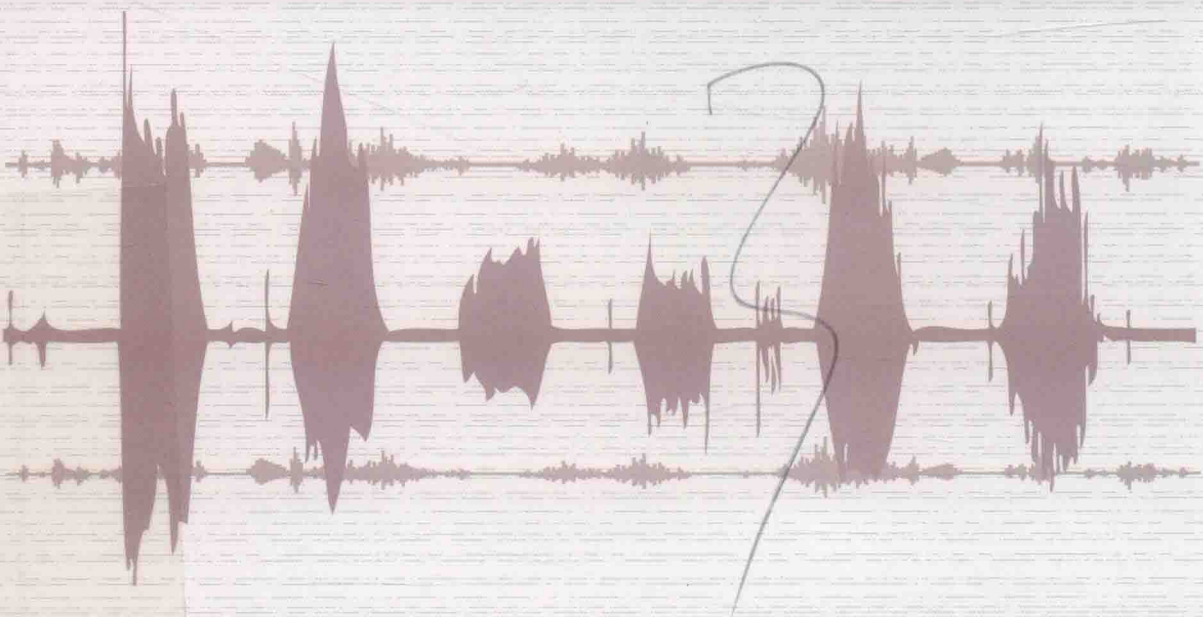
语音信号识别技术与实践

姜囡 著



東北大學出版社
Northeastern University Press

语音信号识别技术与实践

姜　囡　著

东北大学出版社
·沈　阳·

图书在版编目（CIP）数据

语音信号识别技术与实践 / 姜囡著. — 沈阳：东北大学出版社，2019.12
ISBN 978-7-5517-2380-0

Ⅰ. ①语… Ⅱ. ①姜… Ⅲ. ①语音识别－研究 Ⅳ. ①TN912.34

中国版本图书馆 CIP 数据核字(2020)第 005380 号

出 版 者：东北大学出版社
地址：沈阳市和平区文化路三号巷 11 号
邮编：110819
电话：024－83683655(总编室) 83687331(营销部)
传真：024－83687332(总编室) 83680180(营销部)
网址：http://www.neupress.com
E-mail: neuph@neupress.com
印 刷 者：沈阳市第二市政建设工程公司印刷厂
发 行 者：东北大学出版社
幅面尺寸：170mm×240mm
印 张：16
字 数：287 千字
出版时间：2019 年 12 月第 1 版
印刷时间：2020 年 1 月第 1 次印刷
责任编辑：郎 坤
责任校对：刘乃义
封面设计：潘正一
责任出版：唐敏志

