ZNUTTS 语音合成系统的实现方案研究

On ZNUTTS Speech Synthesis Software Application System Implement Scheme

赵建民　朱信忠

（浙江师范大学计算机科学与工程学院　金华321004）

Abstract　The representative technology of the Speech Synthesis System is the TTS (Text_to_Speech), especially the Chinese TTS is important for our nation information development. Our ZNUTTS system is based on PSOLA(Pitch Synchronous OverLap Add) arithmetic,the CTTS's principle and application are discussed in this paper

Keywords　Speech synthesis,TTS,CTTS,PSOLA,PDA

1　语音合成技术概论

当前，语音合成的代表技术是文语转换技术 TTS（Text-To-Speech），TTS 实现将文本自动转换成语音并加以输出。它在语音文稿校对、多媒体用户界面 MUI（Multimedia-User-Interface）、计算机电话集成 CTI（Computer-Telephony-Integration）应用、交互式语音应答 IVR（Interactive-Voice-Response）系统、互联网中的语音服务、信息发布系统、信息家电及掌上电脑的中文发音等方面都有着广阔的应用前景。

我国汉语语音合成技术的研究开始于八十年代初，相对于国外语音合成技术的研究来说起步较晚。在国内科研院所汉语 CTTS 技术的研究中，最突出的是清华大学，中科院声学所，中国科技大学、语言所等单位。清华大学的 Sonic 系统是一个基于波形编辑的优秀的文语转换系统。该系统基于词库进行分词，并且根据语音学研究的成果建立了语音规则，对汉语中的某些音变现象进行了处理。整个系统采用 PSOLA 算法修改超音段语音特征，提高了言语输出的质量。TTS 系统的核心是语音合成，主要有参数合成和波形编辑方法。在实施语音合成时，要处理好基元及参数的提取，实时合成及输出的平滑滤波等一系列问题。清华大学的 Sonic 语音合成系统对我们浙江师范大学语音与模式识别研究室自行研制开发的汉语文语转换系统 ZNUTTS 有很大的启发。ZNUTTS 是一个有一定实用价值的中英文混合文本智能朗读系统，我们的 CTTS 系统与其他文语转换系统相比具有独特的音库压缩和人工自然音色转换等特点，而正是由于这些自身所具备的特点使我们的 CTTS 技术向嵌入式操作系统的移植具备了一定的条件，如果与掌上电脑，信息家电、PDA（Personal Digital Assistant）及网络开发商等合作，可为其增加语音输出功能，推动计算机应用的普及。同时还可公开我们研究的标准 CTTS 接口，与其它智能接口软件融合，建立和谐的人机交互方式，提高计算机的应用效率。另外，在我们自己独立开发的"浙江师大指纹考勤管理系统"中，成功地将 TTS 技术应用于指纹比对成功语音智能提示系统，在 TTS 技术向其他硬件系统上的移植（如 TTS 在电话语音卡上的开发）方面等也积累了一些经验。

我们认为，实现一个语音合成系统要考虑三个方面的因素：合成语音的质量，即其可懂性和自然程度；合成语音的流畅程度，包括词汇的多样性以及语音中的重音，音调等性质；系统的复杂程度，反映在硬件的运算量和存储空间的要求上。实际的语音合成系统的设计往往是在这三个因素之间做折衷考虑，而且对某个语音合成系统的性能也应该在以上三个因素确定的指标范围内进行评价。

2　一个最简单的语音合成系统

一个最简单的语音合成系统的结构如图1所示，这个系统可用于电话应答服务中，例如用户拨的无用电话号码为0579-2281234，系统可以给出提示："The number you have dialed，0579-2281234，has been disconnected"。另外，为了使合成的语音更接近自然语言，每个数字都有几个版本，这样可以使电话号码的首位数有较长的持续时间和较高的音调，听起来更清晰、自然一些。

