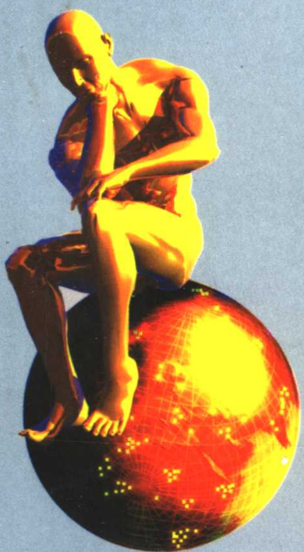
21世纪高等院校优秀教材

实用语音识别基础



王炳锡 屈丹 彭煊 著

国防工业出版社
http://www.ndip.cn

21世纪高等院校优秀教材

实用语音识别基础

Practical Fundamentals of Speech Recognition

王炳锡 屈丹 彭煊
侯风雷 王波 徐望 王炜 马占武 著

国防工业出版社
·北京·

图书在版编目(CIP)数据

实用语音识别基础/王炳锡等著．—北京:国防工业出版社,2005.1
21世纪高等院校优秀教材
ISBN 7-118-03746-X

Ⅰ.实... Ⅱ.王... Ⅲ.语音识别—高等学校—教材 Ⅳ.H012

中国版本图书馆CIP数据核字(2004)第127395号

国防工業出版社出版发行
(北京市海淀区紫竹院南路23号)
(邮政编码 100044)
新艺印刷厂印刷
新华书店经售
*
开本787×1092 1/16 印张24 539千字
2005年1月第1版 2005年1月北京第1次印刷
印数:1—4000册 定价:38.00元

国防书店:(010)68428422 发行邮购:(010)68414474
发行传真:(010)68411535 发行业务:(010)68472764

序

计算机技术是二十世纪最伟大的发明，是当代发展最为迅猛的科学技术，它几乎渗透到人类社会活动的每一个领域；计算机网络的出现使人类的信息交流超越了时间和空间，知识能够共享，引发了经济结构和生活方式的深刻变革，极大地推动着人类社会的发展和进步。

计算机科学和众多科学交叉、融合、渗透，产生出许多新学科，推动着科学技术向前发展。

多媒体技术的兴起，成为计算机与人之间信息交流的关键技术，由此引发的多媒体信息处理领域的研究课题涉及模仿人类感官的信息采集和模仿人类智能的处理、理解、判断，因此不仅对生理器官要有深刻的了解，而且要对神经中枢的作用、心理作用做相应的研究，这是一个极具挑战性的研究领域。

语音识别是人机语音通信的关键技术之一，也是难题之一，经过广大科技工作者的不懈努力，已在不同层面上有突破性进展，并取得了可喜的成绩。

《实用语音识别基础》对语音识别的理论和关键技术作了回顾和总结，从实用出发，引入了现代非线性处理理论和方法，为把语音识别从实验室推向实际应用，提供了必要的解决思路和方案，同时也反映了学科前沿和发展趋势。我认为，本书的出版对语音识别的研究将起到推动作用，对信息处理的发展将做出有益的贡献。

中国工程院院士
2002年中国国家最高科技奖获得者　金怡濂

2004年5月

前　　言

人类有个理想，让机器具有"听"、"说"人类语言的能力。这个理想，在信息时代正逐步变成现实，童话般的神奇世界正慢慢地向我们走来。语音识别正是解决机器"听"懂人类语言的一项研究。

新世纪伊始，信息革命如火如荼，语音识别也进入了全面发展时期，适时地回顾语音识别在理论和实践方面的发展历程，总结研究成果，理清未来发展思路，无疑对语音识别的研究是十分重要的。作者从事语音识别教学、科研二十多年，做了一些工作，积累了一些经验，有些想法和思路，希望与学术界同行交流。当然其中有偏颇之处，只是管窥之见，一家之言。