ISBN 978-7-5517-2380-0 定 价：58.00 元

前　言

“大弦嘈嘈如急雨，小弦切切如私语。嘈嘈切切错杂弹，大珠小珠落玉盘。间关莺语花底滑，幽咽泉流冰下难。冰泉冷涩弦凝绝，凝绝不通声暂歇。别有幽愁暗恨生，此时无声胜有声。”白居易妙笔巧慧，以声绘声，一曲《琵琶行》读来既让人如闻其声、如临其境，又能够深切地体会到作者与演奏者的强烈情感共鸣，这曲千古绝唱充分展示了声音的魅力。

声音是我们体验世界的重要感觉之一，听觉和视觉起着相互补充的作用，听觉甚至比视觉更重要。我们常常能在看见声源之前就听见它，还能通过声音了解那些看不见的信息，比如是谁在说话，声音是否有遮掩，说话者处于怎样的一种情绪状态。

在大数据时代，我们拥有越来越强的计算能力、越来越低的计算成本，人工智能逐渐渗透进我们的生活，在这个智能技术蓬勃发展、熠熠生辉的时代，关于声音、语音的研究，能为当下及未来做些什么呢？我们已经非常熟悉手机里的语音助手，她可以帮助我们查找信息，作日程提醒，甚至可以陪我们聊天，但是目前的语音助手只能识别我们的语言信息，不能辨别我们的情绪，反馈给我们的是音调平平、不带任何情感的声音。想象一下语音助手的未来发展，她能从我们的一声叹息里识别我们的抑郁情绪，然后温柔地告诉我们时间会治愈一切伤痛；她能在一个阴雨的午后，提醒一名独居老人吃药，并用滑稽的语气为老人唱上一段他喜欢的老歌，用声音让家里洒满愉快而明媚的阳光；她能被藏在办案警员的口袋里，偷偷地告诉他嫌疑人是否情绪紧张，是否在说谎，是否为审讯突破的最佳时机……探索不停，未来可期。

本书作者与其团队成员对语音及其情感信息识别具有浓厚的兴趣，在语音识别和情感识别方面迈开了探索的一小步。本书内容是作者及其团队成员在初步研究成果的基础上，按照语音识别的步骤，由浅及深，由易到难，加以归类

和整理而成的，旨在为对语音及情感识别感兴趣的初学者提供学习脉络和研究思路。

本书内容分为 8 章。第 1 章为语音识别技术概述，介绍了语音识别技术的原理和发展与应用。第 2 章为语音信号处理基本技术，包括数字化预处理、短时时域处理和频域处理的内容。第 3 章是语音信号的端点检测和分割，介绍了端点检测的原理和常规检测方法，提出了基于复杂背景条件下的端点检测算法，包括算法流程和实验方法。第 4 章是语音分割聚类，研究了如何获取一段多人对话语音中说话人身份变动的信息，以及如何确定哪些语音段是由同一个人发出来的。详细介绍了三种方法，包括基于混合特征的分割聚类方法、基于改进双门限端点检测的分割法、基于自组织神经网络的改进 *K*-means 聚类算法。第 5 章为基于神经网络的语音识别，详述了基于自适应免疫克隆神经网络的语音识别算法原理、流程和实验方法。第 6 章是伪装语音识别，探讨了在语音被采用伪装手段（如在耳语、假声、模仿他人讲话、捏鼻子讲话以及用手绢或口罩等物品捂嘴讲话等）情况下，如何正确进行语音鉴定的问题。提出了基于 GFCC 与共振峰的声纹提取方法和基于深度置信网络模型的声纹提取方法。第 7 章是基于语音信号的心理压力分级与识别，探讨了反映心理压力的生理信号和分级实验方法，以及基于语音信号的心理压力识别方法。第 8 章是不同情感的语音声学特征分析，通过对生气、害怕、高兴、中性、惊讶、悲伤六种情感语音的共振峰频率特征、共振峰走向特征、音节间的过渡特征、音节内的过渡特征、基频曲线特征以及振幅曲线特征进行语音声学特征分析，探索了同一个人的语音在不同情感下表现的特征差异。

