出版基金
PUBLISHING FUND

The new generation of artificial intelligence and voice recognition

新一代人工智能与语音识别

马延周 著
Ma Yanzhou

清华大学出版社

The new generation of artificial intelligence and voice recognition

新一代人工智能与语音识别

马延周 著
Ma Yanzhou

清华大学出版社
北京

内 容 简 介

有关俄语语音识别的研究在中国尚处于起步阶段，此技术在中俄两国的民间交流和军事交往中发挥着重要作用。本书充分利用了新一代人工智能技术的研究成果，介绍了基于新闻语料的俄语连续语音识别技术。本书的目标是建立基于 Kaldi 环境设计并实现的俄语连续语音识别原型系统，使其同时具备在线识别功能和离线识别功能，以验证声学模型和语言模型的优化算法的有效性，进而为面向特定领域的俄语语音识别实用系统的研发提供理论方法、实验数据和关键技术支撑。为了实现上述目标，本书详细介绍了俄语语音语料的采集、加工、处理，俄语文本语料的采集、清洗、过滤，俄语发音词典的自动预测、生成，声学模型建模基本单元（音素集）的确定，声学模型和语言模型的优化。

本书可作为高等院校外国语言学及应用语言学专业、电子信息和通信类专业本科生及研究生的教学参考书，也可供语音信息处理与应用开发等领域的研究人员使用。

图书在版编目（CIP）数据

新一代人工智能与语音识别/马延周著. —北京：清华大学出版社，2019
ISBN 978-7-302-52384-0

Ⅰ. ①新… Ⅱ. ①马… Ⅲ. ①人工智能－应用－俄语－新闻语言－研究 ②语音识别－应用－俄语－新闻语言－研究 Ⅳ. ①TP18 ②TN912.34 ③G210

中国版本图书馆 CIP 数据核字（2019）第 038778 号

责任编辑：郭　赛
封面设计：何凤霞
责任校对：梁　毅
责任印制：宋　林

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：北京国马印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10　　**字　　数**：238 千字
版　　次：2019 年 8 月第 1 版　　**印　　次**：2019 年 8 月第 1 次印刷
定　　价：44.50 元

产品编号：082543-01

序

自动语音识别(Automatic Speech Recognition,ASR)是自然语言处理(Natural Language Processing,NLP)的一个重要领域。

世界上第一台能够自动识别语音的机器当属一种名为 Radio Rex 的玩具。这种玩具出现于 20 世纪 20 年代。Radio Rex 是一个用赛璐璐材料制作成的玩具狗,它受到一根弹簧的控制,弹簧在 500Hz 的声音频率下会释放,弹簧一旦释放,玩具狗就会动起来。由于 500Hz 的频率粗略等于单词 Rex 中元音的第一个共振峰的频率,因此当人们说出 Rex 的时候,这只叫作 Radio Rex 的玩具狗就会在人们的呼唤声中自动走过来。

20 世纪 40 年代末至 50 年代初,美国建立了一系列机器语音识别系统。早期,美国贝尔实验室中的系统可以识别一个单独说话人讲出的 10 个数字中的任何一个,这个系统存储了不依赖于说话人的 10 个模式,每个数字各有一个模式,每个模式都代表每个数字中的前两个元音的共振峰,研究人员通过选择与输入语音存在最高相关系数的方法使数字的语音识别正确率达到了 97%~99%。

英国伦敦大学的 Fry 和 Denes 建立了一个音位识别系统,根据模式识别原则,该系统能够识别英语中的 4 个元音和 9 个辅音。Fry 和 Denes 研发的系统首次使用了音位转移概率对语音识别系统进行约束。

20 世纪 60 年代末至 70 年代初出现了许多重要的创新性研究成果。

首先,出现了一系列特征抽取算法,包括高效的快速傅里叶变换(Fast Fourier Transform,FFT)、倒谱(cepstrum)处理在语音中的应用以及语音编码中的线性预测编码(Linear Predictive Coding,LPC)的研制。

其次,提出了一些处理翘曲变形(warping)的方法,当与存储模式匹配时,通过展宽和收缩输入信号的方法处理说话速率和切分长度的差异。解决这些问题的最自然的方法是动态规划(dynamic programming)。在研究这

个问题的时候，同样的算法被多次重新提出。最早把动态规划应用于语音处理技术的人是 Vintsyk，尽管他的成果没有被其他研究人员提及，但是后来有很多研究者都再次重复了他的发明。随后，Itakura 把这种动态规划的思想和 LPC 系数相结合，并首次在语音编码中使用，他建立的系统可以抽取输入单词中的 LPC 特征，并使用动态规划的方法把这些特征与存储的 LPC 模板相匹配。这种动态规划方法的非概率应用是对输入语音进行模板匹配，称为动态时间翘曲变形(dynamic time warping)。

