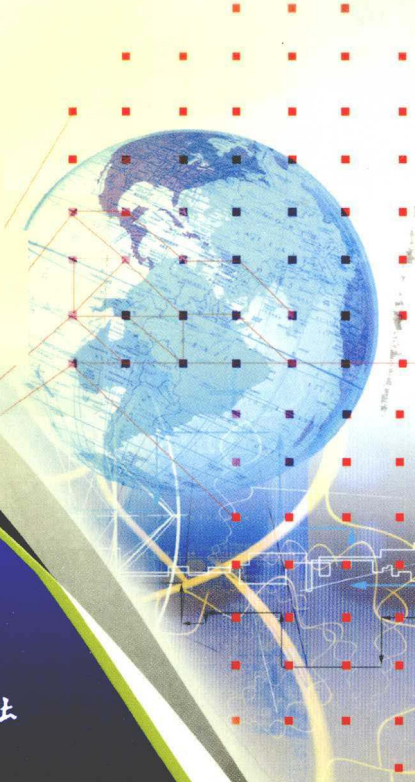
高维信息几何与语音分析

曹文明　郑能恒　冯　浩　著



科学出版社

高维信息几何与语音分析

曹文明　郑能恒　冯　浩　著

科学出版社
北京

前　言

自然语音是人与人之间以及人与计算机之间最方便的一种信息交换方式。因此，使计算机具有类似于人的听觉功能和发音功能，是人们长期追求的目标。计算机语音识别就是为着这个目标而提出的。它是智能计算机系统的重要特征。这一技术的应用将从根本上改变计算机的人机界面，从而对各行各业的发展以及人们日常生活的方方面面都产生深远的影响。目前，语音分析方法都是假定分类信息是完全包含在训练样本内，以多类不同样本的最优划分为基础，分类器的训练过程实际上可以看作对样本的划分过程。这些方法从非同类被识别对象的“差别”出发，在实现算法时，都是侧重于不同事物的“区别”，即一类样本与有限类已知样本的区分。这与人类对事物的认知方式存在很大差异：人在认识事物时是一类一类地认识，重视同类事物之间的联系，或者说，注重一类样本同无限类未知样本的区分。以“区别”为出发点的传统模式识别，必然导致以下两个局限：一是对于首次遇到的未学习过的新事物，容易误认为是某一类已学习过的旧事物；二是在对于未学习过的新事物进行新的学习时，往往会打乱旧的知识，即破坏对原已学习过的对旧事物的识别。这正是传统的模式识别理论在实际应用中难以取得真正理想效果的原因所在。王守觉院士从人类认知事物的角度出发，重新研究了神经网络模式识别问题，创新性地提出了以多维流形的拓扑学理论为基础的强调“认识”的模式识别——仿生模式识别；并以语音分析实用为目标，发展了一种对神经网络行为的高维空间几何分析方法，同时还提出了实现认识事物为目标的高维空间非超球面复杂几何形体覆盖进行模式识别的原理。因此，本书目的是根据语音分析中的高维信息几何能力，研究高维信息几何的仿生信息学的理论与应用。

本书共分 7 章，包含以下内容：

第 1 章讨论了要改变当前语音识别技术停步不前的状况，就必须从根本问题入手，对于语音识别而言就是要从语音模型的建立入手。也就是说，当前的语音识别技术需要迫切解决的两个首要问题是：

① 建立一个稳健的声学模型，这是语音识别系统性能优劣的决定因素。

② 建立一种对被识别目标的快速搜索策略，并且要求这种策略可以准确剔除不在识别范围内的干扰信息。这不但是决定系统识别速度的重要因素，同时也是增加系统稳健性的重要方法——高维信息几何覆盖方法。

第 2，3 章首先简要介绍了语音识别的相关知识——语音识别的分类和语音

识别的基本框架；然后就语音信号的短时特性和窗函数的选取等方面进行了讨论，这是语音可以进行短时分帧处理的前提，也是一段连续语音可以映射成高维空间中近似连续的许多点的前提；接着介绍了几种基于语音短时特性的特征提取方法；最后介绍了现行的一些主要的识别算法。

第 4 章研究了高维信息几何的欧氏空间理论，并分析了高维信息几何的欧氏几何的度量关系，给出了高维信息几何的欧氏空间的复杂几何体神经元应用，研究了基于高维信息几何的同调连续原理的相关概念和性质，根据同调连续原理构建空间子流形，并由拓扑流形的概念和性质并结合仿生模式识别给出了依据流形学习的模式识别方法。

