计算机科学1998Vol. 25№. 6

TDPSOLA 文语转换技术和建立汉语音节单元库的研究

A Method of TDPSOLA and Research of Building Chinese Syllable Voice Bank

荀恩东　蔡　萌　李　生　赵铁军

（哈尔滨工业大学计算机系　哈尔滨150006）

摘　要　A method and principles of building voice bank used in text-to-speech system (TTS), which based on TDPSOLA method and made use of Chinese syllable with tone, were described. The quality of voice bank has direct relation to the clearness and naturalness of TTS system, specially to the system based on the syllable concatenation. In this TTS system, some method were adoptted to ensure the quality of syllable voice bank, such as making detail classifications to some synthesis unit, abstracting recording example sentences, recording the voice bank in the quasi-naturalness speech and abstracting the valid synthesis unit, etc. The result showed that the naturalness of the TTS system had been improved greatly.

关键词　Voice bank, Syllable, Text-to-Speech, TDPSOLA

1　引言

文语转换系统具有将输入的文字自动地转换成语音的功能，广泛地应用于文稿校对、语音应答系统、电子邮件的语音服务等领域。

根据合成模型的不同，可以将文语转换技术划分为三类[3]，即基于声学模型、发声模型和对自然语音进行编码模型的文语转换技术。基于声学模型的文语转换技术以语音产生的源-滤波器理论为基础，最典型的是共振峰合成方法；基于发声模型的文语转换技术是通过模拟人的发声过程来合成语音；对自然语音进行编码主要由对波形进行压缩存储和解码回放组成。

与其它语言相比，汉语普通话中的音节主要有以下几个特点[5]：

(1)普通话中最自然和最基本的语音单位。一般来说，普通话的音节和汉字是一对多的映射。也就是说，一个音节是形意的结合体。

(2)数量小。普通话中的音节比较少，不计声调时只有四百多个，而英语多达4030个，俄语也有2960个。

(3)声学相对稳定性。虽然在音节相连的语流中，同样存在着音节之间的协同发音效应，但是与其他语种相比，在汉语中这种效应的作用范围比较小，人在听感的基础上可以分出一个个音节，每个音节的声学表现有相对的稳定性。

根据上述特点，本文采用第三类文语转换技术，以汉语普通话带调音节为基本音节单元。然而在合成语句时，如果简单地将一个个音节串接起来，得到的合成系统虽然能够保持较高的清晰度，但自然度和流畅性方面往往不尽人意。原因在于简单的波形拼接技术虽然能较好保持音段特性，但无法根据上下文来调整超音段(基音、时长和能量)特征。

法国电信公司在1986年提出了基音同步叠加(PSOLA)技术[6]，既能保持原始发音的主要音段特征，又能在拼接时灵活调整基频、时长和强度等超音段的特征，有效地保证了以音调为基本音节单元的合成系统的质量。同时，如上所述，汉语普通话中的音节具有音段特征比较稳定、超音段特征变化复杂等特点，很适合采用基于 PSOLA 技术的波形拼接方法来合成。

PSOLA 技术也有它的不足之处。具体说，采用 PSOLA 技术需要存储预先录制好的音库，占据较大的硬盘空间；以“不变应万变”的音节单元录制的好坏直接关系到文语转换系统的清晰度和自然度；采用 TDPSOLA 技术还需要计算语音单元的基频、确定基音的标记和元辅音之间的界限。

本文系统地介绍了 TDPSOLA 技术，提出了一

种计算基音和基音标记的半自动标注的方法，着重论述了建立语音音节单元库的策略和方法。

2 TDPSOLA

2.1 介绍

基音同步叠加(PSOLA)最早是由法国电信公司在1986年提出的，这种技术并不能合成语音本身，它只是使语音单元内能够平滑地连接，使语音单元的基音，时长和能量得到改变，它的出现改变了以前文语转换中普遍使用的LP方法。