最后是隐马尔可夫模型(Hidden Markov Model，HMM)的兴起。1972 年前后，美国的研究人员分别在两个实验室独立应用 HMM 研究语音问题。其中一部分的应用是由一些统计学领域的工作引起的，Baum 和他的同事在普林斯顿国防分析研究所研究 HMM，并把它应用于各种预测问题的解决。James Baker 在于卡内基-梅隆大学(Carnegie-Mellon University，CMU)攻读硕士期间研究了 Baum 等人的工作内容，并把他们的算法应用于语音处理。同时，在 IBM 公司的 Thomas J. Watson 研究中心，Frederick Jelinek、Robert Mercer、Lalit Bahl 独立把 HMM 应用于语音研究，他们在信息模型方面的研究受到了 Shannon 的影响。IBM 的系统和 Baker 的系统非常相似，都使用了贝叶斯(Bayes)算法，不同之处是早期的解码算法。Baker 的 DRAGON 系统使用了维特比(Viterbi)动态规划解码，而 IBM 系统则应用了 Jelinek 的栈解码算法。Baker 在建立 DRAGON 系统之前曾经短期参加过 IBM 小组的工作。IBM 的语音识别方法在 20 世纪末期完全主导了语音识别领域，IBM 实验室是把统计模型应用于自然语言处理的推动力量，他们研制了基于类别的多元语法模型，研制了基于 HMM 的词类标注系统，研制了统计机器翻译系统，他们还使用熵和困惑度作为评测系统的度量指标。

HMM 逐渐在语音处理界流传开来，原因之一是美国国防部(U. S. Department of Defense)高级研究计划署(Advanced Research Projects Agency，ARPA)发起了一系列相关研究和开发计划。第一个“五年计划”始于 1971 年，目标是建立基于少数说话人的语音理解系统。这个系统使用了一个约束性语法和一个词表(包括 1000 个单词)，要求语义错误率低于 10%。ARPA 资助了四个系统，并且对它们进行了比较，这四个系统是：系统开发公司的系统(System Development Corporation，SDC)、Bolt，Beranek & Newman (BBN)的 HWIM 系统、卡内基-梅隆大学的 Hearsay-Ⅱ系统和 Harpy 系统。其中，Harpy 系统使用了 Baker 基于 HMM 的 DRAGON 系统的简化版本，在评测系统时得到了最佳成绩。对于一般任务，Harpy 系统的语义正确率达到了 94%，是唯一一个达到了 ARPA 计划目标的系统。

自20世纪80年代中期开始，ARPA陆续资助了一些新的语音研究计划。第一个计划的任务是资源管理(Resource Management，RM)，与ARPA早期的课题类似，其主要进行阅读语音(说话人阅读的句子的词汇量包含1000个单词)的转写(即语音识别)，但这个系统还包括一个不依赖于说话人的语音识别装置。该计划的另一个任务是建立《华尔街杂志》(*Wall Street Journal*)的句子阅读识别系统，该系统的初始词汇量被限制在5000个单词以内，到最后，系统已经没有了词汇量的限制。事实上，大多数系统的词汇量都已经有了约6万个单词。后来的语音识别系统能够识别的语音已经不再是简单的阅读语音了，而是更加自然的语音。其中，广播新闻识别系统可以转写广播新闻，甚至转写那些非常复杂的新闻，如现场采访；还有CallHome系统、CallFriend系统和Fisher系统，它们可以识别人们在电话交流中的自然对话。空中交通信息系统(Air Traffic Information System，ATIS)属于语音理解领域的课题之一，该系统可以帮助用户预订飞机票，回答用户关于航班、飞行时间、日期等方面的问题。

ARPA计划大约每年进行一次汇报，参加汇报的除了有ARPA资助的课题以外，还有来自北美和欧洲的其他“志愿者”系统，汇报时将分别测试各个系统的单词错误率和语义错误率。在早期测试中，营利型公司一般不参加比赛，但是随着时间的推移，很多公司开始积极参赛(特别是IBM公司和AT&T公司)。ARPA的比赛促进了各个实验室之间的借鉴和交流，因为在比赛中可以很容易地看出大家过去一年的研究进展和成果，这成为了HMM模型能够传播到每一个语音识别实验室的重要因素。ARPA的计划也造就了很多有用的数据库，这些数据库原来都是为了评估而设计的训练系统和测试系统(如TIMIT、RM、WSJ、ATIS、BN、CallHome、Switchboard、Fisher)，但是后来却都在其他总体性研究中得到了应用。

