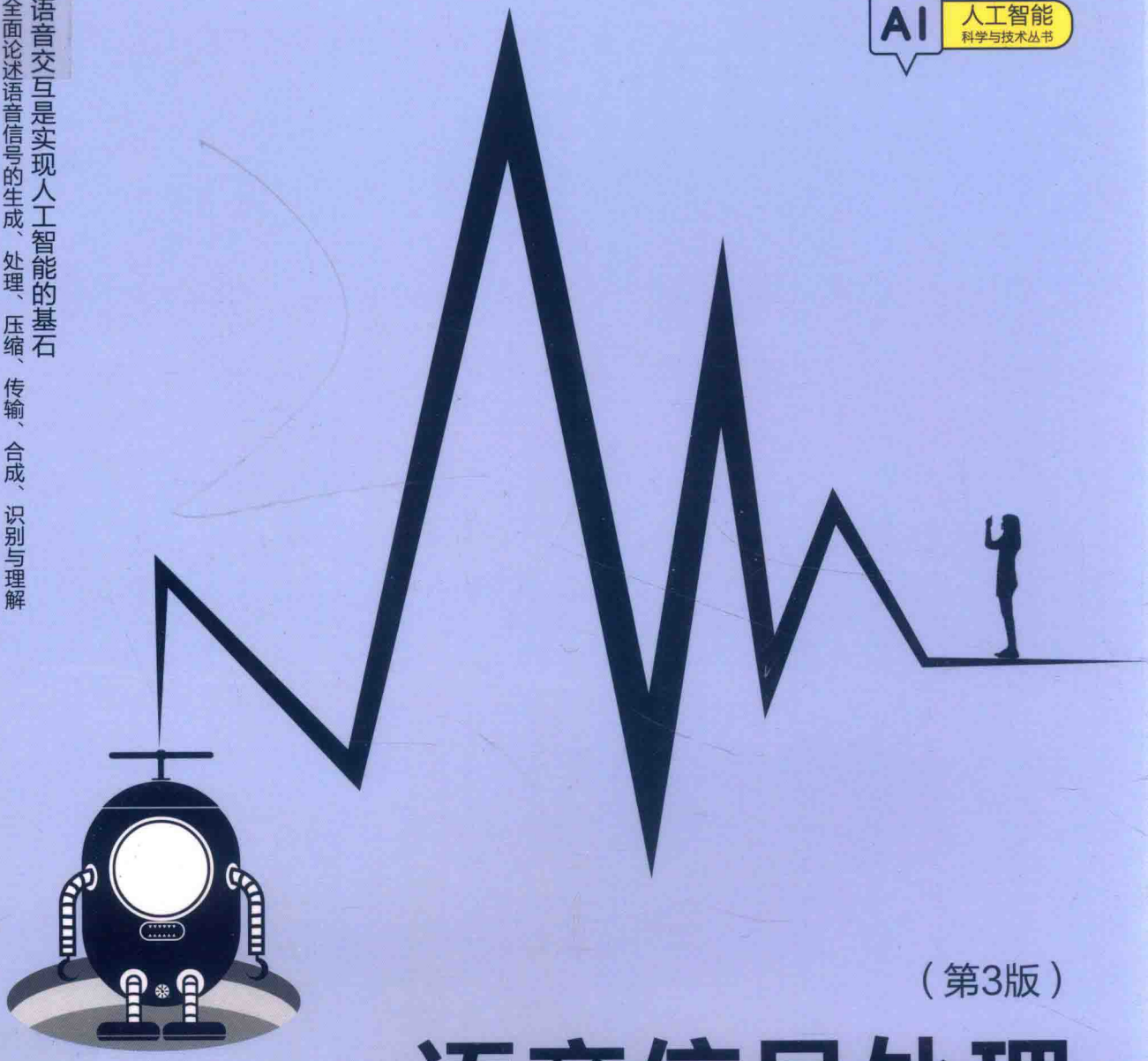
AI 人工智能
科学与技术丛书

语音交互是实现人工智能的基石
全面论述语音信号的生成、处理、压缩、传输、合成、识别与理解



（第3版）

语音信号处理

韩纪庆 张磊 郑铁然◎编著
Han Jiqing Zhang Lei Zheng Tieran

SPEECH SIGNAL PROCESSING
THIRD EDITION

清華大學出版社

AI
人工智能
科学与技术丛书

语音信号处理

（第3版）

韩纪庆 张磊 郑铁然◎编著
Han Jiqing Zhang Lei Zheng Tieran

SPEECH SIGNAL PROCESSING
THIRD EDITION

清華大學出版社
北京

内 容 简 介

本书系统地介绍语音信号处理的基础、概念、原理、方法与应用。全书共分 9 章。第 1 章介绍语音信号处理及其发展过程；第 2 章介绍语音信号的产生与人类听觉的机理，传统的线性语音产生模型，以及非线性语音产生模型；第 3 章从语音信号的时域特征入手，引入时频分析的思想，并进一步阐述时频分析中短时傅里叶变换和小波变换在语音信号特征分析中的应用，最后对广泛使用的倒谱特征以及同态解卷积进行介绍；第 4 章介绍语音信号的线性预测原理、解法、几种推演方法以及线谱对分析法；第 5 章介绍语音编码的相关知识，包括语音的波形编码、极低速率语音编码技术，以及相关编码器的性能指标和评测方法；第 6 章介绍语音识别的基本内容，从基于矢量量化的识别技术到动态时间归正的识别技术，从隐马尔可夫模型技术到基于深度学习的语音识别技术，从孤立词识别到连接词识别及连续语音识别技术，再到关键词检出技术，最后还介绍新兴起的语音识别应用技术，以及用于 HMM 系统构建的 HTK 工具和用于深度学习系统构建的 Kaldi 工具等；第 7 章介绍说话人识别的基本内容，从基于 GMM-UBM 的识别技术到基于支持向量机的识别技术，从基于联合因子分析的识别技术到基于 i-vector 的识别技术，以及近年来受到关注的基于深度学习的识别技术等；第 8 章介绍顽健语音识别技术，从影响语音识别性能的环境变化因素分析开始，介绍噪声环境下顽健语音识别技术，以及变异语音识别的技术；第 9 章介绍语音合成的基本原理、线性预测合成、共振峰合成以及汉语按规则合成，以及基于 HMM 的合成技术等内容。