图1中的系统对十个数字，即0到9的合成采用共振峰合成法或线性预测合成法。共振峰合成法是考虑到人发音时实际上是由喉咙声管对声音进行调制，反映在语音的频谱上即出现若干个共振峰，依次排列在基音谐波的后面，根据各个声音对应的共振峰的位置及

其强弱可以重建信号，线性预测合成是对语音做 LPC 分析，提取 LPC 系数，在解码端利用 LPC 模型，用白噪声做激励得到合成语音。为了提高语音合成的质量，可以做以下的改进：改进声管激励源的模型，这对女声尤为重要；改进音韵学规则及系统的语言分析功能等。

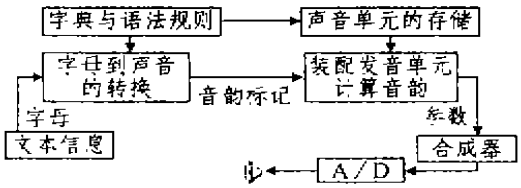

图1 一个简单的语音合成系统结构

3 ZNUTTS 语音合成系统软件特点及功能简介

我们开发的汉语 CTTS 系统 ZNUTTS(可实现中英文混合朗读)，其整个软件分成几个主要功能模块：文本编辑、文语转换朗读模块、拼音管理、词组管理、语音微调、控制面板、脚本编写及帮助和标准安装文件。

任何一个计算机文语输出系统都要有语音数据库，用于存储语音基元。但是其形式不尽相同。在构建语音基元数据库时要重点考虑两个问题：基元的选择和存储形式。本软件的语音数据库全部为标准真人普通语，男女声可选择，音质好、语音清晰、悦耳。我们下赞成采用机械的模拟电子发声。语音数据的存储方式是数字化的语音波形数据，这些数据都经过自适应差分脉冲编码调制 ADPCM、编码和解码易于实时实现。

这里介绍一下 ZNUTTS 语音合成系统的主要功能，本软件是典型的基于 Windows 的32位工具软件，适合运行在 Microsoft 的 Windows 9x/ME/NT/2000 环境下。我们的软件采用的是 Dsound 的混音播放技术，因此使用它机器上必须装有 DirectX5 0或更高版本。它有两种工作状态：一种是在使用键盘输入数字和中英文字符时，计算机声卡可以跟随录入的字符即时发出相应的语音提示，用户可使用桌面上的声音软开关，随时开启和关闭声音。另一种是使用软件的自动文语转换功能将中英文混合文本文件朗读出来，供用户欣赏或进行校对。该软件可进行多项设置，以满足不同用户的需要。如：可以任意设置朗读速度，可在比正常速度快一倍和慢一倍的范围内无级调整，任意选择朗读的标点符号，该软件支持国标汉字一二级字库所有的汉字，运行速度较快，支持剪贴板和编辑窗口。朗读时光标自动跟随文字移动。该软件能够较好地处理多音字的读音(如"睡觉"、"发觉"等"觉"字的不同读音)，用户能够自己向词库中追加多音字词组；汉字和英文单词的读音能够有机无缝地结合(音调几乎一样)、对英文单词有连续功能，可提供中文、英语长篇文章全文朗读、支持对任意句子、单词进行部分语音复读。本系统还能够同音乐进行完美的结合，在朗读文章的同时还可以播放用户所喜欢的音乐，声音不会发生"冲突"，本系统对朗读的各项参数具有开放性，用户可以调节所有有关朗读的参数，包括朗读的音量、频率、每个字的读音、每个读音的长度以及多音字词组控制等，用户还可以自己添加语音库，本系统支持多种文件格式，如：TXT、RTF、HTM、DOC 等。本软件的使用非常简单方便、清楚明了。

当然，本软件功能还很欠缺，很多地方还需要改进，尤其需要在朗读时的音色、音调真人自然化上多下功夫，要解决文章整体阅读平均匀速的问题，另外我们的语音合成系统在实现抑扬顿挫的效果上及解决成语连读的音效问题上都有待提高，还暂时无法达到"真人"聊天一样的逼真效果，这一系列问题还有待我们进一步认真学习和研究。