本书较全面地总结了课题组近年来关于语音识别、语音与心理压力等级识别、语音与情感分析方面的研究内容。主要章节均以理论介绍、算法流程、实验步骤、结果分析为脉络撰写，内容详尽，循序渐进，适合语音识别及语音情感分析的初学者，希望为在此领域有求知欲的学子打开一扇探索之门。

本书的出版得到了国家自然科学基金项目（61304021）、科技部国家重点研发专项项目（2017YFC0821005）、公安理论及软科学项目（2017LLYJXJXY040）、现场物证溯源技术国家工程实验室开放课题（2017NELKFKT08）、中央高校基本科研业务费（D2018004）、辽宁省自然科学基金项目（2019-ZD-0168，2016010808-301，20170540984）、辽宁省博士科研启动基金项目（201601091）、辽宁省教育厅科学研究一般项目

（L2015198）、中国刑事警察学院重大计划培育项目（D2019005，D2019006，D2019007）、中国刑事警察学院教研项目（2018QNZX19）的资助。

特别感谢中国刑事警察学院研究生郭卉的协助，感谢作者的研究生团队姜艳萍、李诚、谢俊仪、刘景天、张阳、郭卉、仁杰、高旭皓、高爽、付彬、贾俊玮、余琳在本书撰写过程中的大力支持！同时，东北大学出版社对本书出版给予了许多帮助和支持，作者谨借此机会表达深切的谢意。

“道在日新，艺亦须日新，新者生机也。”语音及其情感识别的研究之路漫漫，其探索之方向浩瀚如宇宙，本书仅仅撕开了这个神秘领域的一个小小边角，作者及其团队成员为这一角落照射出来的绚烂光芒所吸引，希望能够与广大读者共同探讨，携手追寻科技之光，砥砺前行，开疆拓土。

限于作者水平和能力，不当之处在所难免，恳请各位专家学者给予批评指正。

著者

2019 年 8 月

目 录

第 1 章　语音识别技术概述

语音识别技术是指计算机能够判断出人说话的内容，其根本目的是使计算机可以具有类似于人的听觉系统，能够获得人的语音并理解其中的意图[1-4]。语音识别的研究有重要意义，特别是对汉语来说，汉字的书写和录入较为复杂，因而通过语音来输入汉字信息就特别重要。而且，计算机键盘的操作也远没有语音输入方便，更加显现出语音识别的便捷性，所以语音识别在计算机智能接口及多媒体中有巨大的应用潜力。

基于统计模式识别的语音识别研究技术目前最为常见[5-7]。一个完整的语音识别系统大致有以下三部分。

① 语音信号的预处理，即预先处理原始语音信号。

② 语音信号的特征提取，即特征参数分析，以获得一组可以描述语音信号特征的参数[8-10]。

③ 语音的训练和识别方法，如 DTW、VQ、FSVQ、LVQ2、HMM、TDNN、模糊逻辑算法等，也可以混合使用。

语音识别系统处理流程一般如图 1.1 所示。

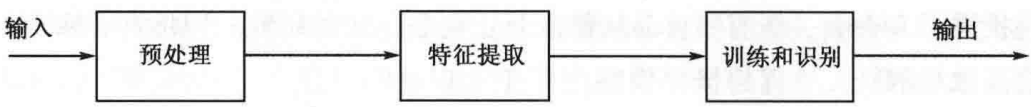


图 1.1　语音识别系统处理流程

1.1　语音识别的基本原理

语音识别分为两步。第一步为训练过程，选择适当的算法，并提取恰当的语音特征参数作为标准模板存储起来，形成标准模板库[11-14]。第二步是识别过

程。语音识别系统基本结构与常规的模式识别系统相同，包括特征提取、模式匹配、参考模板库三部分[15-16]。可将其处理过程看成一个框架，如图 1.2 所示。图 1.2 中，模式匹配主要通过测度估计、识别决策及专家知识三部分实现。