本书可作为高等院校计算机应用、信号与信息处理、通信与电子系统等专业及学科的高年级本科生、研究生教材，也可供该领域的科研及工程技术人员参考。

图书在版编目(CIP)数据

语音信号处理/韩纪庆，张磊，郑铁然编著. —3 版. —北京：清华大学出版社，2019
(人工智能科学与技术丛书)
ISBN 978-7-302-51760-3

Ⅰ. ①语…　Ⅱ. ①韩… ②张… ③郑…　Ⅲ. ①语声信号处理－青少年读物　Ⅳ. ①TN912.3-49

中国版本图书馆 CIP 数据核字(2018)第 271405 号

责任编辑：盛东亮
封面设计：李召霞
责任校对：李建庄
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：27.25　　**字　　数**：663 千字
版　　次：2004 年 9 月第 1 版　2019 年 5 月第 3 版　　**印　　次**：2019 年 5 月第1次印刷
定　　价：89.00 元

产品编号：078637-01

前言

PREFACE

语音信号处理以语音为研究对象，涉及心理学、生理学、语言学、数字信号处理、模式识别、人工智能、机器学习等诸多研究领域，甚至还涉及人说话时的表情、手势等体态语言信息。由于语音是人们日常生活中的主要交流手段，因此语音信号处理在现代信息社会中占有重要地位。

语音信号处理的研究工作最早可以追溯到19世纪70年代，在20世纪得到了长足的发展，并在20世纪90年代，随着IBM、Microsoft、Apple、AT&T、NTT等著名公司为语音识别的实用化开发投以巨资，掀起了语音信号处理技术的应用热潮。进入21世纪，伴随着以深度神经网络为代表的深度学习理论的全面突破、以通用图形处理器(GPU)为代表的硬件技术的迅猛发展，语音识别的性能得到显著提高，从而迎来了语音信号处理技术的蓬勃发展。

目前在语音信号处理领域中不断有新的技术涌现。本书再版的目的就是将这些新的技术融合到已有的相关理论与技术中。全书以语音信号处理过程的总体框架为线索，全面阐述语音信号的前端处理技术、语音编码技术、语音识别和说话人识别技术，以及语音合成技术。相对于上一版，本书补充了基于深度学习的语音识别、基于i-vector的说话人识别等本领域的前沿理论和技术，以利于读者充分了解最新的学术发展动态，并能在学术思想上受到启发。同时，书中也介绍了当前深度学习方法中广泛采用的Kaldi工具的使用技巧，以帮助读者掌握相关的实践手段。