拼音管理是本软件提供给用户自行调整拼音汉字音库的。使用方法为：用户在汉字栏中输入要修改或添加的汉字，并按下[输入]，系统将自动找出相应的拼音放入拼音栏中，这时用户就可以修改它了，修改完了如果想要生效就按[更新]，否则关闭设置窗口并退出，或按[应用]保存到文件中，要注意的是本软件中汉字的"罗马化"采用的是"拼音标准"。由于在计算机里输入四声比较困难，因此汉字拼音与普通的拼音有所区别。在本软件中用拼音后面的数字来表示四声。例如："您"→"nin2"，"我"→"wo3"，"他"→"ta1"。"浙江师范大学"实际读音为："zhe4 jiang1 shi1 fan4 da4 xue2"。

词组管理是本软件提供给用户的又一项自行调整的功能，主要是调整多音字的词库。具体使用方法为：在词组栏里输入所要添加或修改的词语，然后按下[输入]按钮，属性中将出现单独汉字和它的读音，按动[←]、[→]来调整当前修改的汉字，修改完后若想生效就按[更新]，否则关闭设置窗口退出即可。最后按[应用]保存到文件中。

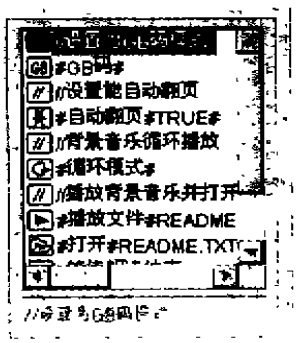

图1

简单地讲，语音微调就是用来调整每一汉字的播放长度的，它可以精确到1/1000秒。微调的具体方法为：首先在拼音栏里选择一个拼音代码，这时在长度栏里就会出现这个拼音代码所对应的长度值，修改长度值，按[应用]保存，按[重置]可恢复到默认状态。

图1为已经开发出的部分脚本控制窗口，本软件脚本是以行为单位的，上下文之间没有直接关系，解释器只对单独的一行进行解释。

4 ZNUTTS系统语音合成关键技术的研究

当前语音合成技术主要采用发音器官参数合成、声道模型参数合成和波形编辑合成等三种技术，其中我们选择的是波形编辑语音合成技术。该技术是在E. Moulines 和 F Charpentier 于80年代末提出的基于时域波形修改的语音合成算法的基础上，并在基音同步叠加技术PSOLA(Pitch Synchronous Overlap Add)方法的推动下发展起来的。波形编辑语音合成技术就是直接把语音波形数据库中的波形相互拼接在一起，输出连续语音流。这种语音合成技术用原始语音波形替代参数，而且这些语音取自自然语音的词或句子，它隐含了声调、重音、发音速度的细微特征，合成为语音清晰自然，质量普遍高于参数合成。基音同步叠加PSOLA技术把基音周期的完整性作为保证波形及频谱平滑连续的基本前提。该算法按以下三步实施：对原始波形进行分析，产生非参数的中间表示；对中间表示进行修改；将修改过的中间表示重新合成为语音信号。由于这种修改的参数不同，又分为TD-PSOLA、FD-PSOLA和LP-PSOLA。

实用的文语转换系统的关键技术不是文本语音替换而是变音、变调的处理，包括轻声和儿化处理。如：处理词组时，由于前后音节协同发音而引起的声调变化，若两个上声相连，前面的上声近似阳平，上声在非上声之前变半上声等都属于变调处理；处理重音音节的音长、音强和音高变化都属于变音处理；轻声处理是指处理音节的弱化问题，如"妈妈"一词中的第二个"妈"字读轻声；儿化处理是指处理音节的儿化问题，如"纸条儿"。ZNUTTS系统采用的是时域基音同步叠加PSOLA合成算法，下面我们简单地介绍一下语音韵律的修改、声调的修饰、语调的修饰等技术的原理。

