计算机科学1993 Vol.20№.5

计算机声音处理技术

胡上序　钱　涛

（浙江大学计算与信息中心　杭州310027）

摘　要

Dealing with voice in computer is becoming an active and important field of computer science.This paper summarizes various kinds of available technologies of handling voice in computer and shows the latest development of this field.Moreover,the speech recognition technology is discussed in details.

一、引言

计算机从处理单一的文字和数据发展到能够处理声音是一个质变。早先计算机通讯多数是采用发送电子邮件或通过网络系统共享数据和文本文件。仅仅限于这种方法是远远不够的。如同人类进行交流的方式除了书写成文的东西，还有语言。计算机声音处理因此也是计算机研究领域的一个重要方向。该技术涉及到诸多其它领域的技术，如电子学、通讯技术和控制学。目前这方面的研究取得了较大的进展，技术也日趋成熟。

二、计算机声音处理技术

计算机声音处理技术可以分为三大类：信号内容处理、连接控制和软件构造。

2.1 信号内容处理

信号内容处理是指创建、操纵和分析音频电路中的信息。这种技术主要包括音频信号数字化、数字信号处理、文字-语言的合成和语言识别等技术。其中，音频信号的数字化编码技术是目前较为成熟的技术。

2.1.1 音频信号数字化　将模拟音频信号数字化是现代化电话系统中的一个最显著的特征。大多数现代化交换机设备都是将音频电路信号转化为数字串流的。

标准电话系统使用3字节脉冲编码调制（PCM），并产生传播速率为64k/s的数据

胡上序　教授，博士导师，主要从事语言识别，计算机神经网络，计算机仿真与智能信息系统等研究。钱涛　博士生，主要从事多媒体技术，计算机语言识别等方面的理论研究和实际系统开发。

参考文献

〔1〕 Bian Shili, Yao Tian Shun, A system of generating target language from interlingua, 5th SSICCA., 1992

〔2〕 R. Simons and J. Slocum, Generating English Discourse from Semantic Networks, Communication of the ACM Vol 15 No 10

〔3〕 左孝凌，刘永才等《离散数学》上海科学技术出版社

〔4〕 殷钟崃等《英语语法理论及流派》四川大学出版社

〔5〕 姚天顺、马黎环，Generating English Text from Chinese Semantic Representation,Proceedings of 1988. International Conference on Computer Processing of Chinese and Oriental Languages Montreal Carad 1988

流。这种方法可以很精确地表示出现在标准电话电路中的脉冲信号。更先进的编码技术不仅可以在保证相同声音质量的前提下降低传播速率，也可以在相同的传播速率下提高声音质量。

就目前技术而言，编码化的信号波形的传播速率不应低于16k/s，否则声音的质量就会下降。当速率低于5k/s时，则音频信号毫无用处。在这一领域，参数化编码技术是一种很有用的技术，它是将数字参数编码而不是将音频信号波形编码。

声音质量与其所需的存储空间是一对矛盾。使用传播速率为16k/s的方法进行编码的声音信息占用的存储空间是使用64k/s方法编码的四分之一。这种优势是以增加用于处理编码和解码的计算机资源的开销为代价的。但是较低的传播速率将导致声音质量的下降。

2.1.2 **数字信号处理** 在大多数情况下，使用数字化技术要比模拟信号技术可以更快更精确地获得信号处理功能。并且我们无需改变或调整滤波器硬件构成就可以改变滤波器的操作参数。

使用数字信号处理技术有很多益处。最主要的好处是不必增加额外的硬件投资，只要简单增加新的数字转换软件就可以增加设备的信号处理能力。目前一些普通的计算机也开始拥有一些基本的信号处理能力，如Next计算机。

2.1.3 **文字-语言合成** 文字-语言合成是一种能将文字流转换成可理解的人类语言的技术。早期在计算机系统中运用的语音技术是事先录音的单词、音节或音素组合连接起来从而产生语句。这并不是真正的语言合成，而是对已存在的语句的重新排列。这种方法对局限于较小词汇范围的情况是非常有效的。但该方法未能解决声音的音调变化问题。随着音调变化，相同的语段必须重复录音并且应用系统还须根据内容的不同选择适当的语调。

2.1.4 **语言识别** 语言识别是指根据已有词汇集识别一段话语的能力。在有些场合，用户会发现用传统的键盘与计算机系统交互是不太方便甚至是不可能的。这时就需要运用到语言识别技术。迄今，语言识别在精确度和词汇量方面还存在较明显的缺陷。