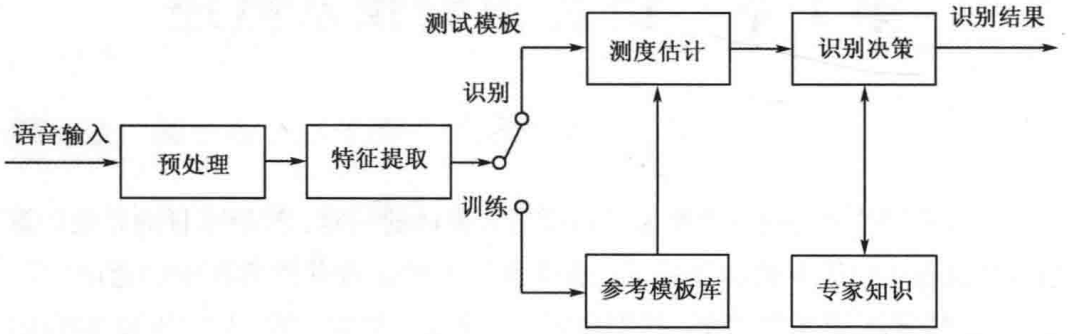


图 1.2　语音识别基本过程

1.2　语音识别技术的发展

最开始对语音识别进行深入探讨是在 20 世纪 40—50 年代，这个时期可以称为语音识别奠定基础的时期[17-19]，计算机也产生在这一时期。这个时期，在两个重要基础(自动化以及概率学或者信息理论模型)上进行了大量的研究。

自动化起源于 20 世纪 50 年代，该算法从 Turning 模型中得到，被很多人认为是现代计算机科学的基础。Turning 的研究首先帮助了 McCulloch-Pitts 神经元的研究，这是一种神经元的简化模型，被用作一种计算元素。Turning 的研究还推进了 Kleene 关于有限自动操作以及正则表达式的研究。1948 年，Shannon 将离散马尔科夫处理的概率模型应用于实现自动语言[20]。1956 年，Chomsky 从 Shannon 的研究中得到灵感，首先考虑将有限状态机作为描述一种语法的方式，并且将有限状态语音定义为由有限状态语法产生的语音。这些早期的模型促使了形式语言理论的产生，其用代数以及集理论将形式语言定义为符号序列[21]，包括与文本无关的语法。这是最初由 Chomsky 于 1956 年为自然语言作的定义，其后又分别由 Backus(1959 年)和 Naur 等人(1960 年)在各自关于 ALGOL 处理语音的描述中提出。这个时期第二个重要的研究是语音处理相关概率算法的发展，这源于 Shannon 的其他研究成果。Shannon 还将热力学中熵的概念引入作为测试通道信息能力—— 或者说语言的信息内容——的一种方式。

英语的熵最先被测量，当时使用的是概率技术。还是在这个时期，语谱图被 Koening 等人提出，他们在仪器语音学领域也进行了重要的研究，成为语音识别后期工作的基石。这促使了 20 世纪 50 年代早期第一个语言识别系统的产生[22]。1952 年，贝尔实验室的研究人员建立了一个基于统计学的系统，该系统可以识别单个说话人所说的任意 10 个数字[23]。这个系统有 10 个与说话人有关的模板库，大体代表数字中的头两个元音共振峰。该系统通过选取与输入有着最高相关系数的模式获得了高达 97%~99%的准确性。