第 5 章中高维空间仿生信息几何学的理论和计算过程都是从高维空间中几何图形概念出发，用空间中的点集来描述子空间，借助于平面几何简单运算组合迭代实现高维空间中复杂的计算，并有了一些成功的范例，但缺乏几何学中公理化体系的优点。本章利用线性代数工具对高维空间几何作系统描述，并形成一个自洽的公理体系，并分析了高维空间中几何图形覆盖性质以及在语音分析中应用。

第 6 章提出了一种基于高维空间点覆盖动态搜索理论的非特定人连续数字语音识别的新算法，这种算法可以不经过端点检测和分割。本章先根据实际连续数字语音的各个不同数字音节，构建连续语音中各个不同数字音节的特征空间覆盖区。在识别时，利用高维空间点覆盖动态搜索理论进行识别，得到了较为满意的识别结果。

第 7 章在语音情感特征的基础上，提出了基于多权值神经网络的语音情感识别，通过实验证明了多权值神经网络在语音情感识别上的可行性，并且通过与 SVM 模型的比较实验证明了多权值神经网络的优越性。

本书反映了高维信息几何的国内外最新动态，书中许多内容都是作者及团队的最新研究成果，其中部分研究结果尚未正式发表。因此，借本书出版之际，要特别感谢中国科学院半导体研究所王守觉院士，同时还要感谢深圳大学的谢维信，衢州学院的张有正，浙江工业大学的丁立军、庄德文等老师，以及何天成、潘晓霞、王建华等同学。

本书得到了国家自然科学基金（No：60901061，60871093，60576055，60872126）、武装预研项目基金（No：9140C8000208C80）、深圳大学自然科学基金的资助，在此一并表示感谢。

书中疏漏和不足之处在所难免，敬请广大读者和专家学者批评指正。

曹文明　郑能恒　冯　浩

2011 年 1 月

目　　录

第1章 绪 论

语音识别(speech recognition)是指采用电子计算机、电子电路等器件自动提取或决定声波信号中最基本的、最有意义的信息(韵质信息)[1],以此来确定语音信号中语言含义的过程。作为一门学科,它与声学、语言学、生理学、心理学、人工智能、数字信号处理理论、模式识别理论以及计算机科学等众多学科紧密相联[2]。

随着科学技术的发展,人机接口、人机交互界面变得越来越重要,语音接口由于其自然、机动(适应性强)、高效性而受到高度的重视[3]。经过近半个世纪的努力,它正逐渐成为下一代操作系统和应用程序用户界面的关键技术。

1.1 语音识别研究的重要意义

自然语音是人与人之间以及人与计算机之间最方便的一种信息交换方式。因此,使计算机具有类似于人的听觉功能和发音功能,是人们长期追求的目标[4]。计算机语音识别就是为了实现这个目标而被提出的。它是智能计算机系统的重要特征[5]。这一技术的应用将从根本上改变计算机的人机交互界面,从而对各行各业的发展以及人们日常生活的方方面面都产生深远的影响。

1. 基于计算机交互界面的应用

有了语音识别技术,用户可以在桌面上用声音命令、控制或操纵计算机,让它成为其最听话的助手。用户可以口述文本并将它转换成为支持剪贴功能的任何应用程序(cut-and-paste function)。用户可以口述文件、报告和邮件,让计算机成为一名出众的打字员。忙时用户可以边做工作边将任务分配给他的计算机伙伴;闲时用户可以以任何姿势与计算机聊天,不必再担心患上电脑综合征。让计算机真正成为其最忠实的伙伴、最得力的助手,且不需劳累手指,远离了辐射较强的显示屏。

2. 基于电话的应用

这是现阶段语音识别技术中最重要的一个应用方向,口述要拨打的电话号码或呼叫一个亲切的名称,就可以连通电话的另一端。即使在洗衣、做饭走不开时,

也可以用一句轻松的命令接通电话。手机有了这一功能也会增色不少。相信在不久的将来，它必定在手机技术中独领风骚。

另外，通过电话使计算机能够直接为客户提供金融、证券和旅游等方面的信息查询及服务，进而成为电子商务进展中的重要一环(voice-commerce)①，这也是当今语音识别技术的热点。现有的许多对话系统(conversational systems)就是专为这一功能而设的。