各种PSOLA算法的工作原理基本是相同的，即在元音段通过与基音同步的汉宁窗(Hanning)、在辅音段加固定音隔的汉宁窗，把一个语音单元分成一系列的短时语音信号，这些短时语音信号再按目标语音特点，重新组合成合成语音。汉宁窗大小的选取直接影响合成语音音质。语音信号的基音的提高和降低，通过改变短时语音信号的重组间距，时长的延长和缩短通过增加和删除一些短时语音信号来实现，原始信号的能量也可以通过对短时语音信号的加权进行改变。PSOLA的一个最简单的实现是在实域中，即TDPSOLA，它也是PSOLA各种算法中计算效率最高的。TDPSOLA可由下列公式表示：

$$S_i(n)=X(n)*W(n-i*T_0) \tag{1}$$

$$S(n)=\sum_{i=-\infty}^{\infty}S_i(n-i*(T-T_0)) \tag{2}$$

式中：X(n)为原始语音信号；W(n)为汉宁窗；$S_i(n)$为短时语音信号；T为要修改的目标语音周期；T_0为原始语音周期。

2.2 实现

在具体实现时本系统采用以下步骤：

(1)取原始待合成的音节单元信号和合成目标音节单元相关的合成参数(包括原始音节单元信号的元、辅音之间的界限、基音标记信息；目标语音单元的时长、基音、能量改变权值等)。

(2)对原始波形按基音标记加二倍于基音周期的汉宁窗，形成短时波形系列。

(3)按照目标音节单元的时长和基音，生成元音段的基音标记，辅音段按照时长变化比例作相应的调整。

(4)在目标音节单元的基音标记处，选择一个适当的短时波形。

计算：(OrgWaveLen/SynWaveLen) * SynWavePitchMarkPos取与该值最近的原始基音标记的一个短时波形。如图1所示。

(5)在元音段、进行短时系列加权叠加；在辅音段，以一定的频率等间隔地进行叠加。

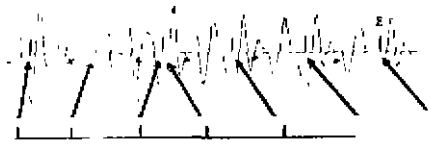

图1 目标语音单元的基音标记处短时波形选取

2.3 基音周期的估计

基音周期是语音信号中最重要的参数之一。是基于TDPSOLA算法进行文语转换基础数据，基音估计的方法很多，在本系统中采用了基于短时自相关函数的方法[7]，根据加窗的短时语音帧来估计的基音周期。这种方法与其他方法相比，具有算法简单、求解精度高等优点。其主要步骤如下：

第1步：计算该语音单元波形的最大采样值MaxPeak。

第2步：进行中心削波

$$Y(n)=C(X(n)) \tag{3}$$

$$C(x)=\begin{cases}0 & -MaxPeak<x<MaxPeak\\ x+MaxPeak & x<MaxPeak\\ x-MaxPeak & x>MaxPeak。\end{cases} \tag{4}$$

第3步：通过加汉宁窗截取计算数据。取汉宁窗长度为20毫秒。

第4步：计算该数据段的自相关函数

$$R(i)=\sum_{n=0}^{N-i-1}Y(n)*Y(n+1)\quad N=1..窗长度$$

第5步：计算R(i)/R(0)，取第一大峰值对应的i值，则该段的基音为11025/i。

第6步：重复1～6步，计算所有段的基音。

2.4 基音标记的标注

由2.2所述，在实现TDPSOLA时，在原语音单元波形上施二倍基音周期长度的汉宁窗，该窗必须位于声带主激励波形的波峰或波谷(本系统采用了后者)，这些点称为基音标记。如果完全由手工标注语音信号的基音标记，工作量巨大，所以，研制自动标注算法和良好的辅助标注界面显得尤其重要。本系统采用了自动标注、人工矫验的方法。如图2所示，步骤如下：