语音信号是非平稳的、时变的、复杂的、信息量大的信号，是语义信息加个人特征的混合信号。我们目前的语音识别是以语音信号为原始素材做三个层面的处理：物理(声学)层面，语言学层面(自然语言理解)，大脑神经(智能)。物理层面的研究较为深入，在理论和实践上有较多的积累；结合语言学的自然语言理解也有相当的研究，但成果不甚显著；结合大脑神经的智能化研究则刚刚起步；而要使语音识别做到像人一样，我认为还应在心理层面开展工作，当前这方面的研究几乎是个空白。

不管是用于什么目的的语音识别，无外乎三个模块：参数描述、模型逼近、推理判决。人们总是希望描述语音某个特征的参数集合稳定，便于提取，但现在仍没找到这种对应关系，各种特征在参数集合中有交织，只是在交织的程度上表现不同。模型逼近具有理论基础，有代表性的是模板匹配、统计概率模型和判别模型，另外还有某些特殊用途的背景模型(反模型、废料模型)。由于对语音信号物理本质的认识的局限性，这些模型只能是逼近。推理判决从简单的最小失真距离判别、假设检验，发展到数据融合、证据理论的应用，应该说这些知识在不同程度上发挥了重要作用，但离实用还有很大的距离。我认为语音信号的物理本质是湍流，是一种复杂的混沌现象，而人耳感知语音的模型仍是个难题，目前只是简单地看作滤波器组对语音信号滤波，语音识别还应有大脑神经的加工处理，以及人的心理活动参与。有人在研究语音情感识别，还有人结合图像处理进行口型识别，我认为这些研究都会对语音识别起到辅助作用。

语音识别既是个理论问题，也是一个工程化的问题。它综合多学科的理论成果，如声学、语音学、语言学、生理学、数字信号处理、信息工程、通信理论、电子技术、计算机科学、模式识别、人工智能等，结合语音信号的特点，产生一系列语音识别的理论。而要实用还有一个工程化问题需要解决，语音识别的成果走出实验室，所面临的问题比语音识别本身还要多，还要复杂，还要难。首先遇到的是各种噪声干扰，其次是各种信道条件下的频谱

畸变,还有各种不同用户的不同需求,应用场合(如工厂、车间、马路、酒吧、歌厅等)的不同,诸如多人话音背景下的语音识别,音乐、广播等背景噪声下的语音识别,等等。

值得称道的是口呼电话拨号、口授打字技术的成功使用,取得了巨大的商业利益,给语音识别研究人员树立了信心。我们的研究也启示我们应立足于实用,把复杂繁乱的应用环境纳入实用语音识别的研究中。科学地简化问题、理性地处理应用环境是语音识别实用化的基础。

由此引发的课题:语音信号表意性稳健参数的研究及提取;语音信号个人特征稳健参数的研究及提取;口音自适应、信道自适应、送话器自适应、背景环境自适应;语音、语言、心理智能模型;多参数、多模式、多模型的融合、推理、判决等。语音识别系统的评测标准及方法研究应引起学术界的重视,它是语音识别系统走出实验室,投入实际使用,进而为某种需求研制专用系统,变成产品,投放市场的重要环节。我认为,评测标准和方法具有导向作用,它是从另一个角度推进语音识别研究的动力。《实用语音识别基础》就是在这样的思考中成书的,之所以称之为基础,它确实是些基本的理论和概念、基本的技术和方法;之所以称之为实用,它试图把语音识别推出实验室,按照实际应用来整合内容。在这个框架下,以介绍新的实用理论为主,尽可能注意在数学上严谨、在逻辑上严密,在工程化的介绍中,以工科大学生的理论基础为基点,以我们的研究思路和成果为主,注重可操作性。