① GALAXY[6,7]。GALAXY 是由麻省理工学院的口语语言系统(spoken language system，SLS)小组开发的一个人机对话系统(human-computer conversational system)，用于在线的信息和服务的口语语音接口。即机器要懂得用户的语音问题，然后找到答案，最后同样用自然语音形式反馈给用户。最初它被用于旅游航空等领域，包括空中旅行(air travel)、本地导航(local city navigation)和天气预报(weather information)等，并支持多种语言的查询。

② VOYAGER。VOYAGER 是 GALAXY 的前一代。它们采用的都是 SUMMIT[8~10]的识别引擎(recognition engine)。

③ WAXHOLM[11]。WAXHOLM 系统被用来查询斯德哥尔摩群岛(Stockholm Archipelago)的游船信息。

④ 还有一些列车时刻表的对话查询系统[12~14]。

⑤ 基于对话的语音识别技术在 ATIS(air travel information system)领域[15,16]、城市导航[17,18]、汽车分类(automobile classifieds)[19]等领域都曾有过应用。

3. 在工业控制方面的应用

语音识别技术作为声控产业，必将对编辑排版、办公自动化、工业过程和机器操作的声控技术起到重大的推进作用。

4. 在多语言翻译工程中的应用

多语言翻译工程[20~24]是语音识别、机器翻译(machine translation)和语音合成(speech synthesis)三种技术结合的产物，其中语音识别是首要完成的步骤。这一工程可以实现各国人们使用本国语言进行互相交流的梦想。这些工程目前仍在进行之中。其中，WERBMOBIL[21]的目的是开发一种便携式的语音对语音的同声翻译系统，该系统目前仅支持日语对英语，以及英语和德语之间的互译。JANUS-Ⅱ[22]是由卡耐基梅隆大学开发的，目前支持从英语或德语到德语、英语、日语中的一种翻译。文献[24]给出的 SLT(spoken language translation)系统是由

① 基于语音的电子商务，有些文献也称其为 v-business。

韩国电子通信研究院(ETRI)开发的,它可以将韩语翻译成英语或日语,用于旅游计划等任务。

待这一工程完成之时,面对面谈判时将不再需要同声翻译人员,一对耳机一个话筒,用户就会感觉像在和老朋友交谈一样,自如随意。

5. *在其他方面的应用*

语音识别还在信息检索、情报获取方面有重要应用。可以预言,语音识别技术必将对工业、金融、商业、文化、教育等诸方面产生革命性的影响。这是一项具有巨大应用前景的工程。

1.2 研究背景

语音识别技术从20世纪50年代发展至今可以说已经取得了许多可喜的成就,如IBM的VTD(voice type dictation)[25]系列和Dragon的Dragon Dictate系统[26,27]。然而,在对历史的回顾之中,人们才发现每一步走得都是如此的艰难,才体会到语音识别技术的研究确实是任重而道远,而且目前仍然存在许多急需解决的基础问题。

1.2.1 国外语音识别研究的发展概况

1. *早期(二十世纪五六十年代)*

早在1952年,贝尔实验室的Davis、Biddulph和Balashek研究成功了世界上第一个语音识别系统——可以识别10个英文数字发音的实验系统[28]。该系统识别的是一个人说出的孤立数字,并且很大程度上依赖于每个数字中元音的共振峰的测量。1956年,RCA实验室的Olson和Belar研制出了可以识别一个说话人的10个单音节的系统[29],它同样依赖于元音带谱的测量。1959年,英国的Fry和Denes研制出了一个能够识别4个元音和9个辅音的识别器[30],他们采用了谱分析仪和模式匹配器。不同的是,他们对音素的序列做了限制(这相当于现在的语法规则),以此来增加字识别的准确率。同一时期的还有麻省理工学院林肯实验室的元音识别器[31],它以一种非特定人的方式来识别10个元音,这里同样运用了语音波形的频率谱和谱序列的信息。

20世纪60年代,出现了一些语音识别基本思想的雏形,主要有三项成果。首先是美国RCA实验室的Martin和他的同事提出了一些有关语音的不均匀尺度问题的解决方案,发展了一些基于语音的起点和终点检测的时间规范化的方法[32]。同时,苏联的Vintsyuk提出了利用动态规划的方法将两段语音的时间进