中国在语音自动处理领域也取得了很不错的成绩。于1999年6月9日成立的安徽科大讯飞信息科技股份有限公司(简称科大讯飞)是一家专门从事智能语音及语音技术研究、软件及芯片产品开发、语音信息服务的国家级骨干软件企业。科大讯飞推出的产品包括大型电信级的应用到小型嵌入式的应用，电信、金融等行业到企业和家庭用户，PC到手机再到MP3、MP4、PMP和玩具，能够满足不同的应用环境。科大讯飞占有中文语音技术市场60%以上的市场份额，以科大讯飞为核心的中文语音产业链已经初具规模。

由以上介绍不难看出，自动语音识别是一个交叉学科，需要具备语言学、计算机科学、声学等领域的知识。

本书作者马延周不惧困难，他努力进行知识更新后的再学习，根据俄语语音的特

点优化了声学层的 HMM 模型，采用较好的算法解决了训练数据不足和训练速度慢的问题；他还在具有较强背景噪声和多个说话人的环境下采用了降噪技术，增强了俄语语音识别的健壮性；此外，他还利用了各种能够辅助俄语语音识别的语言信息，除了俄语语音的频谱特征参数、能量参数、韵律参数以外，他还综合利用了俄语构词规则、变格变位规则、句法表现形式以及语义辨析和语境条件，有效地降低了俄语语音识别的错误率。

在研究过程中，作者建立了基于众包的俄语语音标注平台和语音语料库，设计了面向俄语新闻网页文本数据过滤清洗系统的俄语文本语料库，为俄语连续语音识别系统的研究开辟了新途径。作者还构建了一个具有一定规模的俄语发音词典，可以将俄语文本转写为相应的俄语标准发音，并对俄语语音识别中的音素集和字音转换规则进行了优化，降低了声学模型的训练难度，提高了模型的训练效果。最后，作者设计并实现的俄语连续语音识别原型系统同时具有在线识别功能和离线识别功能，这在一定程度上填补了中国俄语语音识别研究领域的空白。

本书详细阐述了作者的创新性研究，值得我们认真学习，是为序。

冯志伟

2019 年 6 月 5 日

前言

随着人工智能、计算技术和信号处理技术的飞速发展，以及自然语言与计算机网络的结合，语言的功能已由人际交流延伸至人机交流和机机交流，而实现这一目标的重要前提是计算机能够听懂并识别和理解人类的语言。当前，作为人机交互的关键技术，语音信息智能处理已成为网络空间环境下世界各国研究者广泛关注的热点问题之一。尤其是随着新媒体的出现和大数据的兴起，人们迫切需要对具有多通道、多来源、多语言特征的海量语音信息技术进行深化研究与创新突破，此项技术的战略意义和安全价值日渐突显。

近年来，国内外众多科研院所和企业都对英文和中文语音识别进行了深入的探索和研究，开发了一系列实用化系统，但是在俄语语音识别领域，尤其是对连续语音识别的研究还相对薄弱。本书通过考察分析国内外语音识别技术的研究现状及存在的难题，重点研究俄语连续语音识别的基本原理和关键技术，尝试采用深度神经网络(DNN)的声学模型优化训练方法，设计俄语连续语音识别原型系统。

本书试图解决以下三个问题：

(1) 俄语新闻语音语料和文本语料的采集、过滤、清洗、标注及建库方法；

(2) 建立基于 DNN 的声学模型和基于 SRILM 的语言模型，分析两类模型的训练算法优化和训练结果，并通过对比预测生成适用于语音识别的俄语发音词典；

(3) 设计与实现兼具在线和离线识别功能的俄语连续语音识别原型系统，并对原型系统的性能进行测试验证。

本书取得的主要成果如下：

(1) 在俄语声学模型训练过程中设计了基于众包的语音标注平台，建立

了 360 小时的俄语新闻标注语音语料库，形成俄语语音识别音素集，采用 DNN 的优化训练方法生成了大小为 59.7MB 的声学模型；

（2）在俄语语言模型训练过程中设计了俄语新闻文本语料过滤清洗系统，建立了 10GB 规模的纯净可训练俄语文本语料库，采用 SRILM 的优化训练方法生成了大小为 1.21GB 的四元剪枝语言模型；

（3）通过数据驱动的方法预测生成包含 76277 个词形的俄语发音词典，利用该词典的数据资源，并基于 Kaldi 进行二次开发，实现了具有在线识别和离线识别功能的俄语连续语音识别原型系统，可以为面向特定领域的俄语语音识别实用系统的研发提供基础理论和关键技术支撑。

马延周

2019 年 7 月

图书资源支持

感谢您一直以来对清华版图书的支持和爱护。为了配合本书的使用，本书提供配套的资源，有需求的读者请扫描下方的“书圈”微信公众号二维码，在图书专区下载，也可以拨打电话或发送电子邮件咨询。

