·博士论坛·

汉语文语转换系统(TTS)*

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)
E-mail:cwj@s1000e.cs.tsinghua.edu.cn

摘　要　文章讨论了一个典型的汉语文语转换系统的实现。首先介绍了系统的整体框架及其各个功能模块,然后分析了系统的特点及其存在的问题,最后从两方面讨论了改进系统的具体思路:提出了一种简单而有效的基音周期提取算法,验证了上下文环境在提高合成语音自然度中的作用。

关键词　文语转换　基音周期　合成单元　自然度

A Chinese Text-to-Speech System

Chen Weijun　Li Jianmin　Lin Fuzong　Zhang Bo
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)
(State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084)

Abstract: This paper discusses the implementation of a typical Chinese Text-to-Speech (TTS) system. Firstly, It presents the frame structure of the system and discusses all of the modules in detail, then analyzes its characteristics and problems existing. Lastly This paper does some experiments in order to improve the quality of the system: It puts forward a simple and effective algorithm to calculate the base period of speech waveforms and verifis the importance of context in improving the naturalness of synthesis speech.

Keywords: Text-to-Speech, Base Period, Synthesis Unit, Naturalness

文语转换(Text-To-Speech)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、通信、资讯、家电等领域有着广泛的用途。例如,在人机交互方面,可以把它与语音识别技术结合起来,实现人机的语音交流,使人机界面更为友好。同时它还能帮助残障人士,可以作为盲人的“眼”、发音障碍者的“嘴”;在文本录入时,可以用它实现实时的语音纠错;在通信中,可以用它来实现语音自动应答系统,使用户可以通过普通电话来查询各种信息、上网冲浪、收发电子邮件等。在文中,作者实现了一个典型的汉语文语转换系统,并提出了一些改进合成语音自然度的方法。

1　系统的框架结构

一个典型的汉语文语转换系统包括文本格式化、分词处理、标音处理、语音合成等模块,系统的框架结构如图 1 所示。

1.1　文本格式化

一个实用的文语转换系统会面临各式各样的输入:从简单的纯文本到带有特定格式的文字,如排版好的文章、WWW 网页、电子邮件等等。它们不仅包含通常的字符,还可能会包含一些数字和运算符号等特殊字符。此外,有些语种还有缩写词等特殊用法。文本格式化的任务就是把这些特定的格式、特殊的字符和用法转换成口语中相应的文本。在笔者的系统中,文本格式化的内容主要包括:

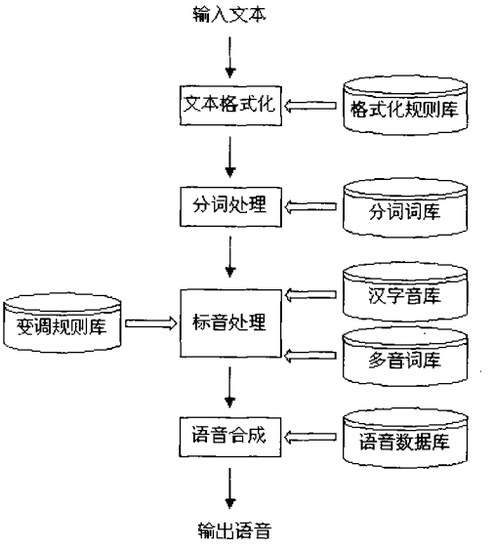

图 1　文语转换系统的框架结构图

(1)各种数字字符串。第一类是纯数字串。包括整数、电话号码、年份等。按中文习惯,整数的念法是分别在各个数字后面

*该课题得到国家自然科学基金(编号为 69523001)的部分资助。

作者简介: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。

插入相应的单位，如百、千、万等。电话号码按数字串来念，年份亦如此。而月份、日期等皆按整数的念法。第二类是杂有各类符号的数字串。包括浮点数、科学计数法、分数、货币、百分数、日期的数字格式（三个数字间以“/”或“-”或“.”隔开）、时间的数字格式（两个数字间以“:”隔开）等。浮点数的念法分为两部分，小数点前面的部分按整数念，后面的部分按数字字符串念。科学计数法的念法比较固定。分数的念法为分母“分之”分子，其中分母和分子皆按整数念。货币的念法一般是在数目后再加上该货币名称。百分数则先念“百分之”然后再念数值。日期的数字格式可首先转换成“××年××月××日”的形式，再设定相应的数值。时间的数字格式可先转换成“[上午|下午|晚上|凌晨]××时××分”的形式，其中的数字按整数念。