本书的内容安排如下。本书共分 4 个部分 17 章。第 1 章简要介绍了语音识别的发展历程以及语音识别技术的研究现状和未来趋势。第 1 部分:基本理论(第 2 章 ~ 第 5 章),介绍了语音识别的基本理论。其中,第 2 章介绍了听觉机理和汉语语音基础;第 3 章 ~ 第 5 章详细讲解了语音信号的处理方法,包括时域处理、时频分析、倒谱同态处理。第 2 部分:语音识别系统(第 6 章 ~ 第 10 章),详细讲述了实用语音识别系统的建立过程,分别介绍了语料库的建立原则、语音信号的预处理、特征提取、特征变换、识别模型。其中,第 10 章识别模型中着重讲解了 5 种常用模型:动态时间规整,隐马尔可夫模型,支持向量机,人工神经网络和高斯混合模型。第 3 部分:语音识别中关键处理技术(第 11 章 ~ 第 13 章),针对语音识别系统实用化的问题,给出了一些改善语音识别系统性能的关键技术。其中,第 11 章介绍了说话人自适应和说话人归一化技术;第 12 章给出了当前一些有效的噪声抑制的方法;第 13 章针对不同信道条件下的语音识别系统的不匹配问题,提出了信号补偿的方法。这些技术是语音识别走向实用的重要环节。第 4 部分:语音识别应用(第 14 章 ~ 第 17 章),介绍了语音识别系统的 4 个主要应用,即说话人识别、关键词识别、语种识别和连续语音识别。这 4 个部分中,每一部分内容可独立使用,在教学中可灵活安排。每章后面附有支持本章内容的参考文献,供读者深入研究之用,书后附有英汉名词对照,供读者查阅外文资料参考。第 1 章、第 6 章、第 7 章、第 11 章、第 12 章、第 16 章由屈丹编写,第 2 章、第 3 章、第 5 章、第 9 章、第 15 章由彭煊编写,第 8 章、第 10 章由王波、彭煊共同编写,第 4 章、第 13 章、第 14 章、第 17 章分别由马占武、王炜、侯风雷、徐望编写。全书内容由王炳锡统筹指导,由屈丹整理修改,最后由王炳锡审校定稿。

本书的特点是:讲解了目前最前沿的语音识别理论和技术,反映了语音识别技术的最

新技术与发展趋势，具有与本学科学术水平相适应的先进性；由浅入深地安排章节，先理论后应用，知识结构合理，章节之间紧密配合、前后呼应，具有很强的科学性和系统性；以自己的研究和实验成果为主，提供了大量的实际参数、图表，与实际工作联系紧密，具有很强的可操作性与实用性。通过使用本书，读者可以了解当前最前沿的语音识别技术，可以掌握语音识别的基本知识和系统理论，可以获得应用实用语音识别技术的基本技能。另外，通过我们在书中描述的启发式研究过程，读者既可以提高自学能力，又可以在创新思维方面得到很大的提高。

在撰写本书的过程中，得到了国内同行学者的支持和帮助，和著名专家进行过有益的交流、研讨，这些使作者受益匪浅。他们的卓越见解提升了本书的理论价值和可用性，在此向他们表示深切感谢。我的博士生、硕士生结合自己的研究课题展开的实用性的研究为本书打下了坚实的理论基础，他们卓有成效的工作，已经迈出了实用化的第一步，使本书言之有理、言之有物，在此向他们表示感谢。书中引用了大量的文献资料，是原作者的辛勤工作，是他们把语音识别推向前进，向他们表示感谢。最后感谢解放军信息工程大学、信息工程学院各级领导和同事们在教学和科研工作中有力的支持和帮助，使我有信心努力拼搏，战胜困难，顽强奋斗。

最后要特别感谢国家自然科学基金委员会对“电话信道自然语音的语言辨识技术研究”项目（批准号:60372038）的支持，诚恳地希望读者对书中不妥之处批评指正。

在新的世纪，语音识别一定会同相关学科共同发展，在信息科学的百花园里，绽放出奇异的光彩，本书作为一块“铺路石”，将“待到山花烂漫时，它在丛中笑”。

王炳锡

2004 年 4 月

于解放军信息工程大学

Preface

Man has long dreamed of having a machine that can "listen to" and "speak" human languages, which enables him to enjoy a fairy world of wonder. This ideal of man, in the information era, is gradually becoming a reality with the state-of-the-art technology in speech recognition, the task of which is to solve the problem of machine understanding the human speech.