在电话系统中运用语言识别技术特别有实用价值。如果在计算机应用系统中安装了语言识别和合成装置，那么它就可直接利用电话作为输入/输出设备。

现代语言识别技术的研究始于五十年代。主要运用模拟信号-数字信号转换器和声谱图，计算机从一段话语中摘取语言特性来识别不同的单词。六十年代，在新的模式匹配和语言分类法方面的研究取得了进展。到了七十年代，出现了一系列主要的语言识别技术并推出了实用语言识别系统（表1）。

有五个重要因素可以用来控制和简化语言识别：

①独立单词。所谓独立单词是指该单词与相邻单词之间有短暂而明显的停顿。由独立单词组成的一段语言的识别要比连续话语来得容易。在发音连续的话语中，难以识别单词的边界。同时，单词发音常常会有所改变，如“want”单词在与“to”放在一起读时将丢掉“t”音节。

②说话人数。因为语言的参数表示对特定的说话人语言特征是敏感的，所以对单个说话人的语言识别要比对混合了多个说话人的话语来得容易。在计算机中，不同人的语音模板是不同的。目前运用模式匹配技术进行语言识别的系统还限于特定的一些说话人。也就是说，语言识别的对象的语音模板必须事先存储在计算机中。

③词汇量。词汇量的大小也是影响语言识别正确率的一个重要因素。不失一般性，较大的词汇量意味着包含含糊单词的可能性也增大了。含糊词汇是指它们的语言模式是相似的，使识别器难以区分。

④环境。背景噪音、音量的大小和麦克

风的稳定性等会影响语言识别的正确率。大多数识别系统只有在环境条件可控和安静的情况下，才能保持较好的识别率。

⑤语法。

名　称	特　征	词汇量(单词)	正确率
Dragon Voice Scribe 400	说话人相关 可识别独立单词	4,000个	>95
Dragon Dictate	说话人相关 可识别独立单词	30,000个	>90
IITAR ,25J IC	说话人相关 可识别连续话语	2,000个	98
Phonetic Engine (语言系统公司)	说话人无关 可识别连续话语	10,000 -40,000个	90
Telereo (声音控制系统)	说话人无关 可识别连续话语	30 -10,000个	>98
Voice Card (Votan)	说话人无关有关 可识别连续话语	300个	99(有) 96(无)
Voice Comm Unit(富士)	说话人有关,可识别相连接的单词	4,000个	99.9

表1　现有的实用语言识别系统

2.2 连接控制

连接控制是指控制音频电路以实现用户和声音设备的互连。

声音处理设备比较复杂和昂贵，通常由多个系统和用户共享。因此，应用系统必须控制不同用户如何与声音设备相互连接。这一部件称为开关设备。在大多数情况下是指电话交换机。

2.2.1 电话信号发送　早些时候，电话用户是通过发送事先准备好的模拟信号来请求电话连接的。这是指由拨盘机制产生卡嗒声或由数字按键系统产生系列双音频信号。同时接受方电话铃响或忙音信号也会反馈给呼叫方。上述技术被称为同频带信号传输技术。该技术较为复杂但能传递的信息量较大。使用较为广泛的有内部拨号和拨号数字信息服务。在这两种系统中呼叫信息都是通过一组普通的电路传给任意用户。

先进的网络服务器不仅在电话网络内部需要更多的信号命令和排错功能，而且在网络系统与电话用户之间也一样。为了达到该目标，采用了一种称为不同频道信号传输技术。该技术使用单独电路来发送数据。目前主要有"信号系统#7.SS#7协议"和国际电报电话咨询委员会（CCITT）推荐的"协议Q.931和Q.932"。

2.2.2 第一方和第三方信号发送　所谓第一方信号发送指在同一根电话线上完成服务请求和信号的发送。目前大多数公共电话网络系统采用该种方式。在这种方式下，用户发送呼叫信号之后在同一设备上得到响应。

第三方发送信号是指在不同的电话线或设备上完成呼叫和响应。

2.2.3 点到点命令连接(图1)　通常实现信号发送的方法是由交换机产生一个特殊的点到点命令来连接交换机和计算机设备。这种连接允许两个环境交换请求和状态信息并提供使声音和计算机功能协调的途径。目前主要有贝尔通讯研究所的简便消息桌面接口软件(SMDI)和PBXs、Northern Telecom的ISDN应用协议、AT&T的添加/转换应用接口和Mitel的Host-Command Interface。

