第 25 卷 第 7 期 电 子 与 信 息 学 报 Vol.25 No.7
2003 年 7 月 JOURNAL OF ELECTRONICS AND INFORMATION TECHNOLOGY July 2003

基于 DSP 的嵌入式汉语文语转换系统及其 VLSI 设计方案 [1]

戴礼荣 王仁华

(中国科技大学电子工程与信息科学系 合肥 230027)

摘 要 简要讨论了嵌入式文语转换 (ETTS) 系统的概念，介绍了一个基于 DSP 实时实现的嵌入式汉语文语转换 (ECTTS) 系统。基于 DSP 实现的结果，分析了 ECTTS 系统的 VLSI 实现方案，提出了基于动态内存管理的 ECTTS 系统前端处理 VLSI 实现方案，基于解码语音帧的 ECTTS 系统后端合成 VLSI 实现框架并对 ECTTS 系统的 VLSI 实现中的存储器及总线结构进行了讨论。

关键词 嵌入式文语转换， DSP, VLSI, 基音同步叠加

中图号 TN391.42, TN4

1 引 言

自基音同步叠加 (Pitch Synchronous Overlap-Add, PSOLA) 方法的提出后 [1]，基于波形拼接的较高自然度文语转换 (Text To Speech, TTS) 系统逐步走向实用阶段。基于 PSOLA 技术的法语、德语、英语、日语、汉语等语种的文语转换系统也相继推出。但此类系统一般针对 PC 机或工作站硬件平台的应用，为获得高质量的合成语音，此类系统需要几十兆，乃至上百兆字节的存储容量的资源支持。由于当前信息技术发展的新的特点，这类 TTS 系统在实际应用中面临很多的挑战。

当前信息技术发展的一个新特点是：无线互联网技术等其它通信技术使得人们通过个人智能移动终端设备可在任何时间和地点获取互联网上的信息，而个人智能终端设备越来越小型化从而使通过语音交互信息更为必要，即更为迫切要求将通过各种信息获取方式获取的文本信息在个人智能移动终端设备上转换为语音。这种新的特点提出了对嵌入式文语转换 (Embedded Text To Speech, ETTS) 系统的要求。 ETTS 系统是一种能够融入到资源有限 (运算能力和存储空间均相对有限) 的个人或其它智能终端设备上的文语转换系统，它应能和智能终端设备成为一体。从嵌入方式来看， ETTS 系统可分为软件嵌入式 (如嵌入到无线通信手机，手持个人计算机，个人数字助理等具有一定有限资源的设备上) 及硬件嵌入式。硬件嵌入式是指文语转换系统用一个芯片或一个芯片组实现后嵌入到应用系统中。显然，硬件嵌入式对应用系统的资源要求很低。从 ETTS 系统实现的方式看， ETTS 系统可分为分布式后端实现方式及前、后端捆绑实现方式。前、后端捆绑实现方式是将 ETTS 系统的前端处理和后端处理作为一个整体实现，输入是文本，输出是语音；分布式后端实现方式是将文语转换系统的前端处理和后端处理分开实现， ETTS 系统仅完成后端处理，其前端处理由语音服务器等完成，这在分布式语音合成 (Distributed Speech Synthesis， DSS) 中是广泛采用的实现方式。

由于以上讨论的当前信息技术发展的新特点，国外对 ETTS 系统已经展开了研究，也推出了相应的实际系统 [2,3]。但嵌入式汉语文语转换 (Embedded Chinese Text To Speech, ECTTS) 系统的研究尚未见报道。本文将介绍一个采用前、后端捆绑实现方式的基于 DSP 的硬件 ETTS 系统。在 ECTTS 系统的 DSP 实现基础上，本文还将进一步分析和探讨 ECTTS 系统的 VLSI 实现方案。

[1] 2002-01-14 收到， 2002-05-28 改回
国家自然科学基金资助 (批准号： 69975018)

2 基于 DSP 的 ECTTS 系统

2.1 ECTTS 系统资源要求

2.1.1 存储容量 ECTTS 系统对存储资源要求主要由前端资源库和后端音库决定。本文中实现的 ECTTS 系统的前端资源库包括规则库，词典，词语类别库，词性库，词性同现概率库，总容量为 260.5kbyte 。其中规则库为 19.1kbyte ，词典为 202kbyte(包括了 6762 条单字词条， 14239 条多字词条) ，词语类别库为 17.5kbyte ，词性库是 18.8kbyte ，词性同现概率库为 3.1kbyte 。