行对齐[33]，这实际上是动态时间弯折(dynamic time warping，DTW)方法的最早版本，20 世纪 80 年代才为外界知晓。第三个是连续语音识别的先驱 Reddy，他采用的是音素的动态跟踪方法[34]。

2. 中期(二十世纪七八十年代)

20 世纪 70 年代，语音识别技术取得了新的突破性进展。

首先，在小词汇量、孤立词的识别方面取得了许多实质性的进展。

其次，IBM 的语音研究小组在 20 世纪 70 年代开始进行大词汇语音识别研究工作。

最后，贝尔实验室也开始了一系列有关非特定人语音识别的实验。这一研究历经 10 年，其成果是建立了用于非特定人语音识别的标准模板的方法。然而，目前的语音识别器对非特定人的正确识别率还不是很高。

这一时期的语音识别方法基本上是采用传统的模式识别策略。其中，以 Velichko和 Zagoruyko[35]、迫江和千叶[36]以及板仓[37]等的研究工作最具有代表性。Velichko 和 Zagoruyko 的研究为模式识别应用于语音识别这一领域奠定了基础；迫江和千叶的研究则展示了如何利用动态规划(dynamic programming)技术在待识语音模式与标准语音模式之间进行非线性时间匹配的方法；而板仓的研究则提出了如何将线性预测分析(LPC)技术加以扩展，使之用于语音信号的特征抽取的方法。

20 世纪 80 年代也是一个硕果累累的年代，这时期的研究重点逐渐开始转向大词汇量、非特定人的连续语音识别上。

首先，连接词识别方面取得了很大的成就。包括 NEC(Nippon Electric Corporation)的 Sakoe 两级动态规划(two-level dynamic programming)法[38]、贝尔实验室的 Myers 和 Rabiner 的分层构造(level building)法[39]、贝尔实验室的 Lee 和 Rabiner 的帧同步分层构造(frame synchronous level building)法[40]，以及英国 JSRU(Joint Speech Research Unit)的 Bridle 和 Brown 的一次通过法[41]。其基本思路都是将连续单词语音识别系统的参考模式看成是由孤立单词的参考模式按时间顺序动态接续组合而成，识别系统将待识别的连续语音和被接续起来的单词模式序列进行匹配比较，找到距离最短的单词参考模式的序列就是识别结果。

其次，20 世纪 80 年代最具代表性的成就就是将语音识别的技术思路从基于模板匹配的方法转移到了基于统计模型的方法，尤其是隐马尔可夫模型(hidden Markov model，HMM)[42,43]的方法。在 20 世纪 80 年代中期，HMM 技术已经得到了全世界的认可，并不断成熟和完善，最终成为语音识别的主流方法。

接着，DARPA 的又一个 10 年计划“(1000 个单词)连续语音数据库管

理”[44]在这一时期开始了。在这一时期，以知识为基础的语音识别的研究日益受到重视。在进行连续语音识别时，除了识别声学信息外，更多的是用各种语言知识[45]，如构词[46]、句法[46,47]、语义[48]、对话背景等方面的知识来帮助进一步对语音作出识别和理解。同时，在语音识别研究领域还产生了基于统计概率的语言模型[49]。

最后，人工神经网络(ANN)[50,51]在语音识别中的应用研究也在这一时期兴起。在这些研究中，大部分采用基于反向传播法(BP算法)的多层感知器网络。ANN具有能够区分复杂分类边界的能力，显然它十分有助于模式划分。由于它克服了HMM的自适应能力弱、稳健性不强等特点，因而在语音识别的研究中也占有很重要的位置。20世纪80年代末，Kohonen利用SOM(self-organizing map)[52]和LVQ(learning vector quantization)[53]成功地设计了一台针对芬兰语的音素打字机[54]。

3. 近期(20世纪90年代至今)

进入20世纪90年代以后，语音识别的研究主要集中在提高非特定人的大词汇量连续语音识别(large vocabulary continuous speech recognition，LVCSR)[55]的性能上，在语音识别的系统框架方面并没有重大突破。在这一时期，尽管真正理想的识别系统距问世还尚待时日，但许多产品已竞相浮出水面，这是语音识别技术初露光芒的一个时期。

首先，基于电话的语音识别系统[6~18]，由于其广泛的应用前景，已成为当前语音识别应用的一个热点，但它们基本上都是基于特定人甚至是单一一个人的小词汇量的识别系统。