4.1 语音韵律的修改

为了实现对韵律的修改，首先要实现对语音基本数据的修改，基音的修改过程如下：首先进行加窗分析，取Hanning窗，其表达式为：$H(m)=1+\cos(2\pi m/N)$，$|m|\leqslant(N/2)$，其中N为窗宽。窗叠率为2～4。当基音模式的基音周期数与目标模式相同时，是不需要修改的。我们采用简单叠加法进行叠加合成，$\bar{S}(n)=\sum_q a_q S_q(n)$，其中 a_q 取窗 h_q 与平均窗宽的商作为能量补偿。我们可以利用基音周期的相关性，但仅能在波形的相似段进行时长变化的修改和增/删处理，这是由汉语的复合韵母和声调决定的。时长变化的百分比计算公式为：

$$percent=100\cdot\frac{datalength}{changedlength}-100$$

其中 *datalength* 为波形数据的原始长度，*changedlength* 为改变基频后的波形数据长度。若 $percent>0$，则时长增加；若 $percent<0$，则时长缩短；若 $percent=0$，则时长不变。假设数据稳定段的周期内平均幅度为 *amp*，则当幅度降低时，加权系数 $delta=\frac{amp-n\times 3}{amp}$。幅度升高时，加权系数 $delta=\frac{amp+n\times 3}{amp}$。当幅度提高时，首先要计算数据被加权后是否会失真，如若失真，则根据最大允许幅度修改系数 *delta*。

4.2 声调的修饰

声调主要由调型、调高和调长构成。调型是声调的表现形态，它是声调音高曲线的形状，反映声调调值的变化过程，普通话的声调调型主要分为平调、拱调和平拱结合调。在连续语流中，声调的调型发生变化，形成多种调型。分析声调的调型曲线，可以用下面的参数来描述，语调的修饰实际上就是改变下面的几个参数：H—音调上限。调型曲线的最高点；B—音域下限。调型曲线的最低点；I—调值中心的基频；R—调域；D—调型时长。调型的变化是指声调的模式和调长的变化，对于汉语普通话来说，如果把轻声也算作是一种声调，则有阴平、阳平、上声、去声、轻声5种基本声调。各种目标基频的模式是预先设置的，基数据结构为：int pitchForm[*m*][*n*][*k*]。其中，*m* 为调类序号，也代表声调编码，5种基本的声调的最大调类数是 *m*，*n* 为目标调型序号，每一个目标调型是一串数字，每个数字表示一个基频周期所包含的采样点数，其上限和下降由语音库的平均音域决定，*k* 为基频周期数。设某词后去声音节的基音标注为指针 *pp*，其周期数为 *pp→number*，音域上下限为 *pp→maxpitch* 和 *pp→minpitch*，可求出其音频中线为：

$$middlepitch=FS/\frac{1}{2}\cdot(FS/pp\rightarrow\max\ pitch+FS/pp\rightarrow\min\ pitch)$$

其中，*FS* 为采样频率。另设 *FormMax*，*FormMin*，*FormMiddle* 分别为预置目标基频模式的音频上限、音频下限和音频中线，则可求得该音节进行基频模式修改

后的基音周期序列为：

$form[i]=middlepitch+Ratio\cdot(pitchForms[m][n][j]-FormMiddle)$

$i=0,\cdots,pp\rightarrow number-1,j=0,\cdots,toneNumber-1$

其中 $Ratio$ 的定义为：

当 $pitchForms[m][n][j]<FormMiddle$ 时，

$$Ratio=\frac{middlepitch-pp\rightarrow max\_pitch}{FormMiddle-FormMax},$$

否则，$$Ratio=\frac{middlepitch-pp\rightarrow min\_pitch}{FormMiddle-FormMin}$$

下标 i 和 j 的关系为：

$$j=i\cdot\frac{toneNumber}{pp\rightarrow number}$$

这是一种均匀的线性插值。其中 $toneNumber$ 为声调编码的归一化周期数。要提高声调值，首先必须提高基频。表现在时域上，就是周期内的采样点数减少。设每变一级对应为周期内采样点数减少 m 个，则：

