

为了考察和验证韵律关联模型对于语音基元选取结果的影响,首先进行了选音对比实验。由于基频曲线的过渡是影响自然度的重要因素,而且基频特征便于可视化表示和理解,因此本实验主要对使用不同基元选取策略选音得到的基频特征的结果进行了分析比较。

图 2 给出了其中一句选音实验的对比结果。为便于比较,图 2 中(b)以虚线的形式给出了原始自然录音语句的基频信息,(a)为不考虑韵律关联信息的基元选取结果,(c)则显示了基于韵律关联模型的基元选取结果。可以看出,基于韵律关联模型的基元选取,在通常意义上的语境及韵律特征参数的匹配的基础上,进一步考虑了相邻音节韵律特征的关联关系,使得前后音节间的韵律特征的过渡(此处显示了基频特征)更为平滑,且其总体基频信息更接近原始的自然录音语句,合成的结果与考虑韵律关联前的结果相比更为自然。

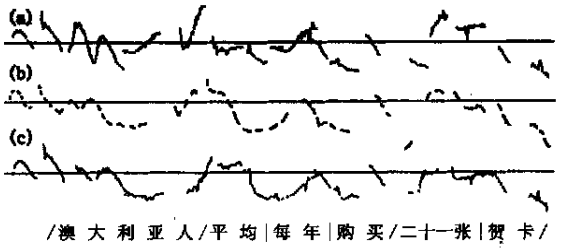

图 2 基频特征选音实验结果比较

图中虚线(b)为自然录音语句的基频信息;(a)为不考虑韵律关联的选音结果;(c)基于韵律关联模型的基元选取结果

5.2 主观听辨实验

为了从总体上进一步考察使用韵律关联模型前后合成语音自然度的变化情况,设计进行了主观听辨实验。共 100 组(300 句)语音用于听辨实验,每组语音包括使用韵律关联模型前后各 1 句合成语音,以及进行比照的相应自然录音语音 1 句。实验时,将每组语音以随机的先后次序播放,要求被试(听辨人)根据自己的听辨结果给出其认为的自然度,自然度以 5 分制方式给出,分别为:5 自然,4 较自然,3 可以接受,2 较差,1 不可接受。共 10 名被试参加了实验。

主观听辨实验的 MOS 得分结果如图 3 所示。图中给出了每名被试的 MOS 得分情况(横坐标为被试序号),以及对于全部被试进行平均的总体 MOS 得分结果(三种不同虚线类型的横线)。可以看出,考虑了韵律关联以后,合成语音自然度的 MOS 得分有一定的提高:不考虑韵律关联信息时合成语音的平均 MOS 得分为 3.11,而考虑了韵律关联信息后其平均 MOS 得分为 3.49,后者比前者平均提高了 0.38,相对提高百分比为 12.22 %。

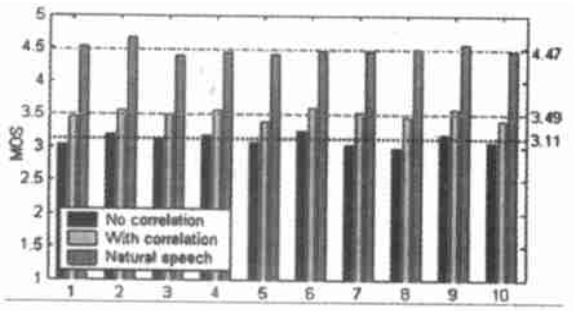

图 3 主观听辨实验结果比较

不考虑韵律关联时平均 MOS 得分 3.11;考虑韵律关联时 MOS 得分 3.49;自然语音的 MOS 得分 4.47

5.3 实验分析

以上实验表明,本文提出的基于韵律关联模型的基元选取策略,由于考虑了音节的韵律特征并充分挖掘了语音单元间韵律特征的关联关系,因此基元选取的结果更能满足韵律特征的要求,前后韵律特征的过渡更加自然平滑;主观听辨的实验结果表明本文的基元选取策略提高了合成语音的自然度。

本文提出的基于韵律关联模型的基元选取策略,主要有如下几个特点:(1)重点研究了连

续语流中的韵律关联现象,提出了包含韵律参数和关联参数的汉语韵律特征参数;(2)基于语音数据库和数据挖掘算法得到韵律关联规则模型,以数据驱动和定量描述的方法刻画了语流中韵律特征的关联关系,克服了传统方法只能给出定性描述的不足;(3)考虑到拼接处韵律特征的过渡及变化是影响拼接自然度的重要因素,将韵律关联规则模型应用于韵律拼接代价的计算,为拼接代价的估计提供了新的方法;(4)基于韵律关联模型的基元选取,从韵律角度出发,挖掘韵律上下文的关联关系,更符合人的知觉感受。

另一方面,尽管考虑了韵律关联信息后合成语音的自然度有所提高,但是和自然语音间还有差距(图3中给出了自然语音的平均MOS得分为4.47),仍需进行更多的工作,比如权重的进一步调整和训练以真实地反映语音数据库中语音韵律特征的分布情况等。

6 总结与讨论

本文重点研究了连续语流中的韵律关联关系,提出了基于韵律关联模型和韵律代价函数的基元选取策略。实验表明,该策略由于考虑了语音的韵律特征并有效反映了语音中前后单元韵律特征的相互作用,使得语音合成的质量和自然度有所改善。

还需看到,目前的合成语音和自然语音相比,还有较大的差距,仍需进一步的工作,如权重的调整和训练等。另外,人的感知是一个奇妙的过程,若能进一步从感知的角度来研究韵律特征参数,将基元的选取和人的感知联系起来,将能得到更好的结果。

致谢 本文中的部分工作得益于与陶建华博士的讨论[9,10],文中韵律关联规则的数据挖掘方法及结果得益并部分来源于王玮博士后的工作[7],在此表示衷心感谢。

参 考 文 献:

[1] Andrew J. Hunt, Alan W. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database[A]. ICASSP96[C]. Atlanta, Georgia, 1996. 373 - 376.

[2] CHOU Fu-chiang, TSENG Chiu-yu, LEE Lin-shan. Selection of Waveform Units for Corpus-based Mandarin Speech Synthesis Based on Decision Trees and Prosodic Modification Cost[A]. Eurospeech99[C]. Budapest, Hungary, 1999. 2295 - 2298.

[3] 陶建华, 蔡莲红等. 汉语文语转换系统中可训练韵律模型的研究[J]. 声学学报, 2001, 26(1): 67 - 72.

[4] CHU Min, PENG Hu. An Objective Measure for Estimating MOS of Synthesized Speech[A]. Eurospeech 2001[C]. Denmark, 2001. 2087 - 2090.

[5] 初敏. 韵律研究与合成语音的自然度[A]. 第五届全国现代语音学学术会议. 新世纪的现代语音学[C]. 北京: 清华大学出版社, 2001. 295 - 301.

[6] G. Fant. 言语产生中的相互作用现象[M]. 1987.

[7] 王玮, 蔡莲红. 基于数据挖掘算法的汉语合成韵律参数预测方法[J]. 声学学报, 2003, 28(1): 1 - 6.

[8] 周讯溢, 王蓓, 杨玉芳等. 语句中协同发音对音节知觉的影响[J]. 心理学报, 2003, 35(3): 340 - 344.

[9] 吴志勇, 蔡莲红, 陶建华. 基于汉语韵律参数的语音基元选取[A]. 第六届全国人机语音通讯学术会议[C]. 深圳, 2001. 199 - 202.

[10] 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(1): 1 - 6.