(2)常用符号。包括各种常用的外文字符，数学运算符号等。如a,b,+,×。

(3)缩写词。这是外文所特有的，如IEEE、HTML等，只须把字母逐个念出即可。

1.2 分词

词语是人类语言的最小语法单位，也是自然语音中的最小韵律单位。汉语中的多音字、变调等韵律特性都是在词语一级完成的。然而，与印欧语系不同的是，汉语和其他一些亚洲语言是以单个的字作为语言的基本组成单位，词语由一个或多个字组成，且词语与词语之间没有明显的界限。因此，对这些语种来说，首先要做的就是分词。

分词问题包括两方面的内容：词库和算法。词库中存放的是符合汉语分词规范的词条，而算法则利用词库，对输入的句子进行分析，将之分割成一个个的词条。在该的系统中，词库中存放有最常用的约五万个汉语词条，词库结构采用了DBF数据库文件格式，并按照词条的顺序排好序，同时对词条的首汉字建立了索引。这样一来，既保证了词库的通用性，使它能得到各种编程语言的支持，又提高了检索的速度，保证了分词的效率。在分词算法上，采用的是简单直观、十分有效的反向最大匹配法，即从句子的末尾开始处理，首先匹配最长的可能的词条，若不成功，则逐个去掉最前面的一个汉字，直至匹配成功。这种方法简单、快速，精度也能保证，比较适合文语转换这类对分词精度要求不太高的系统。

1.3 标音处理

标音处理负责给句子中的每一个汉字标注拼音、声调、停顿等音律信息。这涉及到如下几个问题：单音字的标注、多音字问题、轻声、变调、停顿等。

1.3.1 单音字的标注

这里参照《新华字典》构造了一个汉字音库，存放了所有国标汉字的各种读音。在标音处理时，对于只有单个拼音的汉字，只须从库中把它所对应的拼音取出即可。

1.3.2 多音字问题

在汉语中，多音字是一个非常普遍的现象。它的特点是：个数多、出现频率高、读音多为两个。多音字问题一般在词语一级便可解决，也就是说，每个多音字虽然有多个可能的读音，但当它出现在词语中时，其发音一般是确定的。因此，构造了一个多音词库，里面存放了分词词库中所有带有多音字的词条，并用手工的方法设定了它们的读音。这样，在标音处理时，如果碰上多音字，只须到多音词库中找到相应的词条，其读音便可确定。

1.3.3 轻声

轻声是指在一定条件下读得又短又轻的调子，这里实现了一些典型的读轻声的情形：

(1)助词“的、地、得、着、了、过”和语气词“吧、呢、啊”等；

(2)叠音词和动词的重叠形式的末尾字，如“爸爸”、“看看”等；

(3)构词用的虚语素“子、头”和表示多数的“们”等；

(4)作为量词的“个”。

1.3.4 变调

在汉语词语的发音中，有些音节由于受到相邻音节的影响，声调发生了一定的变化，而与单字时不同，这种情形叫变调。该系统考虑的变调类型主要包括：

上声的变调。单念或出现在词语末尾时，声调不变；在非上声之前，变为半上声；如果两个上声相连，则前一个变得象阳平；

去声的变调。两个去声相连，前一个变为半去声；

“一”、“不”的变调。单念或出现在语句末尾时，声调不变；在去声之前，变为阳平；在非去声之前，“一”变为去声、“不”仍读去声；

“七”、“八”的变调。在去声之前变为阳平，其余场合不变。

1.3.5 停顿

在说话时，为了明晰语意或缓一口气，就需要在不同位置插入不同长短的停顿。在该的系统中，主要是在标点符号处加入停顿，不同的标点符号停顿长短不同。顿号最短，逗号较长，分号又较逗号为长，句末的标点符号（句号、问号、感叹号等）表示的停顿又较分号为长，而章节段落之间的停顿还要更长一些。