第1步：由试听和观测语音波形确定元、辅音界限。

第2步：在元音段，应用2.3节中的算法计算基音曲线，根据该曲线预标记基音标记。

第3步：以预标记的基音标记为中心，应用该点对应于基音曲线的基音值三分之一的长度，进行前后搜索最小的波谷，标注为基音标记。

第4步：用第3步得到的基音标记，修正下一个预标记点，重复第3步。

第5步：如果算法标注有误，对错误点进行修正。以此为基点，用类似方法对此基点以后的各基音标记点进行自动修改。

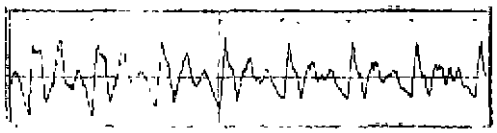

图2 基音标记

3 音库的录制

3.1 基本合成音节单元

基本合成音节单元分为以下6类：

(1)全部汉语带调音节1302个(对一部分音节进行了再分类，见3.2)。

(2)轻声音节513个。

(3)儿化音节63个。

(4)英语26个字母音节。

(5)俄语32个字母音节。

(6)日语82个平假名和片假名。

3.2 部分音节再分类

考虑词在语流中的变化，根据大语料的统计结果，把全部汉语带调音节(1302个)中的一部分音节(468)分为以下几类：

(1)独立音节。往往做为单独使用汉字(词)的音节，如 zai4(在)，bai(把)等。

(2)首音节。经常做为二字词组第一个音，如 shi4(世界)，bu4(部门)等。

(3)尾音节。经常做为二字词组第二个音，如 zhai2(大寨)，gou4(机构)等。

(4)普通音节。

对输入待合成文本进行汉语分词，对于以下几种切分结果进行了处理。含单个汉字的词(字)，如果在独立音节中存在，则优先采用独立音节语音单元。对二字词组前后音节分别优先查找是否在首音节和尾音节音节单元中，如果查找成功，合成时相应使用首音节和尾音节音节单元，否则使用普通音节。对于切分出的三字词组和四字词组，应用词组库中的二字词组进行再切分，对于切分结果，进行类似的处理，选取适当的音节单元。本系统词组库中，词组的最大长度为4。

在对汉语普通话带调音节再分类时，采用了语料库统计的方法。第一步，对系统词组库中的汉字进行拼音标注。对于二字词组区分首音节和尾音节，对三字和四字词组应用二字词组对其切分(对于其中的错误，人工改正)，如果能够切分出二字词组，则对此区别首音节和尾音节，其他的标注为普通音节。第二步，依据该词组库对语料进行分词处理，对分词结果中不含在词组库中的单独汉字，标注为独立音节。第三步，统计独立音节、首音节和尾音节的出现频次。

在统计时，语料摘自《人民日报》，共1000多万字。统计结果见表1。在表中以独立音节为例，当选择前123个高频出现的音节时，占总独立音节出现次数的百分之70。本系统以百分之95为阈值，截取相应出现的高频音节。

表1 音节再分类语料统计结果

数目 / 百分比	独立音节	首音节	尾音节
70	123	161	153
75	156	193	182
80	200	232	217
85	256	283	261
90	337	351	324
95	468	461	426

3.3 在"准"自然语流中录制合成语音单元

在录制系统音节单元库时，如果采取"一音一录"的方法，难免会造成录制的音节单元与自然语流中的发音出现较大的偏差。本系统采用了在"准"自然语流中录制合成音节单元策略。即提供含代录音节的背景句子，在概率意义上，使包含代录音节的词组(或字)在语料中出现较高频次。音库录制者在连续语音背景下"特别留意"要录制的有效音节。抽取参考句子步骤如下。首先，对词组库中的词组进行语料统计，依据出现的频次由高到低进行排序，并标注首音节和尾音节标记。以此为依据，抽取含已统计出来的首音节和尾音节词组。对于独立音节做类似处理，抽取出参考汉字。然后，抽取含这样处理得到的参考词组的汉语句子，对于余下的普通音节采用了一种"贪心"算法，抽取最大覆盖普通音节的最小文本长度的参考句子。听辨试验表明，采用在"准"自然