其次，面向个人用途的连续语音听写机[25~27]技术也日趋完善。这方面最具代表性的是IBM的VTD[25]和Dragon公司的Dragon Dictate系统[26,27]。这些系统具有说话人自适应能力，用户可以在使用中不断提高正确识别率，但由于训练时间较长以及错误率较多，因此还没有得到广泛的应用。

20世纪90年代后期，许多科学工作者开始展开音频和视频相结合的语音识别技术的研究[56]，这一技术思路的出发点[57]是希望能够将人的口形以及面部肌肉运动方向加入语音识别中，以此增强某些模糊语音的判断能力，从而增加语音识别的正确率。

这一时期的另一贡献就是面对各类不同应用方向的系统，人们开始有了一些客观的、统一的评价标准，其用于评估的语料库也逐渐变得更加完备。如1993年和1994年对于大词汇量连续语音识别系统的评估[58,59]，以及1997年对Broadcast News的测试[60]，这些结果对以后人们对语音识别系统性能的认识以及系统今后的发展方向的确立都提供了很大的帮助。

在长期的实践过程中，人们开始认识到语音识别任务的艰巨性。21 世纪初期，在继续走产品化道路的同时，一些公司重新又开始关注语音识别的基础性研究课题。例如，微软亚洲研究研院①语音组将近 10 年内的研究方向设定为：研究能使语音识别器识别更准确、功能更强大的新的声学模型技术，以及语音技术在信息检索中的运用（包括语音档案搜索和语音查询处理），其重点为快速说话人自适应技术（如特定群组的声学模型的建造）和声学模型训练技术（如信道映像和判别训练）。经过几年的努力，他们已经在许多基础课题上有了一定的进展，如对说话人和口音自适应技术[61]的研究，对语音的声调[62]、情绪[63]等方面的研究，以及对语言模型[64, 65]、识别算法[66]的进一步探讨等。另外，他们还对连续语音音长的可变性[67]和训练策略[68]等基础性问题进行了重新的探索和评价。而另外一些公司，如日本的多家公司，则把目光转移到对小型系统的应用上，如基于电话的查询系统和基于硬件的小词汇量的识别系统的研究。还有一些公司仍在继续按着主流方向艰难地行进着。

同时，由于 HMM 在大词汇量连续语音识别中暴露的种种弊端②，人们开始着手对新方法的探索，并不断地将各种方法结合起来以提高识别系统的性能。例如，将 HMM 方法与 ANN 方法相结合[69,70]，将 HMM 方法与矢量量化（vector quantization，VQ）[71]以及 FVQ（fuzzy vector quantization）[72]等方法相结合。

为了提高识别系统在高噪声环境下的鲁棒性（robustness），人们进行了各种各样的尝试[73~75]，并对各种方法进行了详尽的比较研究[76,77]，希望可以找到最佳方案。可以说到目前为止，仍没有找到在连续语音和抗噪声方面独具优势的方法。

1.2.2 汉语语音识别研究的发展概况

在我国，最早的有关语音识别的研究开始于 1958 年，中国科学院声学研究所（当时属于电子学研究所）利用电子管电路识别了 10 个元音。然而，直至 20 世纪 70 年代后期，当国际上的语音识别技术已经迅速发展起来，在理论上也渐渐成熟之时，我国许多单位才开始投入这项研究工作中去，如中国科学院自动化研究所、清华大学等。进入 20 世纪 80 年代以后，北京交通大学、复旦大学、哈尔滨工业大学、北京邮电大学、四川大学等高等院校也开始加入语音识别研究的行列。

---

① 1998 年 11 月 5 日，微软公司投巨资在北京成立微软中国研究院，并于 2001 年 11 月 1 日将其正式更名为微软亚洲研究院。微软亚洲研究院是微软公司在海外开设的第二家基础科研机构，也是亚洲地区唯一的基础研究机构。

② 这一点将在 1.4.1 节中给出详细的解释。

1986年3月，国家高科技研究发展计划(863计划)启动，语音识别作为智能计算机系统研究的一个重要组成部分而被专门列为研究课题。从此，我国开始了有组织的语音识别技术的研究。

1998年4月，国家863智能计算机专家组对国内大词汇量连续语音识别系统进行了测评。清华大学电子工程系的语音识别系统获得了最好成绩，字正确率为93%，句子正确率为62.5%，这与占国际领先地位的IBM的语音识别系统[78](95%)水平相当。然而需要指出的是，测评所用的语音为863语音库中的连续语音，该语音属于朗读式语音①，发音清晰，信噪比很高，而本书所用的语音库属于自然口语语音，且信噪比较低。