本书涉及作者承担的多项国家自然科学基金项目的部分研究成果，在内容上既注重基本理论的系统性，又兼顾实用性和可读性，可作为高等院校计算机应用、信号与信息处理、通信与电子系统等专业及学科的高年级本科生、研究生教材，也可供该领域的科研及工程技术人员参考。

本书的第1、2、4章由韩纪庆编写，第3、6、9章由张磊编写，第5、7、8章由郑铁然编写。韩纪庆负责全书的总体安排和审定。在新版增加的内容中，郑铁然在基于深度学习的语音识别部分、陈晨在说话人识别部分、史秋莹在Kaldi工具部分的撰写上作出了重要贡献。郑贵滨为书稿的完善做了大量工作，在此表示感谢！

本书虽然是作者从事语音信号处理工作30年的理论与实践的结晶，但因作者水平有限、时间仓促，缺点和错误在所难免，敬请读者批评指正，提出宝贵意见。

作　者

于哈尔滨工业大学

2019年1月

目录

CONTENTS

第1章 绪 论

CHAPTER 1

语言是人类最重要的交流工具，它自然方便、准确高效。随着社会的不断发展，各种各样的机器参与了人类的生产活动和社会活动，因此改善人和机器之间的关系，方便人对机器的操纵就显得越来越重要。随着电子计算机和人工智能机器的广泛应用，人们发现，人和机器之间最好的通信方式是语言通信。而语音是语言的声学表现形式；要使机器听懂人的语言并能使用人类的语言进行表达，需要做很多工作，这就是研究了几十年的语音识别和语音合成技术。而随着移动通信的迅猛发展，人们可以随时随地通过电话进行交流，其中语音压缩编码技术发挥着重要的作用。上述这些应用领域构成了语音信号处理技术的主要研究内容。

语音信号处理是语音学与数字信号处理技术相结合的交叉学科，它和认知科学、心理学、语言学、计算机科学、模式识别和人工智能等学科联系紧密；语音信号处理技术的发展依赖这些学科的发展，而语音信号处理技术的进步也会促进这些学科的进步。

1.1 语音信号处理的发展

语音信号处理的研究工作最早可以追溯到1876年贝尔发明的电话，它首次完成了用声电—电声转换来实现远距离传输语音的技术。1939年，Dudley研制成功了第一个声码器，从此奠定了语音产生模型的基础，这一工作在语音信号处理领域具有划时代的意义。1947年，贝尔实验室发明了语谱图仪，将语音信号的时变频谱用图形表示出来，为语音信号的分析提供了一个有力的工具。1948年，美国Haskins实验室研制成功“语图回放机”，它把手工绘制在薄膜片上的语谱图自动转换为语音，可以进行语音合成。共振峰合成方法就是源于这一思想。

对语音识别而言，它的研究相对较晚，起源于20世纪50年代。语音识别技术的根本目的是研究出一种具有听觉功能的机器，能接收人类的语音，理解人的意图。由于语音识别本身所固有的难度，人们提出了各种限制条件下的研究任务，并由此产生了不同的研究领域。这些领域包括：按说话人的限制，可分为特定说话人语音识别和非特定说话人语音识别；按词汇量的限制，可划分为小词汇量、中词汇量和大词汇量的识别；按说话方式的限制，可分为孤立词识别和连续语音识别等。最简单的研究领域是特定说话人小词汇量孤立词的识别，而最难的则是非特定说话人大词汇量连续语音的识别。

1952年,贝尔实验室的Davis等研制了特定说话人孤立数字识别系统。该系统利用每个数字元音部分的频谱特征进行识别。1956年,RCA实验室的Olson等也独立地研制出10个单音节词的识别系统,系统采用从带通滤波器组获得的频谱参数作为语音的特征。1959年,Fry和Denes等尝试构建音素识别器来识别4个元音和9个辅音,采用频谱分析和模式匹配来进行识别决策,其突出贡献在于,使用了英语音素序列中的统计信息来改进词中音素的精度。1959年,MIT林肯实验室的Forgie等,采用声道的时变估计技术对10个元音进行识别。

