第 22 卷 第 1 期
2003 年 2 月

情 报 学 报

Vol.22, No 1
February, 2003

利用语音识别进行信息检索[1)]

陈海英 于金辉
(浙江大学图书馆,杭州 310027)
(浙江大学 CAD&CG 国家重点实验室,杭州 310027)

摘要 本文先简要介绍语音识别技术的主要思想,然后着重讨论利用该技术进行网上文字、图像、语音和音频信息检索并针对目前受语音识别技术水平所限而带来的问题提出解决方法。

关键词 语音识别 文字 图像 音频 检索

Voice-enabled Information Retrieval

Chen Haiying
(*Library of Zhejing University, Hangzhou* 310027)
Yu Jinhui
(*State Key Lab. of CAD&CG, Zhejiang University, Hangzhou* 310027)

Abstract This paper introduces briefly the main ideas of speech recognition, then discusses how to use them for information retrieval and proposes solutions to the problems associated with current technology of speech recognition.

Keywords speech recognition, word, image, audio, information retrieval.

随着 Internet 的迅速发展,网上的文字、图像以及音频信息日益增多。目前网上检索主要通过用户向计算机键入文字进行,例如我们已经非常熟悉的 Yahoo! 和 Google 这样的搜索引擎。语音识别是机器通过识别和理解过程把语音信号转变为相应的文本文件或命令的高技术,经过 40 多年的发展,它已经显示出巨大的应用前景,高性能的语音识别系统相继问世。如果我们能实现通过语音在网上检索,就可以大大减轻用户手和眼的负担,甚至盲人也能够像正常人一样跟上信息时代的步伐在网上自由浏览、检索有关信息,并通过语音合成技术把检索到的文字转化成语音供学习之用。本文先简要介绍语音识别技术的主要思想,然后着重讨论利用这些技术进行网上文字、图像、语音和音频信息检索并针对目前受语音识别技术水平所限而带来的问题提出解决方法。

1 语音识别系统

一个典型的语音识别系统如图 1 所示。

其中,预处理包括语音信号采样,反混叠带通滤波、去除个体发音差异和设备、环境引起的噪声影响等,并涉及到语音识别基元的选取和端点监测问题;特征提取部分用于提取语音中反映本质特征的声学参数,如平均能量、平均跨零率、共振峰等;训练在识别前进行,通过让讲话者说出一些句子,有时需多次重复某些语音,从原始样本中去除冗余信息,保留关键数据,再按照一定规则对数据加以聚类,形成语音模式库;模式匹配部分是整个语音识别系统的核心,它是根据一定的准则(如某种距离测度)以及专家知

收稿日期:2002 年 1 月 18 日
作者简介:陈海英,女,馆员。于金辉,男,研究员。
1) 国家自然科学基金资助项目(60073024)

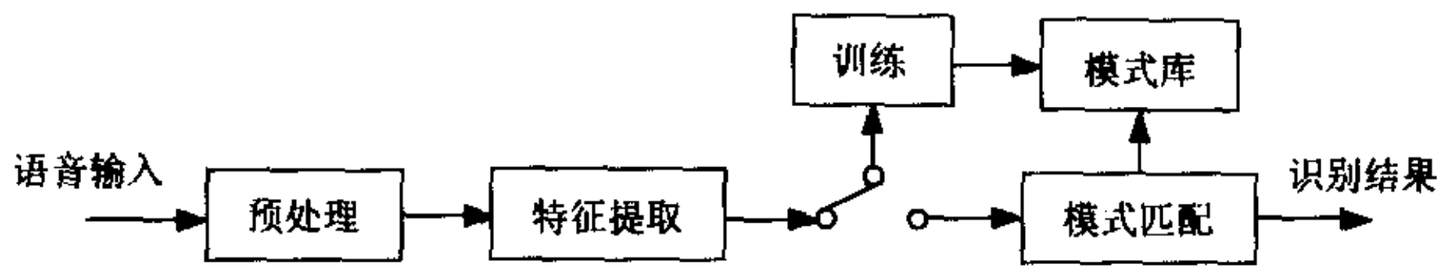

图 1

识(如构词规则、语法规则、语义规则等),计算输入特征与库存模式之间的相似度(如距离匹配、似然概率),判断出输入语音的语义信息。

目前具有代表性的语音识别方法主要有特征参数匹配法、隐马尔科夫法和神经网络法。

1.1　特征参数匹配法

特征参数匹配法是一种传统的模式识别方法,其要点是:在训练过程中从训练语句中提取出特征参数,这些特征参数代表了语音的本质,称为相应语音的模板,然后在识别过程中从待识别语音信号中按同样的处理方法提取出语音参数,最后应用某种不变的测度(如距离测度)寻求语音参数与模板参数之间的相似性,用似然函数进行判决。特征参数匹配法在中小词汇识别方面的运用很成功。