20 世纪 50 年代末 60 年代初，语音处理被划分为经典模式和随机模式两种。经典模式在两项研究后开始迅速发展。第一项是 Chomsky 等人从 20 世纪 50 年代后期到 60 年代中期对形式语言理论以及句法方面的研究，以及很多语言学家和计算机科学家关于解析算法的研究，刚开始为自上而下以及自下而上算法，其后发展为动态编码算法。最早的一个完整解析系统是 Zelig Harris 的转换与语篇分析项目(TDAP)，并于 1958 年 6 月—1959 年 7 月安装在宾夕法尼亚大学。第二项研究是人工智能(AI)新领域的研究。AI 一直吸引着小部分研究者对随机和统计算法(包括概率模型以及神经网)进行探索，新领域的主要关注点转向为以 Newell 和 Simon 关于逻辑理论和通用问题解算机为代表的推理和分析方面的研究。早期的自然语言理解系统就建立于这个时期。这些简单系统主要通过结合模式识别以及以简单启发式推理和问答为基础的关键词搜索来研究单一领域。20 世纪 50 年代，贝叶斯方法开始被用于解决最优字符识别问题。Mosteller 和 Wallace 在 1954 年将贝叶斯方法用于解决《联邦党人文集》的著作权归属问题。Bledsoe 和 Browning 在 1959 年研究出一个使用大字典以及通过将每个单词的相关性相乘来计算字典中每个观测单词序列的相关性的贝叶斯文本识别。到 20 世纪 60 年代末，人们研究出更为正规的逻辑系统。随机模式主要集中在统计学以及电子工程领域。

20 世纪 60 年代，基于转换语法的第一个用于处理人类语言的正式可测试的心理模型开始兴起，建立了第一个在线语料库——美国英语语料库，其中包括来自不同种类(报纸、小说、非小说等)的 500 个书面文本样本。William S-Y. Wang 在 1967 年完成了计算机字典和在线中文方言词典。

1970—1983 年是语音和语言处理研究井喷的时代，也推进了许多研究模式的发展，这些模式现在仍然是语音识别中常用的模式。随机模式在这个时期的语音识别算法研究中起着极为重要的作用，尤其是 Jelinek，Bahl，Mercer，IBM

的 Thomas J. Watson 研究中心的研究人员以及 Carnegie Mellon 大学的 Baker 分别探索的隐马尔科夫模型(HMM)的使用、噪声信道的模拟以及解码。AT & T 的贝尔实验室是另一个研究语音识别和合成的重要研究中心[24]。HMM 描述语音信号过程是 20 世纪 80 年代的一项重大进展，HMM 已构成现代语音识别的重要基石[25]。Colmerauer 和他的同事关于 Q 系统和变形语法的研究开启了基于逻辑的模式的发展。同时，Kay 在 1979 年关于功能语法的研究，以及 Bresnan 和 Kaplan 于 1982 年关于词汇功能语法的研究，树立了特征结构归一化的重要性。自然语言理解领域也是在这个时期开始发展的，最先产生的是 Winograd 的 SHRDLU 系统。其中的程序可以处理极其复杂的自然语言文本指令。他的系统也是第一个尝试建立一个基于 Halliday 的系统化语法的广泛英语语法系统。Winograd 的模型使解析问题被很好地理解，并开始关注语义学和话语。Roger Schank 和他的同事以及学生打造了一系列专注概念领域的语言程序，如脚本、计划、目标和人类记忆组织。话语建模模式专注话语中的四个关键领域。Grosz 和她的同事介绍了关于话语中子结构和话语重点的研究；许多研究者开始研究自动参考解析，并且针对语音行为的基于逻辑研究的 BDI 框架也得到了发展。

1983—1993 年，两类在 20 世纪 50 年代后期和 60 年代早期没落的模型开始再次出现在人们眼前，部分源于反对它们的理论观点。第一类是有限状态模型，该模型在 Kaplan 和 Kay 于 1981 年关于有限状态音系学和形态学的研究以及 Church 于 1980 关于语法的有限状态模型的研究后开始重新得到关注[26]。第二类是经验主义的回归模型。最值得注意的是贯穿整个语音处理的概率模型的发展，这大部分是由于 IBM 的 Thomas J.Watson 研究中心关于语音识别概率模型的研究而引起的。这些概率方法和其他数据驱动的方法从语音开始扩展到词性标注、解析歧义、附着歧义以及语义学。这个经验方向还伴随着对模型评估的新关注点，这个评估基于使用输出数据得到评估量化矩阵，以及重点关注这些矩阵和之前研究的性能比较。