1.4 语音合成

语音合成的方法主要有三类：发音参数合成、共振峰合成、拼接合成。其中，拼接合成法简单直观、合成语音的自然度较高，是当前的主要方法。其基本思路是：首先以波形编码的方式存储大量的涵盖广泛的合成单元，在合成时，根据需要从中挑选一些最合适的单元，然后把它们拼接起来，并调整其基频、时长和幅度，最终生成语音。

拼接合成法的首要问题是合成单元的选择，这需要综合考虑自然度、存储量和灵活性等三方面的因素。合成单元越大，语音的自然度就越高，但存储量大，处理也不够灵活；合成单元越小，存储量就越少，处理越灵活，但自然度会受到影响。在汉语中，音节是最小的语音单元，共有1200多个有调音节和一些轻声音节。因此，这里采用了有调音节来作为合成单元。这样，一方面能保证较好的自然度；另一方面，由于音节个数有限，所需的存储空间也不多，而且在音节一级的处理也是非常灵活的。

在语音数据库中存放着所有有调音节的波形编码。为了便于查找，在此创建了索引文件，记录了每个音节的起始位置和长度。在语音合成时，根据每一个汉字的拼音、声调、停顿等音律信息，从语音数据库中挑选出合适的波形编码，然后拼接起来成为输出语音。

2 系统的改进思路

该系统是一个典型的文语转换系统，它实现了最基本的从文本到语音的转换功能，但也存在着诸多问题。如声音缓慢、生硬、单调，没有连贯感，缺乏人说话时的那种抑扬顿挫的节奏感，不够自然。究其原因，在于缺乏必要的韵律信息以及将这些韵律信息实现在输出语音中。现有的合成单元都是在单字发音

情形下录制的，而在自然语音流中，每一个音节都会受到上下文其它音节的影响，因而必须根据上下文做相应的调整。下面，从两个方面提出了改进系统的思路。

2.1 基音周期的提取

一个音节由声母和韵母两部分组成，从语音波形上看，声母部分类似于白噪声，而韵母部分则是一种周期信号，其周期在基音周期左右摆动。因此，可以通过在时域上增减周期信号的方法来调整音节的时长，从而调整语速。

2.1.1 基音周期提取算法

作者构造了一种简单有效的基音周期提取算法，其中心思想是特殊点的匹配。笔者认为：对于一个周期信号中的相邻两个周期，它们的对应点的值应该是匹配的（由于各种因素作用，不太可能是相等的，而是有一定的偏差），因此只需验证连续的两个点集之间是点点匹配的，便是找到一个周期了。此外，从语音波形图的分析可知，事实上只需验证两个点集中的若干个特殊点，如幅度的局部峰值，便足以判断它们是否匹配，而无须逐点验证。这样便极大地减少了计算量，而精度也不会损失。

在具体实现算法时，有几点值得注意：

(1)匹配条件的设置，即定义了一个允许偏差值 DELTA，当两个点的差值小于该偏差时，则认为是匹配的。

(2)声母部分是白噪音而不是周期信号，但由于幅值很小，任何两点的差值都满足匹配条件。为避免这种情况，设置了一个幅度最小值 VALUEHEAD，如果当前峰值点的值小于该值，则说明是声母部分而不予考虑。

(3)一个周期中的峰值点个数是有一定范围的，太小了会带来差错，太大了则可能将多个周期当成了一个周期。通过对语音波形的分析，可以观察到每个周期的峰值点个数在 4 到 11 之间。

(4)每个音节的周期并非常量，而是在小范围内变动，有的持续时间长一些，有的短一些。

(5)所有参数的值都是在不断的试验中挑选出来的，而且偏差值 DELTA 是一个相对值，它可以根据波形的具体情况进行调整。

算法流程：

(1)计算出所有的峰值点，得到一个峰值点序列，设为，X_1，X_2，…，X_n。将 X_1 设为当前峰值点。

(2) 对当前峰值点，从后面第 L_{min}(=4)个峰值点到第 L_{max}(=11)个峰值点查找与之匹配的点。

(3) 如果没有找到，则将下一个峰值点设为当前点，转(2)。

(4) 如果找到一个匹配点，设 X_m 为，则调用多点匹配算法 MPMatch 验证点集$\{X_1, X_2, \cdots, X_{m-1}\}$与$\{X_m, \cdots, X_{2m-2}\}$是否匹配。

