汉语文语转换系统地址映射算法的设计与实现*

张大军, 陈肇雄, 黄河燕

(中国科学院 计算机语言信息工程研究中心,北京 100080)
E-mail: {djzhang,heyan.huang}@263.net
http://www.cclie.com

摘要: 针对多样本文语转换系统中的语音合成实时性问题,提出了对合成系统语音库的改进策略和语音单元之间相似度的计算方法,在此基础上设计并实现了查找语音单元的地址映射算法.实验表明,地址映射算法和音库的重新组织有效地提高了合成系统的实时性.

关 键 词: 文语转换;语音库;地址映射算法
中图法分类号: TP391 文献标识码: A

近几年来,随着计算机技术的发展,语音合成技术日益成熟.它的广泛应用促进了合成系统向更深层次的发展.目前,语音合成的方法众多,各有优劣,根据研究的方法可分为发声器官的参数合成、声道模型参数合成和波形拼接合成 3 种[1].

发声器官参数合成[2]是对人的发声过程进行直接的模拟.它通过一些参数来确定声道的截面函数,进而计算声波.声道模型参数语音合成[3]主要是基于声道谐振特性来合成语音.这两种方法出现在语音合成技术发展的早期,合成音质较差.

目前,波形拼接技术[4]在合成的清晰度和自然度方面有明显的改善,音质也有很大的提高,所以语音合成技术越来越趋向于使用这种方法.它的实现原理主要是在系统中存放适当的基本语音单元(基元),合成时经过解码、波形拼接、平滑等处理,输出自然的语音.对这类合成系统而言,基元的选择十分重要.它一般是从自然语流中切分出来的,这样可以尽可能地保留基元的韵律特征;假如在合成过程中,语音单元不能满足对输出语音的要求,我们就通过算法对语音单元进行修改,所以,这种方法有较大的灵活性.

汉语是一种典型的声调语言[5],选用单音节作为基本语音单元是合适的.目前,针对一个音节建立多个样本已经成为提高汉语语音合成自然度的有力手段,但是,随着样本的增多,系统的音库结构和基元地址的计算方法日益成为制约合成实时性提高的瓶颈.本文针对华建文语转换系统 HJ-TTS(HuaJian text-to-speech system)合成的实时性问题提出了音库的组织方法和对两个语音片段相似性度量的计算方法,并在此基础上设计了基于地址映射的语音单元检索机制.实验表明,这些方法有效地提高了合成系统的自然度和合成的实时性.

1 问题分析

在单样本文语转换系统中,一个汉语音节只和语音库中的一个基元相对应,所以,在语音单元拼接中不存在对样本的选择问题.当这个基元的韵律特征与期望值不相吻合时,就用算法进行调整,但这会带来合成音质下降的问题[1,2].所谓多样本语音合成,主要是指在音库中对一个音节存放多个不同的样本,在合成时尽可能直接用

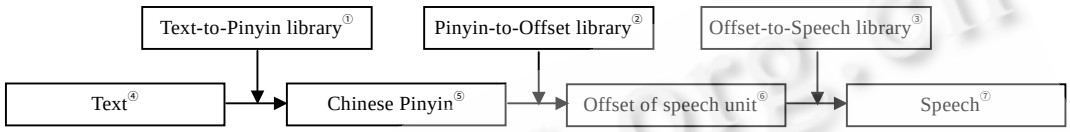
* 收稿日期: 2000-04-05; 修改日期: 2000-07-20
基金项目: 国家自然科学基金资助项目(69835003)

作者简介: 张大军(1973－),男,山东泰安人,博士,工程师,主要研究领域为自然语言处理,文语转换;陈肇雄(1961－),男,福建蒲田人,博士,研究员,博士生导师,主要研究领域为自然语言处理,机器翻译;黄河燕(1963－),女,湖南攸县人,博士,研究员,博士生导师,主要研究领域为自然语言处理,机器翻译.

音库中的样本进行拼接,从而避免算法调节.为了保证合成效果,HJ-TTS 采用了多样本合成的方法,由于音库中有丰富的样本,系统的自然度和可懂度有了明显的提高,但随之而来,在合成系统音库的组织和候选基元计算上都产生了许多问题.

(1) 音库的组织.多样本音库不再和单样本音库一样固定不变,更多的基元随着语音韵律研究的深入,不断加入到语音库中,所以对音库的设计必须考虑到它的动态调整和扩充.此外,怎样在音节的各个样本中迅速查找所需要的样本也成为设计音库时必须考虑的问题,这将涉及语音单元相似度的计算.