另外，由于我国经济的发展和对外开放的扩大，以及受汉语语音识别广泛的应用前景的吸引，汉语语音识别技术必将成为国际集团公司竞争的热点。目前，已有IBM、苹果、微软、摩托罗拉等公司相继投入财力、物力和人力，专门从事汉语语音识别技术的研究和开发。苹果公司在1995年推出商用的连接词语音识别系统，之后，IBM也推出了汉语连续语音识别系统ViaVoice。这些都是进一步推进我国汉语语音识别技术发展的动力。

目前，虽然小词汇表特定人的孤立词汉语语音识别系统已经走出实验室，但较实用的连续语音识别系统依然很少[5]，或者说理想的连续语音识别系统依然很少。IBM的ViaVoice在实验室条件下虽然有高达95%以上的正确识别率[78]。但其训练时间长、抗噪声能力弱，在实际的环境中出错频繁，更难以忍受的是，一旦在你的录音模式下有了另一个人的声音，那么你以前的录音很可能就白费了，这是因为他采用的是说话人自适应的技术，不同人的声音将会破坏你的语音记录和语言模型，从这一点来看，IBM的ViaVoice系统还不能算是非特定人的语音识别系统。基于以上种种原因，ViaVioce从1997年推出至今，仍未得到广泛的应用。

1.2.3 连续语音识别研究中遇到的挫折

20世纪70年代，由美国国防部远景研究计划局资助的为期10年的DARPA(Defense Advanced Research Projects Agency)研究计划均未能达到预期目标。20世纪80年代，再一次的包括噪声下的语音识别和会话(口语)识别系统的任务也没有圆满完成。20世纪90年代至今，有关“航空旅行信息检索”中的自然语言处理部分的任务还仍在进行当中。

---

① 朗读式语音是指符合语法规则的、流畅的、讲话方式和讲话内容都经过特殊准备的语言；而自然口语语音是指随意的、至少没有在讲话方式上经过特殊准备的语言，也就是人们在日常生活中通常所讲的话，它通常是不流畅的，包含许多随机事件。

1997 年，美国 NIST① 用 ROVER 程序[79]对 BBN 公司、GRAGON 公司、IBM 公司、SRI 公司、卡耐基梅隆大学、俄勒冈州大学、英国的剑桥大学，以及法国的 LIMSI 公司、飞利浦公司的语音识别系统进行了统一的测评。取得最好成绩的是剑桥大学的 HTK 系统，该系统在标准语音、口语语音、电话语音、有音乐声、有背景声、有口音等条件下的平均字错误率为 16.2%，IBM 次之，为 17.9%[80]；而对于广播语音的测评结果，最好的两家 BBN 和CMU-ISL的字错误率分别为 44.9%和 45.1%[79]。

对广播语音的错误识别率如此之高，主要是因为现在的各个语音识别系统的稳健性不好。广播语音属于自然口语语音，说话时比较随意，随机事件也比较多（这不同于一般的朗读式语音）；并且广播语音的机械噪声比较多，信噪比相对较低，当系统的稳健性不够强时，正确识别率自然不会高。

另外，从 2003 年 David 对 NIST 历年的测评结果的总结[81]（见图 1-1）来看：对会话语音（conversational speech）的误识率一直是居高不下的；广播新闻语音（BNews）的误识率也一直在 10%以上。这主要是由于在前期工作中对于朗读式语音在大词汇量和抗噪声的性能上本身就不牢固。由图 1-1 可以看出，1994 年以前对大词汇量（20k，见图 1-1 中黑色实心方块）的误识率一直在 10%以上；1994 年以后虽然可以下降到 10%以下，但在噪声环境下（仅仅是换了个麦克风）错误率则呈直线上升状态，由 5.1%上升到15.1%。由此可见，现行语音识别系统的抗噪声能力是非常脆弱的。文献[81]还指出，2003 年的测评结果显示，对基于电话的汉语语音（Mandarin conversational telephone-based speech，Mandarin CTS）的字错误率最高为 42.7%（见图 1-1 中灰色方块），基于英语广播语音的一种低速系统的字错误率最低为 9.9%（见图 1-1 中灰色圆点）。由此可见，汉语语音识别较其他语种而言，其任务更加艰巨。