1.2　隐马尔科夫法

隐马尔科夫法(HMM)的出现使得自然语音识别系统取得了实质性的突破。HMM 方法现在已经成为语音识别的主流技术,目前大多数大词汇量、连续语音的非特定人语音识别系统都是基于 HMM 模型。

HMM 是对语音信号的时间序列结构建立统计模型,将之看作一个数学上的双重随机过程:一个是用具有有限状态的 Markov 链来模拟语音信号统计特性变化的隐含的随机过程,另一个是与 Markov 链的每一相关联的观测序列的随机过程。HMM 语音模型 $\lambda(\pi,A,B)$由起始状态概率(π)、状态转移概率(A)和观测序列概率(B)三个参数决定。π 揭示了 HMM 的拓扑结构,A 描述了语音信号随时间的变化情况,B 给出了观测序列的统计特性。

HMM 语音识别系统的一般过程是:先用 Baum Welch 算法,通过迭代使观测序列与模型吻合的概率 $P(\lambda|O)$(O 指当前样本的观测序列)达到某极限,训练出信号最佳 HMM 模型 $\lambda(\pi,A,B)$;在识别过程中采用基于整体约束最佳准则的 Viterbi 算法,计算当前语音序列和模型的似然概率 $P(\lambda|O)$(O 指当前语音信号序列),选出最佳状态序列,确定输出结果。

1.3　神经网络法

人工神经网络(ANN)本质上是一个自适应非线性动态系统,它模拟了人类神经元活动的原理,具有自适应性、并行性、鲁棒性、容错性和学习特性。目前语音识别神经网络主要有多层感知器网、Kohonen 自组织神经网和预测神经网络。

多层感知器误差反转网络(BP 网)是采用反向传播算法的多层感知器神经网络,它克服了 HMM 对声学上相似的词易于混淆的缺点,已成功地用于音素识别。人们一般趋向于用 BP 网完成静态模式分类,再用 HMM 或动态规划(DP)完成时间对准。

Kohonen 自组织神经网是基于这样一个生理学研究成果:声音或景物对听觉或视觉器官的刺激沿神经通路向大脑皮层投射时会保持一种拓扑结构,从而在大脑皮层形成各种特征区域,Kohonen 提出的自组织特征图(Self Organizing Feature Map)成功地模拟了这一生理特征并对荷兰语的 21 个音素实现了声控打字机。

预测神经网络(PNN)把感知器作为预测器而不是模式分类器来使用,它具有很强的建模能力,可用于大词汇量、连续语音、非特定人的语音识别研究上。它的特点可以概括为充分利用语音模式中的时间相关性作为识别的线索;用 DP 法对语音信号进行时间规正;基于动态规划和反向传播能够找到一种优化算法;容易增加新的识别类等。

基于神经网络的语音识别系统具有很大的发展潜力,但普遍存在训练、识别时间太长的缺点,目前仍处于试验探索阶段。

2　语音检索系统

在 40 余年语音识别研究基础上已经有若干汉语语音识别系统出现,如汉王听写输入系统、天王话音、蒙恬听写王以及 IBM 公司的 Via Voice 嵌入式语音输入系统等。这些系统为我们设计语音检索器和

语音浏览器提供了基本技术基础。

语音检索器的工作方式是通过语音来对文字、图像、语音和音频进行检索,下面我们对它们逐一进行详细讨论。

2.1 文字检索

由于文本是语音的一种脚本形式,利用自动语音识别技术可以把语音转换成文本,从而采用文本检索方法进行检索。

2.2 图像检索

对于图像我们可以在它的文本注解中嵌入图像的内容信息,然后通过语音识别技术把语音转换成文本来对图像进行检索。考虑到在图像内容注解中有主观性和不精确性等缺点,近年来提出基于内容的图像检索,其思路是通过图像自身的视觉内容如颜色、纹理、形状、颜色布局等特征进行计算检索。

其中颜色特征有颜色直方图、颜色矩(Color Moments),颜色集(Color Sets);纹理特征有纹理特征共生矩阵以及视觉纹理特征的近似计算,如粗糙度(coarseness)、对比度(contrast)、方向度(directionality)、线象度(linelikenes)、规整度(regularity)和粗略度(roughness);形状特征有傅立叶描述和矩不变量,Chamfer 匹配,Turning 函数以及小波算子;颜色布局即颜色特征和空间关系。

有了上述图像内容检索特征之后,我们便可以利用自动语音识别技术把相应的语音检索命令转换成文本,然后系统调用对应的检索特征计算模块对图像进行内容检索。

2.3 语音检索

采用语音识别技术还可以以语音为中心进行检索,如电台节目、电话交谈、会议录音等。由于目前连续语音识别系统在实际应用如电话和新闻广播中的识别率并不高,但考虑到检索任务只是匹配包含在音频数据中的查询词句,我们可以把语音对话轨迹转换成文本脚本,然后组织成适合全文检索的形式支持检索。