语流中录制合成音节单元的策略，对系统的合成质量有较明显的改进。

3.4 抽取有效语音单元

以音节为合成单位进行波形连接的文语转换方法，音节单元录制的质量对系统的总体效果有重大的影响，本系统采用了以下策略以保证语音库音节单元的质量，即重复录音，概率择优[6]。对于每个音节单元，录制8遍，在建立合成音节库时，选取其中一个作为音节单元，选取的策略如下：

a)舍弃时长小于所有该类语音单元平均时长80%的语音单元，这样处理的目的是，一方面，这些被删除的语音单元的音质往往不太好，另一方面，在合成时如果使用这些被删除的语音单元，在调整时长时，时长的增加有可能超过限定的时长增长因子(目标语音单元时长/原始语音单元时长)的阈值，造成语音信号的失真或不自然。在本系统中，时长增长因子设为1.22。

b)计算所有语音单元平均采样能量，舍弃平均采样能量大于该值的语音单元，这样处理可以减小在合成时声音的剧烈起伏振荡。

c)对于进行以上两个步骤后剩下的语音单元，通过试听和合成测试，最后，在每一类型语音单元选定一个作为系统使用的语音单元。

(下转第89页)

(上接第4页)

主要参考文献

[1] H. T. Kung et al., Systolic arrays (for VLSI), Sparse Matrix Proc. Society for Industrial and Applied Mathematics, 1978

[2] H. T. Kung, Why systolic arrays, IEEE Computer, vol. 15 Jan. 1982

[3] S. Hauck, Multi-FPGA Systems, PH. D Thesis, Dept. CS&E, University of Washington, 1995

[4] B. K. Fawcett et al., Reconfigurable Processing with Field Programmable Gate Arrays, in Inter. Conf Application Specific Array Processors, 1996

[5] M. Annaratone, et al., The Warp Computer: Architecture, Implementation and Performance, IEEE Trans. Computer, 36(12)1987

[6] 杨超峰，颜玲，傅宇卓，胡铭曾，支持MPEG-2标准的ME芯片设计，体系结构年会'97，1997.10

[7] W. Shang et al., On uniformation of affine dependence algorithms, IEEE Trans. Computer, 45 (7) 1996

[8] Y. Wong et al, Transformation of broadcast into propagation in systolic arrays, J. of Parallel and Distributed Computing, 14(2)1992

[9] J.-C. Tsay et al., Design of efficient regular arrays for matrix multiplication by two step regularization, IEEE Trans Parallel Distrib System, 6(2)1995

[10] W. Li et al., A singular loop transformation framework based on non-singular matrices, Int J. Parallel Programming, 22(2)1994

[11] H. J. Lee et al., Automatic generation of modular mapping, ASAP'96, 1996

[12] D. I. Moldovan et al., Partitioning and mapping algorithms in fixed size systolic arrays, IEEE Trans. Comput, vol. C-35, 1986

[13] W. Shang et al., On time mapping of uniform dependence algorithms into lower dimensional processor arrays, IEEE Trans. Parallel and Distributed Systems, 3(3)1992

[14] S. Y. Kung, VLSI Array Processors, Englewood Cliffs, NJ: Prentice Hall, 1988

[15] G.-J Li et al., The design of optimal systolic arrays, IEEE Trans. Comput. vol. C-34, 1985

[16] K. Ganapathy et al., Optimal Synthesis of Algorithm-specific Lower-Dimensional Processor Arrays, IEEE Trans. on Parallel and Distributed Systems, 7 (4)1996

[17] V. K. Prasanna Kumar et al., Designing linear systolic array, J. Parallel and Distrib. Computing, vol. 7, 1989

[18] A Darte, Regular partitioning for synthesizing fixed-size systolic array, J. VLSI Integration, vol. 12, 1991

[19] K. Ganapathy, et al., Design a scalable processor array for recurrent computations, IEEE Trans. on Parallel and Distributed Systems, 8(8)1997

[20] A. L. Fisher et al., Synchronizing large VLSI processor arrays, IEEE Trans. Computers, C-34(8) 1985

[21] J. Teich, et al., Scheduling of partitioned regular algorithms on processor arrays with constrained resources, ASAP'96, 1996

如果 R1,R2 包含,则只需满足 a<=X∧X<=b;