本文中实现的 ECTTS 系统的后端音库在采用 G.723.1(5.3kbps) 进行压缩处理后 [4] ，在不同的采样率下，其大小为： 16kHz 采样，为 754kbyte; 11kHz 采样，为 559kbyte; 8 kHz 采样，为 441kbyte 。所以，本文中实现的 ECTTS 系统所需的存储资源在 701kbyte~1Mbyte 之间。

后端音库的设计在 ECTTS 系统中占有重要地位。为节省存储空间资源，采用了小音库设计技术，并同时在基元库设计和音库录制上充分考虑了基频和共振峰的协同发音的影响。所设计的高质量后端音库对保证 ECTTS 系统所合成的语音质量具有重要地位。

2.1.2 运算复杂度 系统的运算复杂度主要由 PSOLA 算法和 G.723.1 解码算法的运算量决定。 G.723.1 解码算法在 DSP(Digital Signal Processor) 上实现运算复杂度 [4] 为 2–3MIPS (Million Instructions Per Second) ，加上 PSOLA 算法的运算复杂度，整个系统的运算复杂度在 5MIPS 以内。

2.2 基于 DSP 的 ECTTS 系统硬件结构

为使设计的 ECTTS 系统硬件为 VLSI 实现提供有价值的设计仿真， ECTTS 系统的 DSP 实现应采用定点 DSP 。为此，考虑到尽量提高系统结构紧凑， ECTTS 系统采用 Anaiog Device 公司的定点 DSP 芯片 ADSP-2185M 。该 DSP 的运算速度达到 33MIPS ，同时其在片程序存储器 (PRAM) 和数据存储器 (DRAM) 各 16kword ，因而，能够期望在不扩充外部 RAM 的情况下实现 ECTTS 。此外， ADSP-2185M 有两个支持双缓冲和硬件 A/μ 律语音编解码的串口，其中串口 0 可用于外挂一个 D/A 转换器，串口 1 通过其可编程的特点可用软件实现一个用于嵌入式连接的满足 RS-232 协议的接口，如图 1 所示。另外， ADSP-2185M 有一个自举 DMA

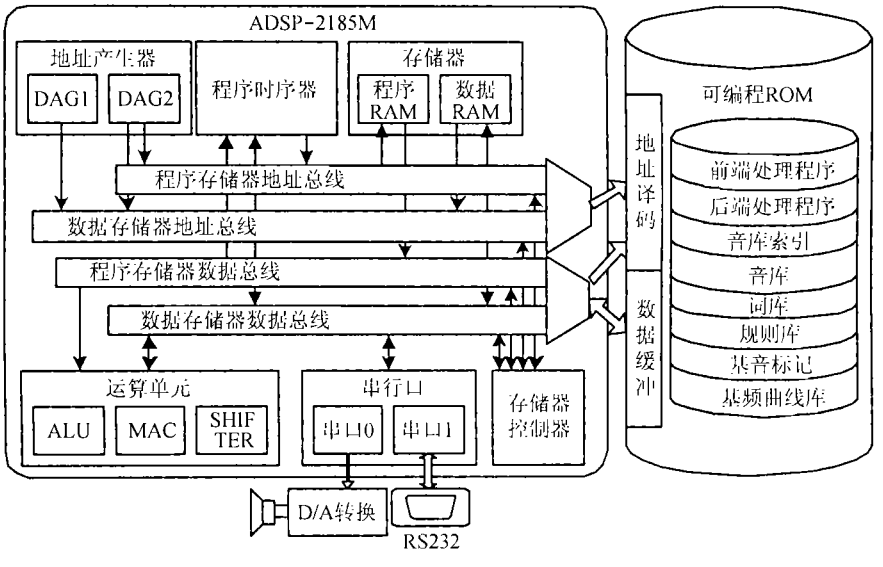

图 1 基于 ADSP-2185M 的 ECTTS 系统的硬件结构

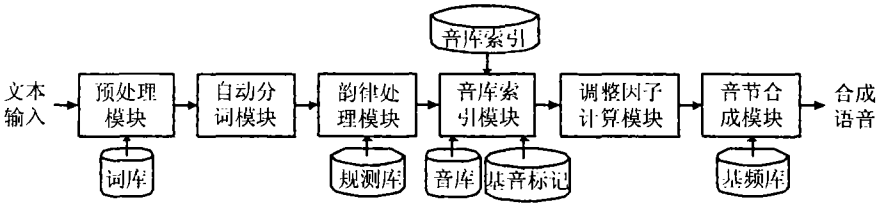

图 2　基于 DSP 的 ECTTS 系统的软件结构

接口，用于程序和数据的装载，因此，ECTTS 系统的程序和系统的数据库可存放在图 1 中的可编程 ROM 中．根据上节的分析，用来存放前后端程序以及词库，音库，规则库等数据信息的可编程 ROM 的容量可选为 2Mbyte．