2.3 软件构造

与声音应用系统相关的软件构造是声音源抽象模型和声音源分布式访问的软件构造。这里声音源包括前面提到的内容处理和连接控制技术。此处重点介绍声音源的抽象

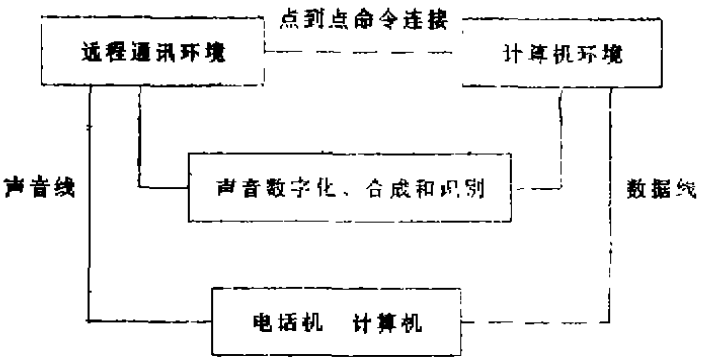

图1　点到点命令连接

模型。

声音源在应用程序员的显示方式应该独立于声音源在程序中的实现细节。这就是声音源模型化的概念。

程序员可以更多地运用这个模型和系统软件，而不必过多地考虑如何将逻辑声音源转化为物理源。所有声音源应由操作系统自动模型化。在某种意义上，程序员面对的是虚拟的声源而非实际的声音。声音源只有通过抽象，其软件才能在更为广泛范围的硬件系统上使用(图2)。

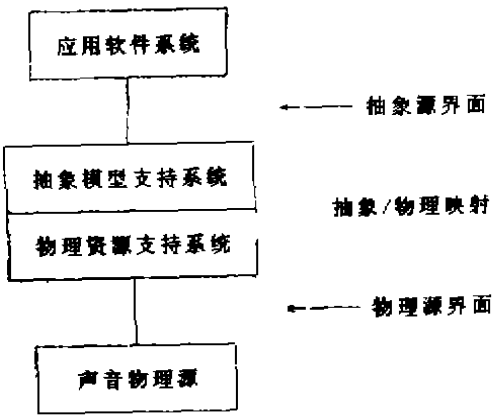

图 2 资源抽象

这种抽象对应用程序员来讲也是非常必要的。在某种意义上，软件上的投资要比在声音设备上的投资大。因此，也希望系统软件无需进行额外的修改就可以与新的声音设备匹配。换句话说，软件操作的对象应该是抽象的声音源而非物理源；在运动的时候，系统又能将抽象资源赋于各类等价的物理资源。

不少计算机公司在声音源抽象模型研究上投入了很大的精力，并取得了不少成果。如王安公司的"Speech and Telephony Environment for Programmers"、数字设备公司的"Computer Integrated Telephony"和IBM公司的"Telephony Applications Services Callpath"。

三、结束语

计算机声音处理技术已有长足的发展，但很多方面仍有待于进一步深入的探讨和研究。如如何使计算机语言识别系统识别更多的词汇、如何进一步提高识别的正确率。

本文没有涉及声音数据压缩技术，这并不是说其不重要，而事实上恰恰相反。随着多媒体技术在国内外的迅速发展，声音数据压缩技术日趋重要，其研究也正成为热点。不少该专题的学术论文见诸报端和学术刊物，可供参阅。

参考文献

[1] Ragui Kamel, Bell-Northern Research, Voice in Computing, Computer, August 1990

[2] C.Schmandt and M.McKenna, An Audio and Telephone Server for Multi-Media Workstations, Proc.Second IEEE Conf.Computer Workstations, CS Press, Los Alamitos, Calif., Order No.810, 1988

[3] Richard D.Peacocke and Daryl H.Graf, An Introduction to Speech and Speaker Recognition, Computer, August 1990

[4] J.Mariani, Recent Advances in Speech Processing, Proc.IEEE Intl.Conf.Acoustics, Speech, and Signal Processing, Glasgow, Scotland, May 1989

[5] V.N.Gupta, M.Lenning, and P.Mermelstein, Decision Rules for Speaker-Independent Isolated Word Recognition, Proc.1984 IEEE Intl. Cont. Acoustics, Speech, and Signal Processing, IEEE, 1984

[6] A.W.Biermann et al., Natural Language with Discrete Speech as a Mode for Human-to-Machine Communication, Comm.ACM, Vol.28, No.6, 1985