(2) 系统的实时性.样本的增多和韵律规律的复杂化导致系统在音库中查找所需语音单元时耗时更多,提高实时性已经刻不容缓.在传统的语音合成方式中,从文本转化为声音的过程可以简单地用图 1 来表示.从图中可以看出这个处理过程有如下缺点:



①文本、拼音对应库,②拼音、偏移量库,③偏移量、语音数据库,④文本,⑤汉语拼音,⑥语音单元偏移量,⑦声音.

Fig.1 Transformation from text to speech in traditional synthesis method

图 1 传统语音合成方式中文本到声音的转化

• 从文本输入到最终语音输出所用的拼音和偏移量在本质上只是从文本的一种表示方式转化到另一种表示方式,这种转变并没有方便基元的查找与拼接,反而增加了时间开销.

• 在系统的运行过程中频繁查找系统的各个库,在查找过程中涉及大量的字符串匹配工作,而字符串操作的特性决定了合成速度不能满足实时性要求.

上面两个问题在本质上是一致的,它们都是由于样本的增多而引起的,因此在多样本合成系统中必须以语音库为核心重新考虑这些问题.在合成系统中,基元相似度计算方法决定了音库中语音单元是怎样根据韵律特征进行区分的,而音库的组织方法在很大程度上决定了基元的查找算法.正是由于这个原因,在 HJ-TTS 中,我们将音库组织和映射算法两者进行综合考虑,提出了解决策略并取得了较好的效果.

2 实现策略

在 HJ-TTS 系统中,我们对音库的组织方法进行了改进,提出了语音片段之间相似度的计算方法.一方面,相似度的计算对于语音单元的选取和音库的组织具有重要的意义;另一方面,系统采用地址映射算法计算基元在音库中的地址,它将字符串匹配代之以数值计算,提高了合成系统的实时性.

2.1 音库组织与相似度计算

在多样本语音合成系统 HJ-TTS 中,我们同时考虑了音库中语音单元的建立和对音库中语音单元的选取规则.自然语流中的句子可以有不同的韵律表达,但是通过实验观察可知,对句子单个音节的韵律特征在一定范围内的变动并不影响整个句子的韵律特征和自然度.所以,只要是在韵律参数允许的范围内,就可以用音库中的语音单元代替自然语流中的语音而不会影响它的韵律.正是基于以上思想,系统在设计样本时要考虑尽可能地用所选样本覆盖这个语音所能出现的场合.虽然在理论上可以断定这种覆盖总是可以找到的,但在实际中,我们的音库仅仅是这个覆盖的一个真子集.目前,HJ-TTS 系统中语音单位的选取主要是根据时长和基频两个因素.因为同一个音节的候选样本很多,所以在音库的结构上与单样本合成系统相比发生了明显的变化.本系统音库中除了包含基本的语音数据之外,还包含了对这些样本的使用规则和默认处理方式.图 2 是这个音库的基本组织形式图.

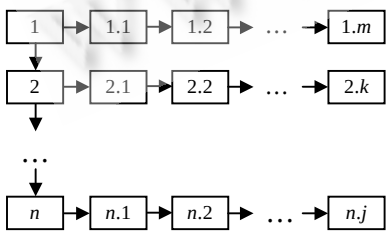


Fig.2 The basic structure of speech database

图 2 音库的基本结构

图 2 中左边的节点为音库中一个音节的开始标志节点,而它所链接的右边的节点是一个音的多个样本.样本的多元化带来了单个样本结构的变化,对于图 2 中的一个样本而言,其结构形式如下:

〈音节单元〉::=该语音单元的名称;
语音单元的大小;
〈Cond〉:=Cond(Not Emph);
Cond(End of phrase);
Cond(Before word “x”);
Cond(After word “y”);
Cond(Change tone);
…
〈音节单元结束〉

在这个音库中不仅记录了一个语音单元的大小和名称等信息,更重要的是记录了这个语音单元的应用条件和使用规则.在合成系统使用一个音节的样本进行拼接时,首先找到这个音节的起始节点,而后根据所有的 Cond 条件判断,在当前的情况下应当使用哪一个样本作为拼接的候选样本.从第 1 个样本开始,如果找不到满足条件的样本,则考察下一个样本的使用条件,在这个音节的所有样本条件都不满足的情况下,就使用默认的样本进行拼接.所以,在一个音节的多个样本之后有一个默认语音单元.这个样本的组织形式如下:

〈音节单元〉::=该语音单元的名称;
音节单元的大小;
〈Cond〉:=NULL
〈音节单元结束〉

对于 Cond 条件的判断,在 HJ-TTS 系统中没有使用字符串匹配,而代之以距离测度[6]计算的方式.距离测度是评估两个语音片段相似程度的一种方法,其定义如下:

定义 **1(**距离测度**)**. 假设源语音片段为 S,将它的韵律特征参数加权后组成一个特征矢量(S_i),目标语音片段为 T,也将它的韵律特征参数加权后组成一个特征矢量(T_i);S 与 T 的距离 $D(S,T)$可按欧氏距离定义如下:

$$D(S,T)=\{\sum_i (S_i-T_i)2\}^{1/2}.$$

$D(S,T)$反映了 S 与 T 之间的相似程度,也可以认为是 S 与 T 的相关系数.

距离测度可以帮助合成系统对音库中的语音单元进行聚类重组,通过保持语音库中任意两个基元之间的距离不大于某个预先定义的阈值,我们可以删除音库中的一些相似语音单元以减少冗余,阈值的大小定义将影响到语音库的规模.

一个基元的 Cond 条件通过影响基元特征矢量的一个或者几个分量反映在距离测度中.基元的选择是依据 Cond 条件计算每个样本与期望语音片段之间的距离测度,然后在语音库中根据距离测度的大小进行选择,这是多样本语音合成技术的关键.一般我们将合成中所需的语音片段称为目标语音片段,而将语音数据库中的基元称为候选基元.在 HJ-TTS 中选择的基元是否合适是从以下两个方面进行考虑的:

(1) 候选基元与目标语音片段的距离测度

查找语音数据库的链表,得到一个音节所对应的所有样本.假设这些样本的集合为$\{S\}$,计算每一个 S 与目标语音片段 T 的距离测度 $D(S,T)$.

(2) 候选基元与其之前的基元的平滑联接程度

假设基元 U 是基元 V 之前的基元.那么,基元 U 与基元 V 的平滑联接程度 $C(U,V)$用 U 的最后一帧的 F_0 参数 U_j 与 V 的第 1 帧 F_0 参数 V_k 的欧氏距离来表示:

$$C(U,V)=\{\sum (U_j-V_k)^2\}^{1/2},\qquad U_j,V_k \text{取矢量形式.}$$

假设候选基元的集合为$\{S\}$,候选基元 S 之前的基元为 U,综合 $D(S,T)$及 $C(U,S)$,如果 $S'\in\{S\}$,且满足

$$D(S',T)+C(U,S')=\min_{S\in\{S\}}\{D(S,T)+C(U,S)\},$$

那么,选择基元 S'作为最合适的语音基元进行拼接.在这里,S 中包含了默认样本,这是因为如果 S'不是默认单元,则 $D(S',T)+C(U,S')$一定比默认样本计算出的距离测度小,所以这与默认样本的作用并不矛盾.

系统采用这种音库组织方式是因为,一方面基元的选取比较容易,另一方面语音库的扩充更加灵活,当总结出一个新的样本及其使用规则以后,可以方便地将一个新的样本加入语音库.从本质上讲,这个音库是开放的,我们可以对音库进行动态的组织与调整.

2.2 地址映射算法

语音库的组织是对音节的多个样本有效地进行组织并加以利用,但随着样本数目的增加,系统在庞大的音库中查找所需基元的时间开销也在增大,合成系统必须根据音库的组织方法设计更快的查找算法,以满足实时性的要求.为了解决这个问题,HJ-TTS 采用了地址映射的计算方法,这与传统的基于拼音流的语音合成相比,合成速度极大地提高了.

HJ-TTS 中采用的基于地址映射的搜索和查询算法将字符串匹配转化成数值计算,从而摆脱了繁琐的字符串匹配操作.下面概述这一算法的主要思想.首先定义一个从无声调的汉语拼音集合到一个整数集合的映射关系,汉语拼音的音码就是与该拼音对应的整数.例如"shi"有 4 个声调(shi1,shi2,shi3 和 shi4),但是它所对应的音码只有一个.这样,我们就建立起一个从所有汉字到它所对应的音码的映射.HJ-TTS 中是用连续的整数来表示汉字的音码.在这个映射的基础上,用一个音码库取代以往汉字和拼音组成的字典拼音库,这个音码库存放的是汉字和它对应的音码.这样,全部汉字和音码之间的关系在音码库中表示为

汉字 m n.