当语音识别系统处理各方面无限制主题的大范围语音资料时,识别性能会变差,尤其当一些专业词汇(如人名、地点)不在系统词库中时。一种变通的方法是利用子词(sub-word)索引单元,当执行查询时,用户的查询先被分解为子词单元,然后将这些单元的特征与库中预先计算好的特征进行匹配。

在无约束的语音中自动检测词或短语可利用关键词的发现(Spotting)技术来识别,或标记出长段录音或音轨中反映用户感兴趣的事件,这些标记就可以用于检索。如通过捕捉体育比赛解说词中"进球"的词语可以标记进球的内容。

2.4 音频检索

音频检索是以波形为对象的检索,这里的音频可以是汽车发动机声、雨声、鸟叫声,也可以是人的哭笑声和音乐等。我们可以把这些声音的样本数据用文本注解后存放在系统中,对每个进入数据库中的声音,先计算其 N 维声学矢量特征,然后计算这些样本的平均矢量和协方差矩阵得出表达某类声音的类模型。在检索时先把语音命令转换成文本,通过文本选出声音样本,然后把网上的候选声音数据的类模型计算出来,与样本的类模型比较得出判断结果。对于音乐一类的音频信号除了文本注释之外,我们可以进一步利用音乐的时间和频率特征如节拍和基本频率来进行检索。

3 存在的问题与解决方法

由于语音信号的变异性很大,在世界上找不到两个人对同一个单词的发音完全相同,所以任何语音识别系统必须解决对不同用户口音的适应问题。目前一般在系统中加入训练模块(如图 1 示)来解决这个问题,这个方法对拥有固定计算机的用户是可行有效的。但对公共终端来讲其用户是变化的,某人今天可以到甲终端,明天可以到乙终端,在一天之内甲终端也可能前后有多个用户上机,若每个用户都在现场在线进行口音适应训练则要占据相当的时间以及内存空间。以 IBM 的 Via Voice 为例,它需要用户朗读 50 ~ 260 个句子进行训练,如果用户的普通话说得很标准,那么录制所要求的前 50 个句子进行训练就差不多了,如果用户的口音较重则需要把 260 个句子都读完。此外在语音识别过程还有可能出现错误,比如语音输入为"入世",但系统却识别成"入室",用户还需与系统交互进行修改。无疑,众多用户在公共终端现场在线对系统进行语音训练将把语音检索带来的效率抵消掉。

解决这个问题的方法是把在线语音训练所用的时间转移到离线,即用户事先进行语音训练并把得到的语音模式存放在网中。在使用语音检索系统时把自己的语音模式调入系统中来,使用完毕系统可

询问用户是否需要保留语音模式供未来几天使用，若需要，则保留，否则系统自动删除该语音模式。对于那些保留过的语音模式存入系统中之后一段时间没有再被使用过，系统可以设置一时间门限（比如一天或一周），超过门限的语音模式则被系统自动删除掉。若某用户从未在网上建立过自己的语音模式，那么只有现场在线花时间建立自己的语音模式之后才能进行语音检索。

由于语音识别系统种类繁多，若在网上调用自己预先建立的语音模式到不同的语音识别系统中涉及到一系列的标准问题，如特征提取、声学模型、语言模型等等。由于它们牵涉到复杂的系统结构与运算方法，在目前统一这些标准是极为困难的事情。从经济和使用角度考虑最好全国采用一种系统，但这样可能形成厂家日后对该市场的垄断。一个现实的办法是国家有关部门对公共终端语音检索制订出全国统一的规划，筛选出少数较好语音识别系统作为候选进行跟踪，在适当的时候采用 2～3 种系统，这样可以激励不同厂家的竞争从而促使系统性能不断改善。在设计网上用语音识别系统平台时则要充分考虑系统的扩展性，使平台有多种系统的接口并允许将来性能逐步提高的新版语音识别系统在现有平台上使用。

4 结　束　语

语音识别是一个专门的研究领域，它涉及到声学、语音学、语言学、人工智能、数字信号处理理论、信息理论、模式识别理论、最优化理论、计算机科学等学科。目前在语音识别研究领域非常活跃的课题为稳健语音识别、说话者自适应技术、大词汇量关键词识别算法、语音识别的可信度测评算法、基于类的语言模型和自适应语言模型等。其中说话者自适应技术的研究已经取得了相当大的进步，稳健语音识别的算法还未取得根本性突破，但其研究意义非常重大。语言模型也是目前研究的一个重要方面。我们相信，随着语音识别技术的不断发展，网络语音检索智能接口必将得到越来越广泛的应用。

参　考　文　献

1　李晓霞，王东木，李雪耀．语音识别技术述评．计算机应用研究，1999，10

2　刘加．汉语大词汇量语音识别系统研究进展．电子学报，2000，28(1)

3　郑方，牟晓隆，徐明星，武建，宋战江．汉语听写机技术的研究与实现．软件学报，1999，10(4)

4　李国辉．基于内容的音频检索．计算机世界，1999，20

（责任编辑　芮国章）