文语转换系统中的韵律研究*)

Study on the Prosody in Text-to-Speech Systems

谌卫军　李建民　林福宗　张　钹

(清华大学计算机科学与技术系　北京100084)

(清华大学智能技术与系统国家重点实验室　北京100084)

Abstract　The paper discussed the problem of prosody in Text-to-Speech(TTS) systems. It firstly analyzed the different markup methods employed in current TTS systems, then discussed the characteristics of prosody in detail, and put forward a comprehensive and integrated Chinese prosodic structure. Lastly a new markup language named Chinese Prosodic Markup Language(CPML) was proposed. The features of CPML are: 1. Abstract and platform-independent, 2. Covering most of important prosodic information, 3. Hierarchical prosodic structure, and 4. Open and updating.

Keywords　Text-to-Speech, Prosody, Prosodic structure, CPML

1　引言

文语转换(Text-To-Speech, TTS)是将文字形式的信息转换成自然语音的一种技术，在人机交互、通讯、资讯、家电等领域有着广泛的用途。然而，当前的TTS系统普遍存在着输出语音的机器味太浓、不够自然的现象，在很大程度上阻碍了它的推广和应用。其根本原因即在于合成语音中缺乏必要的韵律信息。我们认为，一个TTS系统应分为相对独立的上下两层：韵律结构分析和语音生成，上层负责分析语句的韵律结构，并标注相应的韵律标记，下层负责将之转换成相应的合成器参数，并输出语音。因此，当前的首要任务就是要研究韵律的主要特点、韵律的结构和主要内容，在此基础上，制订出一套相应的韵律标记方法。

2　当前的韵律表示方法

当前的TTS系统中的韵律表示方法各不相同，有的是定义了一套标记系统，有的是在各种标记语言的基础上，定义了自己的韵律标记语言，标记语言主要有两种：SGML和XML。SGML[2]是一个国际标准，是在通用标记语言GML的基础上制订的。其基本思想是将文档的逻辑结构和物理结构分开：由文档的作者来设定逻辑结构，再由出版者根据不同的机器/平台生成不同的物理结构，确切地说，SGML本身并非一种标记语言，它实际上是一个用于创建标记语言的语法集合。XML[4]是由环球网协会(W3C)为适应因特网的需要于1996年开发出来的，它是SGML的一个简化版本，与SGML兼容。

2.1　ToBI

ToBI[3]是较早提出的一套韵律标记系统，广泛应用于英语语音的韵律分析和标注。它将韵律标注分为多个层次，每一层都有不同的符号来描述不同的韵律信息和韵律的变化情况。语调层描述的是整个语句的语调变化规律，间隔层描述的是相邻的两个音之间的关联关系，此外还有拼写层和杂类层。ToBI能描述自然语音中最主要的韵律特征，易学易用，是个开放的系统，用户可以根据自己的需要对标记进行添加或调整。

2.2　SSML

SSML(*Speech Synthesis Markup Language*)[1]是爱丁堡大学语音研究中心提出的一种基于SGML的标记语言，其设计目标是成为一种平台独立的语音合成系统接口标准。SSML定义的主要标记有：短语间隔标记，用于设定短语边界的间隔类型和语调的类型如陈述句、疑问句等；重读标记，用于将某个单词设定为重音；发音标记，用于设定单词的读音。

2.3　JSML

JSML(Java Speech Markup Language)[4]是Sun公司开发的一种基于XML的标记语言，用于给文本

*)国家自然科学基金资助课题。谌卫军　博士生，研究方向为文语转换、人工智能理论。李建民　博士生，研究方向为文语转换。林福宗　副教授，研究方向为多媒体。张　钹　教授，博士生导师，中科院院士，研究领域为人工智能理论、神经网络、遗传算法理论等。

标注韵律信息，并以此作为 Java 语音合成器的输入。JSML 的标记分为三类：结构类标记、发音类标记和杂类标记。

结构类标记有两个：段落标记和句子标记，前者设定一个段落，后者设定一个句子。发音类标记有四个：发音标记用于设定单词或短语的读音以及当前文字的类型，如日期、数字；重读标记用于设定当前单词的重音级别，有四个级别；间隔标记用于在连续的语音中设定一个停顿的类别或时长；韵律标记用于设定当前文字的语速、音强、基频和音高范围。杂类标记则包括一些增强功能，如嵌入特定语音合成器的控制命令等。

2.4 SABLE

在 SSML 和 JSML 的基础上，贝尔实验室提出了 SABLE 语言[5]。它综合了前两者所定义的标记，并改进了属性值。每个标记的属性既可以是绝对或相对的数值，也可以是几种类别之一。它还新增了一个说话者标记，用于设定说话人的属性，如性别、年龄等。