2.3 基于 DSP 的 ECTTS 系统软件结构

ECTTS 系统的软件由 6 个模块组成 (图 2)，它们分别是：预处理模块，自动分词模块，韵律处理模块，音库索引模块，调整因子计算模块，音节合成模块．其中前 3 个模块是实现 ECTTS 系统的前端分析，后 3 个模块是实现 ECTTS 系统的后端处理，6 个模块之间是串行工作．预处理模块对通过 RS-232 串行接口接收到的文本字符串根据标点符号进行语句分割，对数字字符进行替换处理等文本预处理；经过预处理的结果由自动分词模块按前向最大匹配分词法进行分词处理，并进一步生成语法分析树．韵律处理模块根据所生成语法分析树添加韵律调整信息，从而得到包含韵律信息的前端文本分析结果．在 ECTTS 系统的后端处理中，首先由音库索引模块根据前端文本分析结果中的拼音信息搜索音库，获得经过 G.732.1 压缩的语音码流并解码，以及获得相应的基频标记信息；再由调整因子计算模块计算音节内的时长、幅度、基频调整因子；最后，在音节合成模块基于 PSOLA 算法合成语音．

2.4 基于 DSP 的 ECTTS 系统实际占用的硬件资源

基于 ADSP-2181M 的 ECTTS 系统实际占用的硬件资源是：占用 DSP 的运算资源为 5 MIPS，其中 G.723.1 解码器占用 3MIPS，PSOLA 算法占用 1MIPS，其它为 1MIPS；程序代码占用 DSP 程序空间 (PRAM) 资源为 12kword(24bit/word)，占用 DSP 数据空间 (DRAM) 资源为 15.3kword(16bit/word)；片外存储器资源占用 1.5Mbyte，其中用于存储 DSP 自举程序占用 0.5Mbyte，数据占用 1Mbyte．

3 ECTTS 系统的 VLSI 实现方案

很多专用集成电路 (ASIC) 的设计均采用首先在通用 DSP 上对所要实现的功能和算法进行实时仿真，从而对专用集成电路的设计方案和实现复杂性提供真实的设计依据．有时，如专用集成电路采用基于通用 DSP 核的实现方案，这种在通用 DSP 上的实时仿真的结果可直接应用到该种方案的集成电路的设计和实现．本节将基于第 2 节中的 ECTTS 系统的 DSP 实现结果分析和讨论 ECTTS 系统的 VLSI 实现在运算和控制流程上的一些关键和主要考虑因素，从而探讨 ECTTS 系统的 VLSI 实现方案．

3.1 基于动态内存管理的 ECTTS 系统前端处理 VLSI 实现方案

ECTTS 系统的前端分析算法的一个重要特点是：数据结构和内存操作比较复杂．前端分析本质上就是对一棵语法分析树的操作，逐步为它添加各种层次韵律信息．为此，就需要一个内存管理机制，实现在语法分析树的操作过程中的内存块分配和回收功能，即实现对树节点动态地添加和删除．

通过 DSP 仿真，证明对所要求的树结构的动态内存管理，采用一种基于双向链表的实现方案较为简单有效．该方案中，整个树形结构由若干条双向链表和相应的树节点组成，其中一条链表对应一个父节点下的所有子节点 (图 3)．此基于双向链表的树结构动态内存分配实现方案可

只利用一个地址寄存器 (I) 和另一个地址修改寄存器 (M) 来实现。I 寄存器用于标示空闲内存块的起始地址。当出现内存请求时，用 M 寄存器记录请求的内存长度，同时把起始地址为 I，长度为 M 的一块内存初始化，并添加到相应的链表中，最后再用 M 更新 I(即 $I \leq I+M$)，以供下一次的内存分配。当出现内存释放时，程序修改链表，标记该链表所对应节点被删除，但暂时不作内存回收。直到语法树分析完毕，得到所需结果后，再将 I 指回语法树分析前所指的位置，实现当前语法树分析占用内存的释放。