如果您在使用本书的过程中遇到了什么问题，或者有相关图书出版计划，也请您发邮件告诉我们，以便我们更好地为您服务。

我们的联系方式：

地　　址：北京市海淀区双清路学研大厦 A 座 701

邮　　编：100084

电　　话：010－62770175－4608

资源下载：http://www.tup.com.cn

客服邮箱：tupjsj@vip.163.com

QQ：2301891038（请写明您的单位和姓名）

资源下载、样书申请

书圈

扫一扫，获取最新目录

用微信扫一扫右边的二维码，即可关注清华大学出版社公众号“书圈”。

目 录

第0章 绪 论

0.1 研究依据

在信息化社会中，以语言信息处理为核心的语言技术已成为当代科技创新的重要基础、动力和源泉。信息技术为人类创造了一个新的虚拟世界，改变了人类的生存方式和生活方式。利用语音技术而开发的智能手机、语音云驾驶系统、语音智能搜索引擎等智能化互动产品，为人们的日常生活和社会交往带来了极大便利。

近年来，高性能计算、信号处理、模式识别及声学技术发展迅速，针对不同应用需求而研究开发语音识别系统已成为可能，因此语音识别技术在工业生产、交通运输、国防安全等众多领域得到了广泛的推广和应用。目前，语音识别所涉及的语种得以扩展。就俄语语音识别而言，对大词汇量、非特定人、连续语音识别的研究仍然面临着许多困难，与人们预期的目标还有较大距离。俄语连续语音识别面临的主要难题有：①在声学层面，俄语的重音变化及自由重音现象难以处理；②俄语语音识别系统的适应性较弱，随着语言交际环境的变化，系统的性能会严重下降；③噪声环境和传输设备会直接影响俄语语音特征的提取，如何排除环境噪声的干扰以提升识别性能也是一大难题；④因发音人不同或随发音人的生理及心理状态的变化，俄语发音特征会产生很大的差异性；⑤在俄语连续语流中，语音的基本单元（如音素、词形等）之间存在协同发音，由于边界模糊而导致难以进行精确的语音分割。

语音信号的端点检测方法是判定语音识别准确率的重要手段，即使在纯净环境下，语音识别系统50%的错误识别均来自端点检测。因此，俄语大词汇量连续语音识别系统的开发必须解决上述难题，才能在一定程度上提高识别的速度和准确率。

鉴于俄语连续语音识别研究中存在的诸多难题，本书集中研究以下三个主要方面：①优化声学层模型，合理利用俄语语音学和计算语音学知识，改进声学模型结构，

采用更好的算法以解决训练数据不足和训练速度慢的问题；②增强俄语语音识别的健壮性，在具有较强背景噪声或多说话人参与的环境下采用降噪技术，进而增强俄语语音识别系统的适应性；③充分利用一切能够辅助俄语语音识别的语言信息。除俄语语音的频谱特征参数、能量参数、韵律参数之外，还要综合利用俄语构词及词变规则、句法表现形式甚至语义辨析和语境条件，从而有效降低语音识别的错误率。

0.2 研究对象与研究目标

本书的研究对象是基于标注新闻的俄语大词汇量连续语音识别的基本原理和关键技术，主要包括以下几点。

1. 俄语语音语料库和文本语料库的构建

大规模语音语料库和文本语料库是语音识别系统的重要基础性资源，实证语料数据的规模与加工质量直接影响着俄语声学模型与语言模型训练的效果。目前，国内外已有一些 IT 企业和研究机构（如 ELDA、LDC、海天瑞声）能够提供大量语音和文本数据库资源，可用于本研究的俄语声学模型和语言模型的构建与训练。

2. 俄语声学建模的基本识别单元的选定

基于计算语音学的理论方法构建俄语声学模型，其目的在于利用高效的算法计算俄语语音的多维特征矢量序列和每一个发音模板之间的距离。充分利用俄语语言学及语音学的知识，设计基于 HMM 的俄语音素模型，提取声学基元，利用有效的相关算法训练 HMM 模型，这对于扩大声学模型的训练数据规模、增强识别系统的准确率和灵活性均具有重要作用。

3. 俄语语言模型中数据稀疏问题的求解

俄语新闻文本语料库的覆盖度不全面，可能导致一些语言现象无法统计，进而导致在已建立的语言模型中检索不到与该模型对应的某些语言现象，即概率为零且无法识别，因此造成语言模型的数据稀疏问题。鉴于此，需要尽可能全面地采集并加工处理俄语新闻文本语料，为俄语语言模型的有效训练提供覆盖面更大的实证数据支撑。

本书的研究目标包括：基于 Kaldi 设计实现俄语连续语音识别原型系统，使之具备在线识别和离线识别功能，以验证声学模型和语言模型优化算法的有效性，进而为面向特定领域的俄语语音识别实用系统的研发提供理论方法、实验数据和关键技术支