2.5 汉语韵律标记系统

蔡莲红[6]定义了一套用于汉语文语转换的标记系统，用来控制相应文字的读音、语速、基频、数字朗读方式、停顿时间、音量、语音风格，字重音级别等韵律信息。

李智强[7]提出了汉语韵律标记系统的几条设计原则和评测标准，并设计了一个普通话韵律标记系统，该系统分五层，每一层标记不同的韵律现象。1. 拼音层：标注每个字的拼音和声调。2. 语气层：标注语句的四种语气：陈述、疑问、祈使、表情。3 声调语调层：标记每个音节的声调变化和全句的语调变化。4. 重音层：标记音节的重音等级，共四种级别。5. 韵律结构层：标记韵律词、韵律短语和语调短语之间的间隔，共四种间隔类型。

3 韵律标记设计中的几个问题

由上可见，由于对韵律的内容和结构看法不一，并受所采用的语音合成器的影响，当前的 TTS 系统有着各自不同的韵律表示方法。它们有的过于简单，不能充分完整地描述所有主要的韵律信息；有的对层次的理解和划分不够准确清晰，不能很好地体现韵律的层状结构；有的未能正确地把握韵律的特点，所定义的标记或属性过于具体，把语音合成器的参数也看成是韵律，如基频的大小、音高的频率范围、停顿的时间长度等等，而事实上，人们在平常说话的时候是不可能去计算自己和他人的声音频率，或者去估计每一次停顿需要多少秒，更重要的是，任何一个韵律信息都不是孤立存在的，它的实现要受到上下文其它韵律信息的影响。例如，同样是重音，在实现方法上既可以是提高音量，也可以是延长音长或运用停顿，因此，不能简单地把韵律和具体的合成器参数等同起来，而有必要在这两者之间增加一个韵律建模模块，它能综合考虑该韵律所处的上下文以及系统所用的语音合成器的特点，将韵律转换成恰当的合成器参数。此外，由于现有的标记方法各自为政，缺乏一个行业通用的公共标准，势必会增加不同系统之间移植和集成的困难，而即便在同一系统中，上层或下层的任何变动都将导致对方的改动。而且在机器翻译等系统的人机语音对话程序中，韵律信息是手工标注的，用户每使用一个新系统，都不得不学习一套新的标记。所有这些都严重妨碍了文语转换技术的发展和推广应用。

我们认为，韵律标记的设计要从韵律自身的特点出发。韵律有三个最主要的特点，一是抽象性，韵律实际上是位于人的思维层面的一种表达语意、表现情感的手段，而不是位于感官层面上的音高、音强等语音的物理属性。事实上，由于年龄、性别、发音器官等方面的不同，每个人都有着各自不同的声音特点，发出的语音信号各不相同，但这并不妨碍我们正常的交流，原因就在于当我们听他人说话时，听觉系统会首先对输入的语音信号进行过滤处理，然后传入大脑并在大脑中抽象出语句的韵律模式，从而使我们能理解对方的意思；反之，当我们说话时，也不是直接与音高、音强等语音的物理属性打交道，而是首先在大脑中形成话语的韵律模式，然后再控制语言系统将之转换成语音信号输出。总之，大脑思维的抽象性决定了韵律的抽象性。二是繁杂性，语音中韵律信息是丰富多采，难以尽述的。三是局部性、上下文相关性，任何一个韵律信息都是作用于一定的局部范围内，而且它的实现方式也受到上下文其它韵律信息的影响。

根据韵律的特点，我们试着提出如下几条设计原则：

首先，要科学地制订标记，既要遵从当前的语言学、语音学和朗读学等方面的相关理论，又要与实验中对语音语料库的观察分析结果相一致。

其次，韵律标记必须是抽象的，是与人们在平常说话时对韵律的理解相一致的，因而也是独立于具体实现平台的。

第三，韵律标记应该能全面、完整地覆盖自然语音中的主要的韵律现象。

第四，韵律的局部性和上下文相关性还决定了韵律结构是一种层状结构，每一层都包含着不同的韵律现象。

第五，标记语言应该是开放的，能够不断地更新和完善。

第六，由于韵律标记是抽象的，而语音合成器能够

接受的却是具体的参数，因此在语音生成部分必须有一个韵律建模模块，它能综合考虑每一个韵律所处的上下文环境以及系统所用的语音合成器的特点，将韵律转换成恰当的合成器参数。例如，由语气来确定语调的大致轮廓，由语速来确定停顿的大致长度等。当韵律的内容有所变化或者系统需要更换语音合成器时，只需要改动这个韵律建模模块即可。