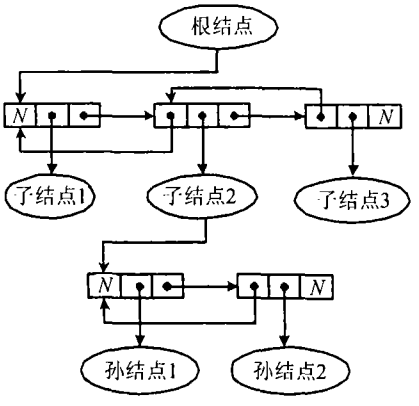

图 3 基于双向链表的语法分析树结构

显然，在所采用的基于双向链表的动态内存管理实现方案中，非常容易实现对树的遍历及树节点动态的添加。但树节点动态的删除实际上是一种伪删除技术，即在一个语法树分析过程中，只标记该节点被删除，而其占用的内存暂时不释放。其优点是动态内存管理实现简单，而通过 DSP 仿真证明，在一个语法树分析过程中，所需动态删除的节点所占的空间不是很大。

3.2 基于解码语音帧的 ECTTS 系统后端合成 VLSI 实现框架分析

为了节省存储空间，音库中存放的只是原始语音的压缩码流，所以在后端合成时，首先要解码获得语音信号，然后才能使用 PSOLA 算法调整出最终的合成语音。在具有大容量存储器的硬件环境下，通常的做法是：先将整个汉语单音节解码完毕，并存放在一个很大的线性缓冲区内，然后用 PSOLA 算法生成合成语音，并将合成语音存放在另一个线性缓冲区内。当所有音节合成完毕后，再将合成句子整体播放出来。很显然，在采用 VLSI 实现后端合成时，应避免采用这种方案：首先，为提高 VLSI 的性能价格比，不应设计过大的内部随机存储器 (RAM) 存放整个汉语单音节的信号样本，更别说整个句子了，所以应采用一边解码，一边合成，一边播放 (D/A 输出)；其次，语音合成的实时性要求也不可能允许在整个音节解码完毕之后才开始合成新的语音样本。

通过 DSP 仿真实验表明，ECTTS 系统后端合成 VLSI 实现的一种可行的解决方案是：采用基于 G.723.1 定长解码语音帧的实现框架。在该实现方案中，语音编解码算法是 G.723.1，它的帧长固定为 240 点 (30ms)。考虑到 PSOLA 算法的特点，此时 ECTTS 系统后端合成 VLSI 实现中只需设置若干长度为 240 的整数倍缓冲区：解码缓冲区 (长度 720word, 16bit/word)，合成缓冲区 (长度 480word, 16bit/word)，放音、工作双缓冲区 (长度各 240 word, 16bit/word)。其中，解码缓冲区用来存放最近 $n(n \leq 3)$ 帧解码语音，合成缓冲区则存放最近 $m(m \leq 2)$ 帧合成语音。合成语音的播放采用双缓冲方式，放音缓冲区存放的是正在输出的 240 点语音，而新合成的 240 点语音将放在工作缓冲区中，当播放完一帧语音后，两个缓冲功能互相切换 (图 4)。

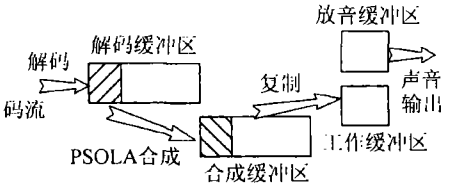

图 4　后端合成基于 G.723.1 定长解码语音帧的实现结构

ECTTS 系统后端合成 VLSI 实现的以上解决方案中，为满足语音合成的实时性要求，其中的语音解码操作，PSOLA 合成操作及放音 (D/A 输出) 操作之间的流程关系应是由放音 (D/A 输出) 操作驱动的：每当语音输出的两个缓冲区进行切换后，如果当前合成缓冲区中存有的合成语音少于 240 点，那么就激发 PSOLA 合成操作，直到其中的合成语音超过 240 点．这时，合成操作和解码操作都会暂停直到新的语音输出要求产生，而刚合成出来的一帧语音从合成缓冲区中清除，并转移到放音工作缓冲区中，处于准备播放状态．在 PSOLA 合成操作时，合成操作又驱动着解码操作：每一次的 PSOLA 合成都是以基音周期为单位进行，如果解码缓冲区无法提供下一个基音周期的语音，就要激发解码下一帧，直到所需语音已经具备．这时，解码操作会暂停，而合成操作则计算出下一周期的合成语音并放入合成缓冲区，当然，解码缓冲区中不再被使用的语音同时会被清除。