到 20 世纪最后五年，语音识别领域经历了重大的变化。首先，概率和数据驱动的模型在自然语言处理中变得极为典型。解析算法、词性标签算法、参考解析算法和话语处理算法都开始融合概率学理论以及使用评估方法学。其次，计算机速度和存储的发展允许语音识别的商业实现。

经验趋势起源于 20 世纪 90 年代后期，并且在 21 世纪以惊人的速度发展。

这种发展主要是由三个协同趋势引发的。首先，大量的语音和书面材料可通过语言数据组合(LDC)和其他类似方法轻松得到。这些资源的存在帮助解决了更为复杂的传统问题，例如有监督机器学习中的解析和语义分析问题。这些资源还有助于解析、信息提取、词义消歧、问答和总结的其余竞争评估的建立。其次，它对学习的日益关注引发了与统计机器学习理论更加正统的相互作用。如支持向量机、最大化熵技术、与多项式逻辑回归类似的公式和图形化贝叶斯模型等成为计算语言学中的经典技术。再次，高性能的计算机系统的普及帮助了系统的训练和实现，这些系统在十年前是无法想象的。最后，无监督统计学方法开始重新被关注。人们开始从统计学方法转向机器学习。大成本以及生成可靠的注释语料库成为有监督方法使用的一个限制因素，转向使用无监督学习技术的趋势将很可能愈演愈烈。

以下五个方面可以用于控制和简化语音识别任务。

(1)孤立词

识别由孤立的单词(每个单词间停一小会儿)组成的语音要比识别连续语音轻松得多，这是因为很难找出连续语音中词之间的界限。连续语音中的协同发音效应导致一个单词的发音会随着它在一个句子相对其他词的位置而产生变化。让说话人在每个单词间停一会儿显然可以降低语音识别的错误率。然而，这一类的限制会给使用者带来不便，也会降低语音识别系统的输入速度。

(2)单个说话人

识别单个说话人的语音要比识别一群人的语音简单得多，这是因为语音绝大部分的参数特征与特定说话人的个性密切相关[27]。这就导致了由某个说话人得到的模式匹配模板对于另一个说话人来说可能非常不好用。因此，很多语音识别系统为面向单个说话人的。非常少的语音识别系统可以在公共场合被有效使用。很多研究者发现，对于同一个语音识别任务，面向单个说话人的准确率是面向多个人的准确率的 3~5 倍。

使一个语音识别系统可以处理多个人的任务的一种简单方法为将很多人训练得到的模板混在一起使用。另一个更为复杂些的方法为寻找说话人间相对固定的语音特征。

(3)词的大小

这里，预处理语音信号是为语音识别算法做准备的[28-29]。待识别词的大小同样会严重影响到语音识别精度。大的词相对小的词而言更可能包含歧义词。

歧义词是这些模式匹配模板中对识别系统中的分类算法而言非常类似的单词。因此，将它们区分开更加困难。当然，包含很多歧义词的小的词也特别不好识别。

语音识别系统搜索语音模型数据库所花费的时间与词的大小有关。包含很多模型模板的系统一般需要剪枝技术来减少模式匹配算法的计算量。由于忽略了理论可能有用的搜索路径，剪枝算法可能会产生识别错误。

(4)语法

识别领域的语法决定了单词可允许的序列。对单词选择上限制的多少被称为语法的复杂度。复杂度低的语音识别系统比起可以让说话人更加自由发挥的识别系统而言要更为精确些，这是因为该系统可以将有效词和搜索空间限制为与当前输入上下文相关的单词。