20世纪60年代初期,日本的很多研究者开发了相关的特殊硬件来进行语音识别,如东京无线电研究实验室Suzuki等研制的通过硬件来进行元音识别的系统。在此期间开展的很多研究工作对后来近二十年的语音识别研究产生了很大的影响。RCA实验室的Martin等在20世纪60年代末开始研究语音信号时间尺度不统一的解决办法,开发了一系列的时间归正方法,明显地改善了识别性能。与此同时,苏联的Vintsyuk提出了采用动态规划方法来解决两个语音的时间对准问题。尽管这是动态时间弯折算法(dynamic time warping,DTW)的基础,也是连接词识别算法的初级版,但Vintsyuk的工作并不为学术界的广大研究者所知,直到20世纪80年代大家才知道Vintsyuk的工作,而这时DTW方法已广为人知。

值得一提的是20世纪60年代中期,斯坦福大学的Reddy开始尝试用动态跟踪音素的方法来进行连续语音的识别。后来Reddy加入卡内基梅隆大学,多年来在连续语音识别上开展了卓有成效的工作,直至现在仍然在此方面居于领先地位。

20世纪70年代之前,语音识别的研究特点是以孤立词的识别为主。20世纪70年代,语音识别研究在多方面取得了诸多的成就,在孤立词识别方面,日本学者Sakoe给出了使用动态规划方法进行语音识别的途径——DTW算法,它是把时间归正和距离测度计算结合起来的一种非线性归正技术。这是语音识别中一种非常成功的匹配算法,当时在小词汇量的研究中获得了成功,从而掀起了语音识别的研究热潮。Itakura利用语音编码中广泛使用的线性预测编码(linear predictive coding,LPC)技术,通过定义基于LPC频谱参数的合适的距离测度,成功地将其扩展应用到语音识别中。以IBM为首的一些研究单位还着手开展了连续语音识别的研究,AT&T的贝尔实验室也开展了一系列非特定说话人语音识别方面的研究工作。

应该指出的是,20世纪70年代,人工智能技术开始被引入到语音识别中。美国国防部的高级研究规划局(Advanced Research Projects Agency,ARPA)组织了有卡内基梅隆大学等五个单位参加的一项大规模语音识别和理解的研究计划,当时专家们认为:要使语音识别研究获得突破性进展,必须让计算机像人那样具有理解语言的智能,而不必过多地在孤立词识别上下功夫。在这个历时五年的庞大的研究计划中,最终在语言理解、语言的统计模型等方面积累了经验,其中卡内基梅隆大学完成的Hearsay-II和Harpy两个系统效果最好。在这两个系统中,引用了"黑板模型"来完成底层和顶层之间不同层次的信息交换和规则调用,成为以后其他专家系统研究工作中的一种规范。但从整体上看,这个计划并没有取得突破性的进展。

20世纪70年代末80年代初,Linda、Buzo、Gray等提出了矢量量化(vector quantization)码本生成的方法,并将矢量量化技术成功地应用到语音编码中,从此矢量量化技术不仅在语音

识别、语音编码和说话人识别等方面发挥了重要的作用，而且很快推广应用到其他领域。这一时代，语音识别的研究重点之一是连接词识别，典型的工作是进行数字串的识别。研究者提出了各种连接词语音识别算法，大多数工作是基于对独立的词模板进行拼接来进行匹配的方法，如两级动态规划识别算法、分层构筑(level building)、帧同步(frame synchronous)分层构筑方法等。这些方法都有各自的特点，广泛用于连接词识别当中。

20世纪80年代开始，语音识别研究的一个重要进展，就是识别算法从模式匹配技术转向基于统计模型的技术，更多地追求从整体统计的角度来建立最佳的语音识别系统。隐马尔可夫模型(hidden markov model，HMM)技术就是其中的一个典型；尽管开始的时候仅有较少的单位采用这种模型，但由于该模型能很好地描述语音信号的时变性和平稳性，具有把从声学—语言学到句法等统计知识全部集成在一个统一框架中的优点，因此从20世纪80年代起，它被广泛地应用到语音识别研究中。直到目前为止，HMM方法仍然是语音识别研究中的主流方法。HMM的研究使大词汇量连续语音识别系统的开发成为可能。20世纪80年代末，美国卡内基梅隆大学用VQ/HMM实现了997词的非特定人连续语音识别系统SPHINX，这是世界上第一个高性能的非特定人、大词汇量、连续语音识别系统。此外，BBN的BYBLOS系统，林肯实验室的识别系统等也都具有很好的性能。这些研究工作开创了语音识别的新时代。