日本在 1981 年的第五代计算机计划中提出了有关语音识别输入-输出自然语言的宏伟目标，1987 年提出高级人机口语接口和自动电话翻译系统。开展研究工作时，他们建立全国的合作体系、分派任务、避免重复；共享通用语音资料、尽可能采用标准化设备和分析技术，交流研究结果和经验。但最终都没能实现预期目标。

IBM 推出 VTD3.0 之后，声称实现了非特定人的连续语音识别。实际上，其仅仅是对话者自适应技术做了一些改进，虽然在速度和精度上较以往的识别系统

① 美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）从 1984 年开始计划对 DARPA 语音计划中的各个识别系统进行测评，并从 1987 年正式开始进行每年一次的评估，NIST 的测评方法是根据 IBM 提出的 BLEU（bilingual evaluation understudy）测评方法的一种改进，BLEU 是一种基于N-Gram的自动评测方法。

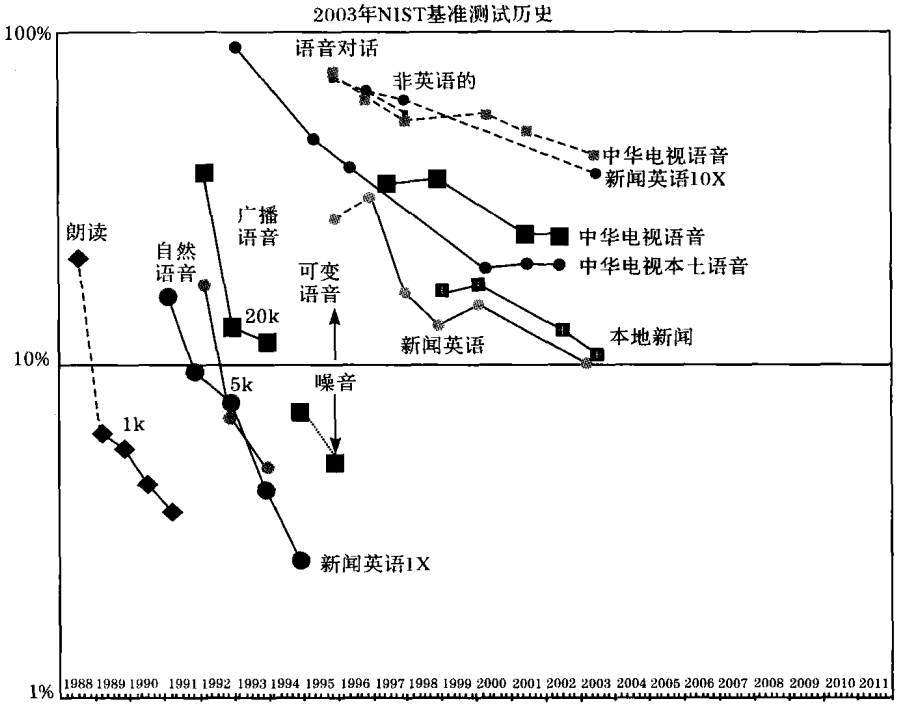


图 1-1 NIST 历年的语音测评结果的总体曲线图[81]

有了很大的进步，但由于在实际应用中仍需要根据使用者的不同而重新进行长时间的训练，因此这并不算是真正意义上的非特定人的语音识别。

总而言之，语音识别技术在近 30 年里的发展是空前的，识别系统的性能在一步步地提高，功能也在逐渐地增多，许多孤立词语音识别系统的正确识别率可以达到 98%[82,83]以上，而且各种含有语音识别功能的芯片[82,84]、软件[25]及网站（如 TOM 及时语网站①）也在不断地问世，一些简单的语音识别系统已在商业、军事、工业控制等领域得到了应用。但是，这些系统基本上仍停留在实验阶段，或者说仍然是在特定环境下的应用。因此，在继承原有语音识别系统框架的同时，还需要对语音识别技术的基础理论进行更进一步的挖掘和探索，尤其在如何提高噪声环境下的正确识别率、减少误识率、扩大词汇量、实现真正意义上的非特定人的连续语音识别等方面仍需作进一步的研究。

① TOM 及时语是北京雷霆万钧网络科技有限责任公司开发的基于互联网平台的集语音识别与合成技术于一身的中文语音门户，其链接地址是 http://www.tom.com/tomvoice，包括酒店查询、订餐服务、航班查询等服务。