$form[i]=form[i]-m\cdot n,i=0,\cdots,pp\rightarrow number-1$，

对应于降低调值，表现在时域上，就是周期内的采样点数增加。设每变一级对应为周期内的采样点数增加 k 个，则 $form[i]=form[i]+k\cdot n,i=0,\cdots,pp\rightarrow number-1$。要进行调域的修改，首先要利用音域中线的公式求出 $middlepitch$。假设每级变化的采样点数是 $interval$，则目标基音模式为：

$form[i]=middlepitch+(form[i]-middlepitch)\times Ratio,i=0,\cdots,pp\rightarrow number-1$

其中，如调域加宽时

$$Radio=\frac{pp\rightarrow\max\ pitch-middlepitch}{pp\rightarrow\max\ pitch+n\cdot interval-middleptich}$$

对于调域上限的变化，上浮可看作是调域加宽，下移可以看作是调域缩窄，则上述公式仍然适用，但在计算目标基音模式的公式中，右边的 $form[i]$ 要小于 $middleitch$，即只对较高基频进行处理。调域下限的变化与上限的变化类似，不同之处在于：一是要把下限的上浮看作是调域缩窄，下移看作是调域加宽；二是计算目标基音模式的公式右边的 $form[i]$ 要大于 $middlepitch$，即只对较低的基频进行处理。

4.3 语调的修饰

语调表现在语句上，而声调表现在音节上。语调是说话人的心理、情感、态度的反映，一般为传递陈述、疑问、感叹、命令等信息。声调类型是词的结构的一部分，是对音高的调节；而语调是对音高的再调节。对于长句来说，一般可分为数个短语，短语自成单元，具有完整的规则语调短语。语调的变化伴随着句子的节奏、速度的改变，其声学特征表现为音高、音域时长、音强和停顿的变化，我们从语调的声学特征，来实现语调的修饰，进而把语调分成以下的一些模式：

(1)平语调　它没有高低升降的字调变化。句中各音节的声调基本上是原状，只是句尾的音高是趋降的，一般用于陈述或说明。

(2)升语调　它的末尾语调上扬，声学征兆是音高上升，调域扩大、调长加长，升语调用于命令、疑惑、惊讶等，另外，在并列陈述句中，并列成分表现为升调，只有在句末才是降调。

(3)降语调　它的末音节略短，略下降，降语调表示说话人的情绪不高，一般用于感叹。

(4)曲语调　它是一种双向语调，主要由升调和降调混合组成。主要用于表述较复杂的情感和语义。如："是我的错，你没有错。"

上述语调模式可反映一些句型的语调，汉语句子发音的最基本的语调就是陈述句、疑问句和感叹句。陈述句的语调是平稳趋降的，只要陈述句的句尾不是被强调部分，其基频就相对要低一些。疑问句往往分为有疑问词和没有疑问词两种，当句子中有疑问词时，语调与陈述句相同，当句子中没有疑问词时，句子语调是上升的，句尾的语调也有所上升。对于感叹句，若句子中有感叹词，则语调与陈述句基本相同，当没有感叹词时，语调尾部是下降的、句尾的字调也有所下降。长的句子可以看作是由短句或语调短语组成，当它们之间有类似于并列、等候下文的关系时，除句尾之外，其它各种语调短语尾部的调值都呈上升趋势，并要适当停顿；否则，语调短语尾部的调值应呈下降趋势，但要比句尾降得少。

参考文献

1　钟玉琢，蔡莲红，李树青，史元春. 多媒体计算机技术基础及应用. 北京：高等教育出版社，1999.6

2　郝杰，吴元清，郑榕. 实用多媒体技术及其C语言实现. 北京　电子工业出版社，1995.11