从20世纪80年代后期和90年代初开始，人工神经网络(artificial neural network，ANN)的研究异常活跃，并且被应用到语音识别的研究中。进入20世纪90年代后，相应的研究工作在模型设计的细化、参数提取和优化，以及系统的自适应技术等方面取得了一些关键性的进展，使语音识别技术进一步成熟，并且出现一些很好的产品。许多发达国家，如美国、日本、韩国，以及IBM、Microsoft、Apple、AT&T、NTT等著名公司都为语音识别系统的实用化开发研究投以巨资。

进入21世纪，基于深度学习理论的语音识别得到了全面突破，识别性能显著提高。2006年，加拿大多伦多大学的Hinton等提出了一种深度神经网络(deep neural network，DNN)模型——深度置信网络模型(deep belief network，DBN)。它由一组受限玻尔兹曼机(restricted boltzmann machine，RBM)堆叠而成，其核心部分是贪婪的逐层无监督学习算法，其时间复杂度与网络的大小及深度呈线性关系。通过先使用DBN来对包含多个隐层的多层感知机进行预训练，然后通过反向传播算法来进行微调(fine-tuning)，能够提供一种解决深层网络优化过程中过拟合和梯度消失问题的有效途径。

通常对DNN等深度模型的训练需要具有强大计算能力的设备，而近年来以通用图形处理器(graphics processing unit，GPU)为代表的硬件技术的迅猛发展，有力支撑了深度学习理论与方法的高效实现。

最早将深度神经网络方法成功应用到语音识别中的研究机构是多伦多大学与微软研究院。他们使用DNN代替传统的GMM-HMM系统中的高斯混合模型，以音素状态为建模单位，提出了DNN-HMM的识别方法，显著降低了误识率，从而引发了基于深度神经网络的语音识别热潮。此后，随着深度学习技术的发展，卷积神经网络(convolutional neural networks，CNN)和循环神经网络(recurrent neural networks，RNN)等网络结构成功地应用到语音识别任务中。它们与传统的DNN方法相比展现出了各自的优势，受到越来越广泛的关注。目前，能够彻底摆脱HMM框架的端到端语音识别技术正日益成为语音识别研

究的焦点,无论是学术机构,还是工业界都投入大量的人力和财力,致力于此方面的研究。

近年来,语音识别研究工作更趋于解决在真实环境应用时所面临的实际问题,这可从作为国际语音识别研究热点风向标的 NIST(national institute of standards and technology)评测情况反映出来:其评测的语音类型已从最初的朗读语音到广播语音,再到后来的交谈式电话语音(conversational telephone speech),发展到目前真实场景的会议语音。相对于广播语音,交谈式电话语音增加了相应的难度,具体表现在:发音多为自发的口语语音,存在着大量的不流利(如犹豫词、重复、更正等)现象,同时,语音内容和词汇的随机性明显增加。此外,针对实际的电话线路,噪声的影响较大。2002 年,美国国防部先进研究项目局(Defense Advanced Research Projects Agency,DARPA)提出了一个"EARS-Effective, Affordable and Reusable Speech-to-text(高效低耗可重用语音文字转化)"的项目,把 NIST 的语音评测推到了又一个新的时代——丰富的语音文本(rich transcription,RT)转写,其要求不仅将语音所对应的文字显示出来,而且要将语音中的其他丰富信息,如文字之间的标点符号、句词之间的停顿、说话人等也能同时识别出来。从 2004 年的评测结果看,对广播语音和电话语音的词错误率(word error rates,WERs) 已分别下降到 10% 和 15% 以下。从 2005 年起,NIST 评测的语音类型转变为英语会议语音,包括磋商式会议(conference meeting)和演讲式会议(lecture meeting),其特点是研究真实会议场景中多人多方对话时的口语语音识别。相对于交谈式电话语音,会议语音又增加了相应的难度,表现在:必须解决会议场景中处于不同位置上说话人语音数据的有效采集问题,以及在多人交谈相互语音有少部分交叠时各自语音的分离问题。为此,NIST 评测中开始提供采用远离用户,且处于空间上多个位置、摆放形式多样的多麦克风或麦克风阵列采集来的现场数据作为评测的语料。从 2007 年进行的评测结果看,会议语音的词错误率在 40%～50%之间。2009 年的评测内容基本与 2007 年相同,所不同的是仅进行磋商式会议语音的评测,同时为各个测试任务定义了视频和音视频的输入条件。