上述驱动机制构成了基于解码语音帧的合成实现框架．这种方法一边解码，一边合成，一边播放，合成语音的输出是以 1 帧为单位，并且同时存在的解码语音不会超过 3 帧．

3.3 ECTTS 系统的 VLSI 实现存储器及总线结构分析

ECTTS 系统的 DSP 仿真表明，为得到较高质量的合成语音，使合成语音具有可接受的合成语音清晰度和自然度，解码后的语音数据应达到 12bit 的精度，PSOLA 合成算法等需要 16bit 的定点运算精度．因此，ECTTS 系统的 VLSI 实现中的处理器应该是 16bit 的，其内部存储器 (RAM) 及相应的寄存器也应是 16bit 的．根据 ECTTS 系统的 DSP 仿真结果，ECTTS 系统所需要的运算量不是很大，因此，ECTTS 系统的 VLSI 实现总线结构可采用诺尔曼结构 (Norman structure)，当然也可采用哈佛结构 (Harvard structure)．

由于 ECTTS 系统的实现需要的各种数据库大约 1Mbyte，在 VLSI 实现中，可固化在 ROM 中．这些数据可按字节组织，也可按字 (16bit) 组织．从 VLSI 实现的性价比看，当然采用按字节组织为好．不过，由于 VLSI 实现中的处理器是 16bit 的，当采用按字节组织时，当对 ROM 中的数据读取时，需要将按字节组织的数据转换为 16bit 字长的数据．因此，VLSI 实现中需要考虑实现这种数据组织机制．

4 结　论

本文在讨论嵌入式文语转换 (ETTS) 系统的概念和分类的基础上，介绍了一个基于 DSP 实时实现的嵌入式汉语文语转换 (ECTTS) 系统．基于 DSP 的 ECTTS 系统的硬件结构紧凑，其尺寸大小是 5.0cm×5.0cm．其音节清晰度，单词可懂度以及句子自然度都达到了实际应用的水平．

研究该系统的目的一方面是该系统可作为汉语语音合成的一种嵌入式应用，另一方面该系统的研究结果为 ECTTS 系统的 VLSI 实现提供了有价值的仿真．本文基于此仿真结果，分析和讨论了基于动态内存管理的 ECTTS 系统前端处理 VLSI 实现，基于解码语音帧的 ECTTS 系统后端合成 VLSI 实现框架及 ECTTS 系统的 VLSI 实现存储器及总线结构．为 ECTTS 系统的 VLSI 实现提供了较为合理的实现方案．

参 考 文 献

[1] E. Moulines, F. Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication, 1990, 9(12), 453–467.

[2] Youhyeon Jeong, Sionghun Yi, A development of the stand-alone audiotex system, Communication Technology Proceedings, ICCT'96., Osaka, Japan, 1996, vol.1, 441–444.

[3] T. Ebihara, Y. Ishikawa, Y. Kisuki, T. Sakamoto, T. Hase, Speech synthesis software with a variable speaking rate and its implementation on a 32-bit microprocessor, IEEE Trans. on Consumer Electronics, 2000, 46(3), 887–895.

[4] 王仁华, 徐超, 戴礼荣, ITU-G.723.1 双速率语音编码器定点 DSP 实现, 信号处理, 1997, 13(3), 199–206.

AN EMBEDDED CHINESE TEXT-TO-SPEECH SYSTEM BASED ON DSP AND ITS VLSI DESIGN SCHEME

Dai Lirong Wang Renhua

(*Dept. of Electron. Eng. and Info. Sci., Univ. of Sci. and Tech. of China, Hefei* 230027, *China*)

Abstract The concept of Embedded Text-To-Speech (ETTS) system is first discussed. Then a real time realized Embedded Chinese Text-To-Speech (ECTTS) system based on DSP is introduced. Based on the DSP realization results, the VLSI design scheme of the ECTTS system is discussed. For the ECTTS system front-end processing, a VLSI design scheme based on the dynamic memory management is suggested. For the ECTTS system post-end processing, a VLSI design architecture based on the decoded speech frame is also proposed. The memory and the bus structure of the VLSI design of ECTTS system are discussed.

Key words Embedded text-to-speech, DSP, VLSI, PSOLA

戴礼荣: 男, 1962 年生, 博士, 副教授, 主要从事语音信号处理、语音编码与通信、人机语声对话的研究及 DSP 技术应用.

王仁华: 男, 1943 年生, 教授, 博士生导师, 中国通信学会会士、理事. 主要从事数字信号处理、语音通信、多媒体通信等方面的研究.