总之，由于韵律标记的抽象性，独立于具体实现平台，这就使得上层和下层的研究完全独立开来，它们的任何变动都不会影响到对方。当前，文语转换技术正处于尚未成熟的阶段，无论是上层的韵律结构分析方法还是下层的语音合成方法，都有许多需要尝试、改进的地方，在这个时候上下层的独立就显得尤为重要了。另外，这种独立性还使得建立一种行业通用的、标准性的韵律标记语言成为可能，只要这种标记语言是开放的，能够不断地更新完善，那么这个目标的实现是完全有可能的。

4 汉语的韵律结构

根据上述设计原则，结合汉语的特点，我们提出了汉语的韵律结构。它所包含的韵律信息来自两个方面：语言学、语音学和朗读学等学科的相关理论知识[8~14]，以及我们对一个语音语料库的分析结果，该语音语料库取材于中央人民广播电台播出的各类节目。

汉语韵律结构分为六个层次：韵律词、次韵律短语、主韵律短语、句子、段落、总体。

韵律词是最小的韵律单位，它不等同于通常意义上所说的作为具有确定语法或语义功能的最小单位的词。韵律词的划分依据是连接紧密，带有最基本的韵律信息。它的长度也有限制，最短为一个音节，最长为四个音节。韵律词所包含的多音字、变调、词重音模式、轻声等韵律信息是汉语特有的，一般都能在韵律词范围内通过各种规则和知识来解决。

次韵律短语由若干个韵律词组成，是一个与语法成分密切相关的相对独立的整体，它所包含的韵律信息主要是韵律词之间的间隔类型。值得注意的是，间隔并不等于停顿，间隔属于韵律结构分析范畴，而停顿属于语音生成范畴，有间隔的地方不一定出现停顿，这还要根据上下文的其他韵律信息来综合判断。

主韵律短语由若干个次韵律短语组成，因而也是一个相对独立的整体，它所包含的韵律信息是各个次韵律短语之间的间隔类型，这种间隔通常表现为长短不一的停顿。

句子由主韵律短语组成，是语言的基本组成单位。句子所包含的韵律信息有：1. 主韵律短语之间的间隔类型——通常表现为较长的停顿；2. 语句重音，在一句话中，体现句子语意目的的词语要用重音加以强调。重音有主要重音、次要重音之分；3. 语气，在不同的语言环境下，说话人的思想感情、态度目的不同，在语气上就会有不同的色彩和份量；4. 语速，它取决于语句的内容、说话人的个性特征和语气等方面。

段落由句子组成，它所包含的韵律信息主要是句子之间和段落之间的间隔类型。由于句子之间、段落之间的关联程度不同，因此间隔的类型也不尽相同。

总体的韵律信息包括：1. 文体类别，指明该段文字的类别，如对话、记叙文、散文等；2. 个性特征，由于发音器官、性格、性别、年龄等因素的影响，每个人的声音特点都各不相同；3. 语言信息，汉语的方言众多，必须标明语言的种类。

一段文字的所有韵律信息归结在一起，就组成了它的韵律结构。

5 汉语韵律标记语言(CPML)

在上述理论的基础上，我们定义并实现了一种用于汉语文语转换系统的韵律标记语言 CPML(Chinese Prosodic Markup Language)。它是一种基于 XML 的标记语言，每一个标记由标记名称和若干个属性组成。主要的标记有：

声调标记〈ShengDiao〉，设定当前音节的调值，属性 TYPE 采用五度标记图来表示声调，如阴平的调值为55，上声的调值为214。主要用于描述变调这一韵律现象。

词重音模式标记〈CiZhongYin〉，设定当前韵律词的重音模式，用属性 MODE 来表示。一般而言，二字词有重重、重中、重轻、中重四种模式，三字词有中轻重、中重轻、重轻轻三种模式，而四字词绝大多数都是中轻中重。

轻声标记〈QingSheng〉，将当前音节设定为轻声。

间隔标记〈JianGe〉，在当前位置设定一个间隔，属性 TYPE 表示间隔的类型。该标记用来设定每一层的各单元之间的间隔类型，属性值包括：韵律词间隔、次韵律短语间隔、主韵律短语间隔、句间间隔、段间间隔。注意：1. 上层间隔蕴含下层间隔，若在同一位置既有下层间隔又有上层间隔，只需标注上层间隔。2. 除韵律词间隔外，每一层的间隔又可分为长、短两种类型。这是因为即便是在同一层，各单元之间的关联程度也有所区别。

语句重音标记〈JuZhongYin〉，把当前文字设定为句子的重音。属性 TYPE 表示重音的类型：主要重音或次要重音。在语音生成部分，重音的具体实现方法是多种多样的，可以重读，也可以轻读，可以放慢速度，也

可以变换节奏，还可以运用停顿或音色的变化等。

语气标记〈YuQi〉，设定当前文字的语气，属性 TYPE 表示语气的类型。汉语中的语气实在是丰富多彩、难以尽述，我们大致划分出了十八类，每一类包括几种相近的语气。例如平淡类包括陈述，疑问类包括疑问、犹豫，感叹类包括慨叹、赞叹，等等。