目前无论从 NIST 评测的内容看,还是欧美发达国家的关注点看,研究真实场景中多人多方对话时的口语语音识别是当前语音识别的研究热点之一。从处理口语语音与朗读语音的方法看,其不同之处在于声学模型的自适应(acoustic adaptation)和发音词典自适应(lexicon adaptation)方面。声学模型自适应常采用基于最大似然线性回归(maximum likelihood linear regression,MLLR)和最大后验概率(maximum a posteriori,MAP)的方法。这两种方法是当前最为有效的自适应方法,许多新的自适应方法都是从二者中派生出来的。发音词典自适应常采用发音变化建模(pronunciation variation modeling)相关技术,主要研究由说话方式、语速、口音等带来的影响。

口语语音识别的另一个挑战是缺乏建立在大量口语文本语料之上良好的语言模型。朗读语音识别器所使用的统计语言模型,实际上都要依赖于大规模的训练语料,但是同样量级的口语语言的文字脚本还难以实现。口语语音中的不连贯进一步增加了语言模型估计的难度。目前研究者正致力于多种口语语言模型的建模方法研究。

当前语音识别研究的另一个趋势是,不再只单纯地关注大词表连续语音识别的精度,而是从实际的应用角度出发,积极探索机器对人类的语音进行感知与理解的途径和方法。而从整个计算领域的发展趋势看,近年的研究热点之一是普适计算,计算的模式与物理位置也正从传统的桌面方式逐步向以嵌入式处理为特征的无处不在的方式发展,比较典型的是移

动计算方式。因此对语音处理而言，探讨在典型的移动方式下的语音感知与理解机制，实现能根据用户的语音内容及所处的音频场景，并借助其他辅助信息（如地理位置、时间等）自主地感知和理解用户的意图及情感倾向，从而提供更智能化、人性化的人机交互手段，具有重要的理论意义与现实意义。同时，随着网络技术和移动计算技术的迅速发展，出现了网络环境下的语音识别技术、嵌入式和计算资源有限时的语音识别技术、语种识别技术、基于语音的情感处理技术等一些新的研究方向。

在国内，20 世纪 50 年代末就有人尝试用电子管电路进行元音识别，而到了 70 年代才由中科院声学所开始了计算机语音识别的研究。在此之后，有关专家也开始撰文介绍这方面的工作。从 20 世纪 80 年代开始，很多单位陆续参加到这一行列中来，它们纷纷采用不同的方法，开展了从最初的特定说话人中、小词汇量孤立词识别，到非特定说话人大词汇量连续语音识别的研究工作。20 世纪 80 年代末，以汉语全音节识别作为主攻方向的研究已经取得了相当大的进展，一些汉语语音输入系统已向实用化迈进。四达技术开发中心、星河公司等相继推出了相应的实际产品。清华大学、中科院声学所在无限词汇的汉语听写机的研制上获得成功。20 世纪 90 年代初，四达技术开发中心又与哈尔滨工业大学合作推出了具有自然语言理解能力的新产品。在国家“863 计划”支持下，清华大学和中科院自动化所等单位在汉语听写机原理样机的研制方面开展了卓有成效的工作。北京大学在说话人识别方面也做了大量的工作。

近年来，随着改革开放的不断进行，我国的国际地位与日俱增，汉语语音识别越来越受到重视，国外很多著名的公司都在国内设立了研发机构，并且都将汉语语音识别作为主攻方向之一。IBM 公司于 1997 年推出了汉语连续语音识别系统 ViaVoice，输入速度平均每分钟可达 150 字，平均最高识别率达到 95%，并具有“自我”学习的功能。2000 年发布的 ViaVoice 千禧版，用户可以通过语音导航到计算机桌面及浏览网页。1998 年，微软（Microsoft）投资 8000 万美元在中国筹建微软中国研究院（2000 年更名为微软亚洲研究院），开发的重点方向之一就是语音识别。1998 年，Intel 提出了基于 Intel 架构发展语音技术的构想，向软件开发厂商提供包括信号处理库、识别库、图像处理库在内的高性能语音函数库支持。1999 年，Intel 和 L&H 公司合作，推出语音识别软件开发包 Spark3.0，其中包括 Spark 语音识别引擎和软件开发工具箱。微软也推出了基于.net 的语音识别引擎。2011 年苹果公司在其 iphone 手机上率先推出了智能语音助理 siri，掀起了语音应用的热潮。国内一些著名企业也投入大量资金开始资助语音识别方面的研究，如百度、科大讯飞、阿里巴巴等。