如果 R1,R2 相等,则只需满足 a<=X∧X<=b(或 X<=c∧X<=d);

如果 R1,R2 相交,则只需满足 c<=X∧X<=b。

不难看出,利用区间之间的包含,相交和相等关系可以使得查询公式化简并且在语法级上存在简单的化简算法。

把本节讨论的结果扩充到其它线性序集上是没有困难的。

结束语 我们采用本文所述算法实现一个数据库引擎,采用此引擎实现了一个数据库系统并命名为 eBASE。

考虑 xBASE 是目前应用最广泛的数据库系统,eBASE 同 xBASE 数据一级兼容,因此 eBASE 属于 xBASE 家族。

eBASE 同 C 语言是无缝耦合的,是一个全 C 语言数据库系统。由于许多系统都提供 C 语言接口,因此这意味着 eBASE 可以同这些系统实现无缝耦合。

由于在 eBASE 的实现中采用本文所述优化算法,因此它的内核很小,仅70-80KB 左右。但它的逻辑查询速度较快,尤其是在模糊查询方面。

实验还证明此系统在网络环境下性能特别优良。

目前 eBASE 已经实现了网络和 WINDOWS 版本,并且实现了同 Visual BASIC,AUTO_CAD 等系统的连接。同时,我们已经采用此系统为国内外客户实现了十余个应用系统,取得了良好的效果。

由于 eBASE 完全内嵌到了 PROLOG 系统中,因此 PROLOG 程序员完全可以通过 PROLOG 程序透明地访问外部数据库中的数据,这给 PROLOG 程序员带来极大的方便。目前此系统已在加拿大 NT 的合作项目中采用,效果良好。

参 考 文 献

[1] D. Ullman, "Principle of Database and Knowledge-base Systems", Computer Science Press, 1988

[2] J. W. Lloyd, "Foundation of logic Programming", Springer-Verlag, 1984

[3] S. Ceri, et al., "Logic Programming and Database", Springer-Verlag, 1990

[4] J. Bocca, "On the evaluation strategy of Educe", Proc. ACM Singmod, Washington, May 1986

[5] S. Ceri, et al., "Interface relational database and PROLOG efficiency", raporto interno No 85-23

[6] 李磊,等,"PROLOG-DBMS 实现中的子句间优化技术",软件学报,1995年第三期

[7] 周龙骧,"数据库实现技术"中国地质出版社,1990

[8] 李磊,"PROLOG—DBMS 实现方法",计算机学报,92年第三期

[9] 施伯乐,"KBASE-P:一个知识库程序设计",软件学报,1995 07 544-550

(上接第69页)

结束语 本文介绍了 PSOLA 算法的原理,给出了 TDPSOLA 实现的基本步骤,计算基音曲线和标注基音标记方法,虽然 TDPSOLA 方法可以改变语音的某些超音段性质,但是音节单元的性质在不同的自然语流中有很大的不同。我们采取了针对音节在词组和句子的位置信息,对常用音节再分类,增加音库中音节单元的个数。在抽取参考词组时,考虑该词组在语料中出现的频次,应用高频词组作为参考词组。这样处理,系统的合成质量在概率角度就得到了一定的提高。

参 考 文 献

[1] 张家录、吕士南等,"汉语文语转换的研究",信号处理,1989年第一期

[2] 倪宏、李昌立、莫福源等,"汉语大词汇量合成系统的研究",第六届全国语音图像通信信号处理学术论文会议集,1993

[3] 初敏,"高清晰度高自然度汉语文语转换系统的研究",中国科学院声学研究所博士论文,1995

[4] Cai Lianhong, Zhao Qiaofeng, Wang Yong, "Research of Prosody Modification Based on PSOLA in Chinese TTS", 1995

[5] 杨顺安,"语音合成与语音学研究",中文信息处理,1992

[6] Robert Edward Donovan, "Trainable Speech Synthesis", The Dissertation of the University of Cambridge, 1996

[7] 杨行峻、迟惠生,《语音信号数字处理》,电子工业出版社,1995