其中左边表示的是汉字,右边是该汉字对应的音码.对于同音字,它们对应的音码必然相同.事实上,汉字右边的整数 n 相同与否反映了汉字的发音是否相同.

除了音码库之外,系统还需要一个记录基元在语音库中位置的偏移量库.为了快速查找基元,系统对偏移量库也进行了相应的改动.汉语是典型的声调语言,所以,偏移量库在记录基元在语音库中的位置时要相应地反映出其声调.汉字一般都有 4 个声调,另外,加上在连续语流中的音变现象以及为了方便将来对系统音库的扩充,我们在本系统中设定了每个拼音有 8 个声调.这些声调在偏移量库中反映为以下组织结构:

…

偏移量 m_1——音码 m 的第 1 个声调对应样本的偏移量

…

偏移量 m_8——音码 m 的最后一个声调对应样本的偏移量

…

偏移量 j_k——音码 j 的第 k 个声调对应样本的偏移量

…

偏移量 n_1——音码 n 的第 1 个声调对应样本的偏移量

…

偏移量 n_8——音码 n 的最后一个声调对应样本的偏移量

…

有了上面的基础,系统可以直接用数值计算的方法而不是通过字符匹配得到基元在语音数据库中的位置.该算法的描述如下:

算法. CalculateOffset()

Step 1. 输入合成的文本 w,j=1.

Step 2. 对文本进行分词 $w=w_1w_2w_3w_4\ldots w_n$,其中 $w_1,w_2,..,w_n$ 为分词单位.

Step 3. 对 $w_j(j=1,2,\ldots,n)$进行处理,将其转化为 mn,其中 m 为汉字音码,n 为声调.

Step 4. 计算 $offset=m*8+n$;$j=j+1$;

Step 5. 在偏移量库中直接定位到 offset 位置,得到该音码和声调所指定的第 1 个语音单元样本在音库中的

位置 position.

Step 6. if (该样本的使用条件与选取规则一致 or 该样本是最后一个) goto Step 9.

Step 7. 将 position 定位到该音码和声调指定的下一个样本.

Step 8. goto Step 6.

Step 9. 从音库 position 位置开始读取语音单元数据.

Step 10. 根据语音合成规则对该语音单元进行拼接处理.

Step 11. if (j<n+1) goto Step3.

Step 12. 合成结束.

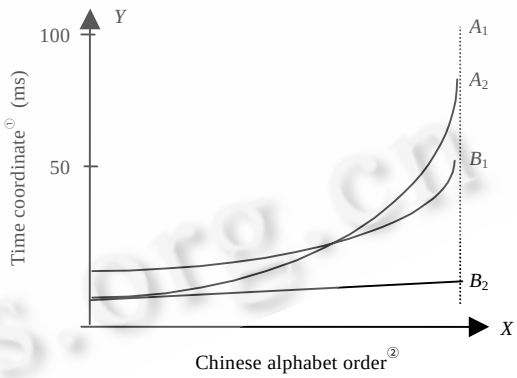
与传统的合成算法相比,上面的算法只进行了一次对文本的搜索就将分词单位转化为音码表示(将转化过程放在这里仅仅是为了叙述方便,该过程在语音合成的分词阶段已经完成),其他计算都是基于数值计算而不是进行文本匹配,这样,合成的速度就得到了提高.

从上面的分析可以看到,HJ-TTS 中音库的改进和地址映射算法是结合在一起提高合成速度的,在音库结构的配合下,地址映射算法加快了从文本到音节第 1 个样本的查找速度,在一个音节的多个样本中,通过距离测度的计算保证了系统迅速定位到所要求的基元.

3 实验结果分析

地址映射和字符串匹配在多样本语音合成中都可以实现对语音单元的查找功能.在 HJ-TTS 中采用的是基于地址映射的查找方式,而它的前期版本采用了字符串匹配的方法,并且基元在音库中按顺序存放.我们就这两个版本进行了比较分析.

在结果分析中,我们采用的是重叠字测试法.设合成 n 个“zuo”音花费了 100 毫秒,以 n 为基准,用合成 n 个相同汉字所花费的时间进行比较分析,如图 3 所示.