尽管语音识别技术研究已经取得了很大的成绩，但到目前为止离广泛的应用尚存在距离。很多因素影响着语音识别系统的性能，如实际复杂环境中的背景噪声、传输通道的频率特性、说话人生理或心理情况的变化，以及应用领域的变化等都会导致语音识别系统性能的下降，甚至不能工作。研究语音识别系统顽健性（robustness）问题受到了研究者的广泛重视，国内外很多单位都开展了大量的工作。但到目前为止，所做的工作大都是针对某一种或两种影响因素进行补偿的研究，综合考虑各种影响因素补偿方法的研究还相对偏少。

语音识别通常是指能识别出相应的语音内容，除此之外，它还有一种特殊的形式——说话人识别。说话人识别不必识别出语音信号的具体内容，而只要鉴别出该语音是哪个说话人发出的即可。从实现的技术手段上看，说话人识别和语音识别一样，都是通过提取语音信

号的特征,并建立相应的参考模板来进行分类判断。说话人识别问题,最初是在第二次世界大战期间,美国国防部向贝尔实验室提出的课题。目的是根据窃听到的电话语音来判断说话人是哪一位德军高级将领,这对分析当时的德军战略部署具有重要的意义。该项目持续进行了三年,但并未达到预期的目的。

说话人识别研究的早期工作,主要集中在人耳听辨实验和探讨听音识别的可能性方面。随着语音识别研究的不断深入,说话人识别研究也获得了突飞猛进的发展。语音识别中很多成功的技术,如矢量量化(vector quantization,VQ)、隐马尔科夫模型等都被应用到说话人识别中。

20 世纪 90 年代,Rose 等提出了单状态的 HMM,即后来的高斯混合模型(gaussian mixture model,GMM),它是一个顽健的参数化模型。Matsui 等比较了基于连续 HMM 的说话人识别方法,发现识别率是状态和混合数的函数。同时,识别率与总的混合数有很强的关联性,但与状态数无关。这意味着不同状态间的转移信息对文本无关的说话人系统而言是没有作用的,因此,高斯混合模型 GMM 得到了与多状态 HMM 几乎相同的识别性能。正是上述工作,使得 GMM 建模方法在说话人识别研究中得到了越来越多的重视。特别是 Reynolds 等对高斯混合模型 GMM 以及通用背景模型(universal background model,UBM)的详尽介绍后,由于 GMM-UBM 具有简单有效,以及具有较好的顽健性等特点,迅速成为当今与文本无关的说话人识别中的主流技术,并由此将说话人识别技术带入了一个新的阶段。20 世纪 90 年代另一项重要的研究工作是,针对说话人确认中,说话人自身的似然度的得分变异的规整技术,出现了很多关于得分规整的算法,比较典型的如基于似然比(likelihood ratio)和后验概率(a posteriori probability)的技术。为了降低计算规整算法的计算复杂性,相继出现了群组说话人(cohort speakers)等方法。与此同时,说话人识别技术与其他的语音研究方向的结合更加密切,比如针对对话/会议中包含多人的说话人分割与聚类技术,音频元数据(metadata)的检索研究等也得到了很多研究人员的关注。

2000 年以来,各种新的说话人识别技术层出不穷,如支持向量机和 GMM 的结合,出现了一系列说话人得分规整的新方法,包括 Z-norm、H-norm、T-norm、Ht-norm、C-norm、D-norm 和 AT-norm。此外,针对信道失配问题,研究者们提出说话人模型合成方法。近年来,又提出了联合因子分析(Joint Factor Analysis),通过将说话人所在的空间划分为说话人空间和信道空间,进而能提取出与说话人相关的特征,并去掉与信道相关的特征。在此基础上,为了压缩说话人特征的规模,研究者又采用一个总变化空间来代替上述两个空间,从而提出了基于 i-vector 特征的方法。由于 i-vector 方法中只使用一个总变化空间来提取特征,因此所提取出的特征中可能同时包含说话人和信道的影响,需要对其进行进一步的信道补偿。通常是采用线性判别分析(linear discriminant analysis,LDA)来去除信道的影响。