(5) 若不匹配，则从第 m+1 个峰值点到第 L_{max}(=11)个峰值点继续查找与之匹配的点，转(3)。

(6) 若匹配，则得到一个周期了。然后不断调用 MPMatch 算法一直匹配下去，直到某次匹配不成功，这说明具有这个周期的信号已经结束了，因此转(2)寻找新周期。

多点匹配算法 MPMatch：

(1) 为了验证任意两个点集$\{X_1, X_2, \cdots, X_m\}$与$\{Y_1, Y_2, \cdots, Y_n\}$是否匹配，首先要找出点与点之间的对应关系，即两个点集中的哪些点是位于周期信号中的相同的峰值位置。这可以根据各个峰值点与初始点的距离来判断。

(2) 找到所有的对应关系后，将它们两两配对，并计算差值。如果某两个点的差值悬殊，则整个匹配失败；否则计算所有差值的算术平均值，如果该值小于一个阈值，则匹配成功。

2.1.2 实验效果

作者对语音数据库中的所有 1659 个音节使用了该算法，标注了每个音节的基音周期。然后做了个实验，通过删去每个音节的若干个周期，得到缩短后的读音。结果表明：在保证声音质量的前提下，99%的音节的时长可以缩短二成至三成，92%可以缩短三成至四成，74%可以缩短四成至五成。这说明该算法是很有效果的。

2.2 合成单元的研究

为了提高自然度，除了对合成单元的时长、基频、幅值等进行调整外，还需要对合成单元本身进行研究。作者意识到，即使是同一个音节，在不同的上下文环境下，发音情况也有很大的差别，因此在语音数据库中，对同一个音节保存不同上下文环境下的多个读音，对提高自然度是很有帮助的。这里研究了词语一级的情形。

在词语一级，每个音节被分为三种类型：重音、中音、轻音。双音节词有两种轻重格式：重轻和中重；三音节词以“中轻重”为主，少数是“中重轻”格式；而四音节词则以“中轻中重”为主。为了验证效果，做了如下实验：

表 1 多类型合成单元实验

（其中合成读音是由读音来源中的相应音节拼接而成的）

合成读音	读音来源 1	读音来源 2	读音来源 3	读音来源 4
月亮(重轻)	月光	明亮		
牙刷(中重)	牙龈炎	刷牙		
西红柿(中轻重)	西风	鲜红	市场	
老头子(中重轻)	老王	头发	儿子	
光明正大(中轻中重)	光复中华	清明时节	政通人和	大方
二氧化碳(中轻中重)	二泉映月	臭氧层	生物化学	木炭

实验表明，将一个音节分为三种类型后，合成语音的自然度有了非常明显的改善。

3 结束语

文章构造了一个典型的汉语文语转换系统，它实现了最基本的从文本到语音的转换功能。为了提高输出语音的自然度，作者还从两个方面进行了改进尝试，取得了较好的效果。今后的研究将集中在如何提高句子一级的自然度，主要内容包括：对句子进行完整的韵律分析，生成抽象的韵律结构；构造韵律建模模块，将抽象的韵律结构转换成具体的韵律参数；提出更好的语音合成方法，构造更为合理的语音数据库，将韵律参数转换为自然的语音输出。（收稿日期：2000 年 5 月）

参考文献

1.Oliveira L C，Viana M C.A Rule-Based Text-to-Speech System for Portuguese.ICASSP，1992:73-76

2.刘源等.信息处理用现代汉语分词规范及自动分词方法.北京：清华大学出版社，1994

3.黄伯荣，廖序东.现代汉语.北京：高等教育出版社，1993

4.李建民.汉语 TTS 系统.清华大学工学博士学位选题报告，1997

5.谌卫军.汉语文语转换系统中的韵律结构生成.清华大学工学博士学位选题报告，1998