分
计算机世界/ ２００１年/ ０７月/ ０９日/ 第Ｄ２６版/
客户服务

语音合成的应用系统设计

清华大学计算机系人机交互与媒体集成研究所　　陶建华　　蔡莲红

合成语音是通过一个声学模块来具体实现的。早期的语音合成技术的研究，往往集中在语音合成算法本身，其研究的方法和手段与语音编码有很多相似之处。其声学模型的构筑，也多通过模拟人的口腔的声道特性来产生。其中比较著名的有Ｋｌａｔｔ的共振峰（Ｆｏｒｍａｎｔ）合成系统，后来又产生了基于ＬＰＣ、ＬＳＰ和ＬＭＡ等声学参数的合成系统。这些方法用来建立声学模型的过程为：首先录制声音，这些声音涵盖了人发音过程中所有可能出现的读音；提取出这些声音的声学参数，并整合成一个完整的音库。在发音过程中，首先根据需要发的音，从音库中选择合适的声学参数，然后根据韵律模型中得到的韵律参数，通过合成算法产生语音。

进入２０世纪９０年代以来，波形拼接（ＰＳＯＬＡ）的算法，越来越被广泛地应用在语音合成系统中。国内外的许多跨国公司和研究机构均投入了大量的人力和物力从事语音技术的开发，如Ｌ＆Ｈ、ＩＢＭ、Ｍｉｃｒｏｓｏｆｔ、Ｌｕｃｅｎｔ、Ｍｏｔｏｒｏｌａ等。清华大学计算机系在汉语语音合成的研究和开发中，也突破性地运用了人工神经网络、决策树、隐马尔可夫模型等方法。这些方法的运用，彻底改变了汉语语音合成研究的研究重点，使汉语语音合成的研究突破了早期重点对单纯算法的研究，而变成一个系统工程的研究。目前我国语音合成的整体研究和开发，迈上了一个全新的台阶，并处在了国际最先进的行列。

一、ＳｉｎｏＳｏｎｉｃ语音合成系统

ＳｉｎｏＳｏｎｉｃ是清华大学计算机科学与技术系和北京炎黄新星网络科技有限公司共同推出的汉语语音合成系统。它采用目前世界最先进的数据驱动技术，利用精心设计的语音语料库对ＴＴＳ系统进行训练，所得到的系统体现了连续、自然的语言特性，从而使系统发音自然、亲切。整个系统的核心技术包括：基于神经网络的韵律模型、基于ＨＭＭ的语音切分和选取模型、基于ＨＭＭ的多语种文本分析和语义分析、汉语语料库设计和标注、语音分析工具的研制等。

该系统的构成分为：用户编程接口以及ＴＴＳ内核两大部分，其中，内核部分又可以按照系统运作的不同过程分为多个子模块，包含了训练模块、文本分析模块、韵律生成模块、语音合成模块以及与语料库之间的通信协议等。同时，ＳｉｎｏＳｏｎｉｃ还考虑了不同类型用户对ＴＴＳ系统功能的需要，提供了丰富的编程接口。

该系统的工作过程如下：

用户提供文本并调用ＴＴＳ系统接口，文本首先被送入系统的文本分析模块，文本分析模块首先对用户输入的文本进行规格化处理，然后运用统计模型算法对其进行分词、分短语

类号：ＢＪ８２

、确定发音、分析标点符号或特殊符号等处理，同时，还要确定文本发音的轻重模式。经过文本分析后得到的参数，被送入到系统的韵律生成模块。在韵律模型中，首先通过统计模型的方法得到韵律中音节的音长和音强参数，然后通过优化的神经网络模型来确定音节的基频曲线，并得到音节停顿模式等信息。将这些信息和参数传送到系统的声学模块，系统的声学模块再根据这些参数，从音库中选择合适的语音单元，并采用PSOLA的方法生成最终的合成语音。

在整个系统工作的过程中，用户可以随时通过系统提供的接口，获得系统的内部状态，进行合成参数设置、随时中断或暂停系统等工作。

SinoSonic系统功能和指标有：可读字、词、句子、文章及标点、数字、运算符和英文字母，语音库覆盖国标一、二级所有汉字；能输出男声或女声；提供丰富的、合理的编程接口，方便用户进行二次开发；语音输出以句子为单位，按词汇停顿，能自动决定多音字的正确读音可随时改变声音的幅度（Volume）、基频（Pitch）、速度（Duration）、词间或句间停顿；读出时，可随时“暂停”、“恢复”、“终止”语音。

二、语音合成系统性能指标

语音合成系统的基本性能指标包括：可懂度、清晰度、自然度、汉字转拼音正确率（分词正确率）。

考虑到实际应用，还有系统的数字、姓氏、特殊符号等方面的处理能力、跨平台处理能力以及语音合成的速度（指单位时间内，通过语音合成系统生成语言的音节数，或语音合成同时支持的并发请求个数）等。