目前,说话人识别的重点已经从实验系统转移到研究针对实际应用面临的问题。NIST 从 1996 年起开始举办每年一度的说话人识别评测(speaker recognition evaluations,SRE)。从其评测内容、评测方式的演变看,正逐步贴近实际的应用情况。例如,麦克风的种类越来越多,语种从单纯的英语,扩展到十几种语言,场景也从简单的单个说话人方式扩展到多个说话人方式。应该指出的是,近些年在 NIST 举办的说话人测试大赛中,识别率最高的单系统是基于 i-vector 的系统。除了 NIST 说话人评测之外,其他机构也组织过类似的评测,比

如荷兰 NFI-TNO(Netherlands forensic institute-TNO human factors)组织的说话人评测，主要针对司法应用方面的说话人识别。中文口语处理会议也在 2006 年组织了不同任务单元的说话人评测。虽然以上两个评测的规模和影响力不如 NIST 评测，但是都针对具体的应用语音环境，通过会议交流的方式，开放式的进行算法的优势对比和分析，不同程度地促进了技术的提高和进步。

目前，国外已经有了一些成熟的产品。如 AT&T 应用说话人识别技术研制出了智慧卡，已应用于自动提款机。欧洲电信联盟在电信与金融结合领域应用说话人识别技术，于 1998 年完成了 CAVE 计划，在电信网上进行说话人识别。说话人识别技术应用最为成功的例子是在伊拉克战争期间，萨达姆在电视上发表讲话后，美国 FBI 宣称讲话者不是萨达姆本人，而德国的科学家应用说话人识别技术证实讲话的人确实是萨达姆。从后来的情况看，德国科学家的判断是正确的。随着 Internet 的发展，网络环境下的说话人识别技术日益受到了重视，已成为当今的一个研究热点。

就语音合成技术而言，最早的语音合成器是 1835 年由 W. von Kempelen 发明，经威斯顿改进的机械式的会讲话的机器。它完全模拟人的发音生理过程，用风箱模拟来自肺部的空气动力，气流通过特别设计的哨时会产生语音中的辅音；气流通过形状可以变化的模拟口腔的软管时会产生元音。风箱、哨和软管三部分机械配合起来就可以产生一些音节和词。这是一个相当完善的机械式语音合成器。最早的电子式语音合成器是前面提到的 1939 年 Dudley 发明的声码器，它不是机械地模仿人发音的生理过程，而是通过电子线路来实现基于语音产生的源/滤波器理论；其中声源包括产生清音的噪声源和产生浊音的周期脉冲声源，它们分别用噪声发生器和张弛振荡器来实现，而声道的滤波作用是通过电子通道滤波器来实现的，滤波器的中心频率是用键盘上的十个琴键来控制。

现代的语音合成器都是利用计算机来实现的。从 20 世纪 70 年代末开始，出现了文-语转换(text to speech，TTS)系统的研究，其特点是用最基本的语音单元，如音素、双音素、半音节或音节作为合成单元，建立语音库，通过合成单元拼接而达到无限词汇的合成。为了保证合成声音具有良好的音质，在这种系统中除语音库外，还有一个相当庞大的规则库，以实现对合成语音的音段特征和超音段特征的控制。20 世纪 80 年代，由 D. Klatt 设计的串/并联混合型共振峰合成器是 20 世纪最有代表性的工作。它可以设置和控制多达八个共振峰，可模拟发音过程中的声道共振，而且还设有单独的滤波器来模拟鼻腔和气管的共振。其中，元音和浊辅音的产生用串联通道来实现，清辅音的产生用并联通道来实现。此外，这种合成器还可以对声源做各种选择和调整，以模拟不同的嗓音。它共可以产生七种不同音色的语音，包括模拟不同年龄、性别和个性的说话人的语音。瑞典皇家理工学院 Fant 实验室在多语种文-语转换系统研究方面也做出了突出的成绩，完成了英语、法语、瑞典语、西班牙语和芬兰语的文-语转换系统。

20 世纪 90 年代末，日本的研究者提出了一种多样本、不等长语音拼接合成技术 PSOLA。它在语音库中存放了大量的真人语音样本，通过选择合适的拼接语音片段来实现高质量的合成语音。在这项技术中，语音合成问题被简化为如何建立一个在语音学上充分覆盖的语音库，如何从语音库中选择合适的语音片段来拼接，以及如何对语音片段之间的拼接部分做适当的调整。

20 世纪 90 年代中期，随着语音识别中统计建模方法的日益成熟，研究者提出了可训练