语速标记〈YuSu〉，设定当前文字的语速，属性 LEVEL 表示语速的级别，共五级：最快，快速，中速，慢速，最慢。

类别标记〈LeiBie〉，设定当前文字的文体类别，用属性 TYPE 来表示。汉语中主要的文体类别有：对话、记叙文字、说明文字、新闻、散文、诗歌、戏曲、报道等等。每一种类别有着各自不同的语言风格。

个性特征标记〈GeXingTeZheng〉，设定说话人的个性特征。属性 CHARACTER 表示说话人的性格，如活泼、内向等；属性 AGE 表示说话人的年龄，如儿童、少年、青年、中年、老年等；属性 SEX 表示说话人的性别，男性或女性。

语言标记〈YuYan〉，设定语言的种类，用属性 TYPE 表示，如普通话、粤语等。

6 例子

为了说明 CPML 的实际应用，我们从新闻广播中摘选了一段文字，仔细分析了播音员在朗读时所用到的韵律信息，然后将其中的一部分用 CPML 语言标记了出来。

〈CPML〉〈YuYan TYPE＝"普通话"〉〈GeXingTeZheng AGE＝"中年"SEX＝"女性"〉〈LeiBie TYPE＝"记述文字"〉〈YuSuLEVEL＝"中速"〉〈YuQi TYPE＝"陈述语气"〉

〈CiZhongYin MODE＝"重中"〉浙江〈/CiZhongYin〉〈JianGe TYPE＝"韵律词"〉〈JuZhongYin TYPE＝"主"〉永康〈/JuZhongYin〉拖拉机厂〈JianGe TYPE＝"次短语"〉根据农村〈QingSheng〉的〈/QingSheng〉需要，〈JiangGe TYPE＝"主短语"〉生成一种"价格相当于〈ShengDiao TYPE＝"51"〉一〈/ShengDiao〉头牛，效率胜过两头牛，牛的活儿它全会"的小型拖拉机，受到各地农民的热烈欢迎。

〈/YuQi〉〈/YuSu〉〈/LeiBie〉〈/GeXingTeZheng〉〈/YuYan〉〈/CPML〉

结束语 长期以来，韵律问题一直是困绕着文语转换技术的发展和推广应用的最大难题，本文便是在这个方面的一些尝试。我们结合汉语语言学、语音学和朗读学等方面的相关理论知识，以及我们对一个语音语料库的观察分析结果，仔细研究了韵律的主要特点，完整、全面地提出了汉语韵律的结构及其主要内容，在此基础上，定义并实现了一种新的韵律标记语言 CPML。它的主要特点是：1. 抽象的、独立于具体实现平台的韵律标记；2. 全面完整地覆盖了主要的韵律现象；3. 层状的韵律结构；4 开放的，能不断地更新和完善。今后的工作主要有两个方面，一是韵律结构分析部分，对于任意的输入文本，如何自动地给它标记韵律信息，生成抽象的韵律结构；二是语音生成部分，如何建立韵律建模模型，采用何种语音合成方法，将标有韵律信息的文字转换成自然的语音。

参考文献

1 Taylor P, Isard A. SSML: A Speech Synthesis Markup Language: [report] Center for Speech Technology Research, University of Edinburgh, 1996

2 Bray T, Paoli J. Extensible Markup Language (XML) [W3C Working Draft] Available at: http: // www. w3. org/TR/WD-xml-970807, 1997

3 Silverman K, Beckman M, et al. TOBI: a standard for labeling English prosody. ICSLP, 1992. 867～870

4 Sun Microsystems Inc Java Speech Markup Language Specification. [report]. Sun Microsystems, Inc. 1997

5 Bell-labs. SABLE: A Synthesis Markup Language. [report]. Available at: http: // www. alphaworks. ibm. com/formula/speechml, 1999

6 蔡莲红，罗恒. 文语转换系统韵律置标方法的研究. 软件学报，1996，7(增刊)：514～518

7 李智强. 普通话韵律标音系统的初步研究. 见：吴泉源，钱跃良编. 智能计算机接口与应用进展 北京：电子工业出版社，1997. 169～173

8 黄伯荣，廖序东. 现代汉语. 北京：高等教育出版社，1993

9 徐世荣. 普通话语音知识. 北京：文字改革出版社，1980

10 林焘 语音学教程. 北京：北京大学出版社，1992

11 冯胜利. 汉语的韵律、词法与句法. 北京：北京大学出版社，1997

12 张颂. 朗读学. 长沙：湖南教育出版社，1983

13 郭锦桴. 综合语音学. 福州：福建人民出版社，1993

14 乌坤明，等. 朗读知识与技巧. 长春：吉林文史出版社，1991