SinoSonic除了满足一般意义上的特性外，还有许多独有的性能，如：

（1）即时性：TTS技术实时完成文本到语音的转换，它实现信息的即时传送。

（2）并发性：TTS技术与电信网络结合，同时处理多个呼叫请求，它实现信息的并发传送。

（3）适应性：TTS系统能在不同操作系统平台下运行，支持Windows9x、Windows2000、Linux和Unix。

（4）可靠性：经过长时间测试，SinoSonic系统性能稳定可靠。

（5）灵活性：根据用户特定需求，SinoSonic系统的输入、输出特性和用户接口极易修改。

（6）拓展性：随着应用领域不断扩展，用户需求不断提高，SinoSonic也可不断更新拓展。

三、语音合成系统的API设计

语音合成系统的API，可以考虑不同层次的开发需要。目前国际上较为流行的方法是面向用户应提供不同层次的用户接口，即High-Level API或Low-Level API。API分层设计的核心思想，是提供语音合成系统以不同层次的开发需要。

High—Level Speech API的目的是使用户不需要进行太多的学习，便能够迅速、简便地使用语音合成系统的大部分功能。API简洁、明了、功能全面，且在不

同的应用平台保持一致性，适用于一般意义上的语音合成系统应用再开发。其提供的基本功能应包括：

（1）系统初始化；

（2）系统卸载；

（3）直接将文字转换为语音，并用声卡或其他声音播放卡将声音播放出；

（4）提供播放、暂停和停止等基本播放功能；

（5）修改语速、基频和能量的功能；

（6）韵律控制符的分析和应用；

（7）可视化功能接口。

Low-Level Speech API的目的是使用户能够进行全面、深入的底层开发，其API接口复杂，功能小而细、复杂、规模大，可按不同功能集进行分类，且系统的几个不同的组成模块（如文本分析、韵律、声学处理）均可以提供单独的接口，能全面满足语音合成系统现在和将来应用开发的需要。其提供的基本功能应包括：

（1）系统各个子模块的初始化；

（2）系统各个子模块的卸载；

（3）文字分词、转拼音或词性标注功能；

（4）用户词典维护接口；

（5）合成语音特色（包括男、女声等）；

（6）韵律控制符的分析和应用；

（7）语速、基频和能量的控制功能；

（8）声音播放卡的控制功能；

（9）语音合成的流控制功能、内存管理功能及消息管理功能；

（10）用户自定义文本分析、韵律及合成算法引擎的接口（合成平台开放性）；

（11）不同应用平台的特殊接口；

（12）不同语言的特殊接口；

（13）可视化接口；

（17）声音同步接口；

（15）出错信息解释接口。

详细基本功能集的定义可根据各单位自己的系统的情况而定，也可以制定统一的标准。接口的设计，还应考虑语音合成产品除了在提供自身发音性能的同时，正向着网络化、多语种、多合成引擎的方向发展。同时，接口还应该可虑方便用户自定义发音风格、系统可训练的实际应用需要。

四、新华音霸

新华音霸是清华大学、炎黄新星和新华世纪联合推出的PC屏幕阅读软件。它可以朗读计算机屏幕中任意出现的文字，增加了人机交互的友好性，同时它还采用了清华大学最新研制的虚拟头像技术，配合声音进行同步播放，极大地提高了软件的趣味性。

五、语音网关

运用语音合成技术，而构筑的语音网关，在很大程度上改变了传统IVR运作模式，为电信网统一消息平台、呼叫中心（CallCenter）注入了全新的活力。它可以为用户实时提供，诸如E-mail、新闻、信息查询等信息，并为用户用清晰自然的语音朗读出来。目前，清华大学和炎黄新是共同推出的语音网关技术，在国内具有相当的优势，并在移动梦网、168平台改造等重大项目中，得到了非常成功的应用。

六、总结

目前就语音合成系统的系统构架来说，它正朝着多语种、网络化和分布式运算的方向发展，其关键的技术牵涉的领域也越来越多。目前，国际上许多大的公司和科研机构，如Motorola、Lucent、IBM等均参与了一种新的XML的一个扩展子集VoiceXML的制定。VoiceXML的出现，将会极大地改变人机交互的通信模式。在分布式运算结构中，将会要求系统的设计更为模块化，并且对模块之间的并行和协调工作提出了更高的要求。现有的语音合成系统研究水平，从一定程度上使系统走向了产品化，其音质和发音效果也被普通人所接受。然而，从另一个角度来说，人的发音各有特色，发音的习惯也不尽相同。能完全像真人一样体现人的说话语气、概念，能体现不同的情感，并能模拟不同人发音特色的语音合成系统的出现，还需要我们投入更大的精力去开拓。下一代的语音合成系统将不再称为“文字到语音转换系统”，而是会被称做“概念到语音转换系统（CTS系统）”。