(5)环境因素

这个算法必须能够计算说话人语音经过预处理后的模板以及所有的存储模板或者语音模型之间的拟合优度的度量。一个选择过程是以最好的匹配要求去选择需要的模板。环境噪声在麦克风特性中会产生变化，音量也能极大地影响识别精度。很多语音识别系统能够在安静且可控的环境条件下获得很低的识别错误率。然而，当存在噪声或者当背景环境与训练参考模板时的环境不一样时，识别性能将急剧下降。为了弥补这一点，说话人通常需要佩戴与训练时使用的麦克风特性一致的头戴式限噪麦克风。

1.3 语音识别技术的应用

比尔·盖茨说过："语音技术将使计算机丢下鼠标和键盘。"随着计算机的小型化，键盘和鼠标已经成为计算机发展的一大阻碍。计算机从超大体积发展到现在占地不到1平方米的微型计算机，想必未来的计算机可能会意想不到得小，那么键盘和鼠标对其来说就是障碍了，这时候就需要语音识别来完成命令。一些科学家也说过："计算机的下一代革命就是从图形界面到语音用户接口。"这表明语音识别技术的发展无疑改变了人们的生活。在某些领域，手机正在逐渐地演变成一个服务者而非简单的对话工具，通过手机，人们也可以使用语音来获取自己想获得的信息，其工作效率也自然而然地提高了一个档次。

语音识别技术渐渐地变成人机接口的关键一步，这样一个极具竞争性的新

兴产业的发展更是十分迅速，发展趋势也在逐步上升。1999—2005 年，语音识别技术市场正在以每年 31%的趋势增长，如今在智能手机中，语音助手已经成为了标配功能，为用户带来了许多的便利，人们也可以通过电话和网络来订购机票、火车票，甚至是旅游服务。因此，语音识别技术在人们的实际生活中也有着越来越广阔的发展前景和应用领域。

在手机与通信系统中，智能语音接口正在把手机从一个单纯的服务工具变成一个服务的“提供者”和生活“伙伴”；使用手机与通信网络，人们可以通过语音命令方便地从远端的数据库系统中查询与提取有关的信息；随着计算机的小型化，键盘已经成为移动平台的一个很大障碍，想象一下，如果手机仅仅只有一个手表那么大，那么再用键盘进行拨号操作是不可能的。语音识别正逐步成为信息技术中人机接口的关键技术，语音识别技术与语音合成技术结合使人们能够甩掉键盘，通过语音命令进行操作。语音技术的应用已经成为一个具有竞争性的新兴高技术产业。

语音识别技术发展到今天，中小词汇量非特定人语音识别系统的识别精度已经大于 98%，对特定人的语音识别系统的识别精度就更高。这些技术已经能够满足通常应用的要求。随着大规模集成电路技术的发展，这些复杂的语音识别系统也已经完全可以制成专用芯片，大量生产。在西方经济发达国家，大量的语音识别产品已经进入市场和服务领域。一些用户交机、电话机、手机已经包含了语音识别拨号功能，还有语音记事本、语音智能玩具等产品也包括语音识别与语音合成功能。人们可以通过电话网络用语音识别口语对话系统查询有关的机票、旅游、银行信息，并且取得了很好的结果。调查统计结果表明，多达 85%以上的人对语音识别的信息查询服务系统的性能表示满意。

可以预测在近 5～10 年内，语音识别系统的应用将更加广泛。各种各样的语音识别系统产品将出现在市场上。人们也将调整自己的说话方式以适应各种各样的识别系统。在短期内还不可能造出具有和人相比拟的语音识别系统，建成这样一个系统仍然是人类面临的一个大的挑战，我们只能朝着改进语音识别系统的方向一步步地前进。至于什么时候可以建立一个像人一样完善的语音识别系统，则是很难预测的。就像在 20 世纪 60 年代，谁又能预测今天超大规模集成电路技术会对人类社会产生这么大的影响？