①时间坐标,②汉字字典顺序.

Fig.3 Result of two synthesis methods

图 3 两种合成方法的分析结果

在图 3 中,横坐标表示汉字,以字母的顺序从左向右排列,纵坐标表示合成所用的时间.图中 A_1 和 A_2 基元的选取是基于字符串匹配的,而 B_1 和 B_2 基元的选取是基于地址映射的方式.A_1 和 A_2 的区别是,A_1 将分词所花费的时间考虑在内,而 A_2 仅仅是从分词结果到语音生成所花费的时间.B_1 和 B_2 的区别也是如此.从图中可以看出,A_2 和 B_2 相比,A_2 随着汉字在字典中位置的后移,合成所花费的时间也在增多,这与在字符串匹配过程中,汉字在字典中的位置越向后,其匹配所需时间越长是一致的.但是,B_2 基本上呈线性关系,这说明地址映射算法受汉字在字典库中位置的影响很小.对于 A_1 和 B_1 的关系也是如此.B_1 和 B_2 相比,在 B_1 中所花费的时间随着汉字位置增长很快,这是因为在 B_1 的分词中进行了字符串匹配.从上面的实验结果来看,在采用地址映射算法之后,合成的实时性有很大的提高.

4 结束语

基于以上方法建立的文语转换系统 HJ-TTS 能够实时地进行语音输出,具有较好的自然度和可懂度.如今,我们已经将这个文语转换系统成功地应用到华建语音翻译系统中.虽然该系统能够很好地满足语音翻译的要求,但是它使用的音库较为庞大.今后,我们将针对如何减小音库规模这个问题,对语音单元的选取规律做进一步的实验和探讨.

致谢 中国科学院计算机语言信息工程研究中心的陈肇雄和黄河燕研究员对本文的工作给予了细心的指导,许洁萍博士和胡春玲博士对本文的完成提出了很多有益的建议,在此一并表示感谢.

References:

[1] Chu, Min. Research on Chinese TTS system with high intelligibility and naturalness [Ph.D. Thesis]. Beijing: Institute of Acoustics, the Chinese Academy of Sciences, 1995 (in Chinese).

[2] Syrdal, A., Bennett, R., Greenspan, S. Applied Speech Technology. New York: CRC Press, 1995.

[3] Fant, G. Acoustic Theory of Speech Production. Netherlands: the Hague Press, 1960.

[4] Liu, Qing-feng, Wang, Ren-hua. A new speech synthesis method based on the LMA vocal tract model. Chinese Journal of Acoustics, 1998,17:153~162

[5] Yang, Shun-an. The Chinese Speech Synthesis Technique Oriented Acoustics-Phonetics. Beijing: Society Science Literature Press, 1994 (in Chinese).

[6] Zhang, Sen. Research on text-to-speech systems based on SC-analysis [Ph.D. Thesis]. Beijing: Institute of Computing Technology, the Chinese Academy of Sciences, 1998 (in Chinese).

附中文参考文献:

[1] 初敏.高清晰度高自然度汉语文语转换系统的研究[博士学位论文].北京:中国科学院声学研究所,1995.

[5] 杨顺安.面向声学语音学的普通话语音合成技术.北京:社会科学文献出版社,1994.

[6] 章森.基于 SC 文法的文语转换系统的研究[博士学位论文].北京:中国科学院计算技术研究所,1998.

Design and Implementation of Mapping Address Algorithm in Chinese Text-to-Speech System*

ZHANG Da-jun, CHEN Zhao-xiong, HUANG He-yan

(*Research Center of Computer and Language Information Engineering, The Chinese Academy of Sciences, Beijing* 100080, *China*)

E-mail: {djzhang,heyan.huang}@263.net

http:// www.cclie.com

Abstract: In order to solve the real-time synthesis problem in a multi-sample text-to-speech system, the strategy of improving the structure of speech database is discussed and a method of calculating the similarities between two speech units is put forth in this paper. On the basis of these, a mapping address algorithm to locate the address of speech unit is brought up. The experimental results show that mapping address algorithm and reconstruction of speech database improve the real-time performance of text-to-speech system.

Key words: text-to-speech; speech database; mapping address algorithm

* Received April 5, 2000; accepted July 20, 2000
Supported by the National Natural Science Foundation of China under Grant No.69835003