At the beginning of the century, information technology develops by leaps and bounds, and speech recognition also steps into an all-round developing period. It is absolutely necessary to review the history of speech recognition in both theory and practice, to summarize the research achievements and to make clear the intended developing thought. The author has been teaching and researching in speech recognition for over twenty years and has obtained some achievements, experience, ideas and thoughts which, biased as they may be due to the author's own limited knowledge, he would like to share with other academic colleagues.

Speech signal is non-stationary, time-varying and complex with a large amount of information including semantic and personal alike. The present speech recognition is to process the original speech signal in three levels: physical (acoustic) level, linguistic level (natural language processing), cerebral nerve level (intelligence). Physical in-depth research has had much accumulation in both theory and practice. The achievements of natural language understanding combined with linguistics are not as notable. The intelligence integrating the human brains is just a start. Speech recognition comparable to man calls for more studies at the mental level, which is now almost a blank.

Speech recognition, whatever its application, involves three procedures: parameter description, model approximation and reasoning decision. It is expected that the parameter set describing speech signals is stable, and easily extracted, but so far the relationship between them is not found. Various parameters interlace in the parameter space, and the only difference lies in the degree of interlacing. The model approximations have the theoretical bases which are representative of model matching, statistical probability model and decision model with other additional background models for some applications (anti-model and filer). Due to the limited knowledge of the physical nature of speech signals, these models are only approximations. Reasoning decision develops from minimum distortion distance decisions, hypothesized test to data fusion and evidence theory. It should be said that these knowledge have played an important role in different degrees, but it is a long way before practical application. Speech signals are essentially turbulence in physical nature, a complex chaos phenomenon, but it is difficult to get human ear perception model of speeches. At present, it is only regarded as the filtering speech signals using filter banks. Speech recognition also involves the brain processing and human psychological actions. Emotion speech recognition and mouth shape recognition combining image processing is also studied in recent years, which plays an auxiliary role to speech recognition.

Not only is speech recognition a theoretical problem, but also an engineering problem. It integrates theoretic achievements of many disciplines, for example, acoustics, phonetics, linguistics, physiology, digital processing, information engineering, communication theories, electronic technology, computer science, pattern recognition and artificial intelligence. Integrated with the characteristics of speech signals, speech recognition brings forward a series of speech recognition theories. But in order to apply it to real environment, there are many engineering problems to be solved. When the achievements of speech recognition come out of the laboratory, they face more complex and difficult problems. The first problem is various kinds of noise interferences. The next is the spectral distortion in various channel conditions. In addition, different requirements of users and different application environments, such as in a factory or a workshop, on the street, in a bar, are new tasks in real environments. For example, speech recognition in the background of many speakers and speech recognition in the noise of music and radio speech are all the difficult tasks.

It is worth saying that the successful applications of oral telephone dialing and spoken typing have obtained great commercial benefits, which adds the confidence of researches in speech recognition. Our researches also enlighten us that we should be established in practicality and bring complex and multifarious application environments into the researches of the applied speech recognition. Scientific simplification of the problems and rational processing of the application environments are the basis of application of speech recognition.

So the evocable tasks include the study and extraction of robust semantic parameters of speech signals, the study and extraction of robust individual parameters of speech signals, accent adaptation, channel adaptation, telephone transmitter adaptation, environment adaptation, the intelligent model of speech, language and mentality, the fusion, reasoning and decision of multiple parameters, patterns and models etc.. The academe should attach importance to the evaluation standards and methods of speech recognition system, which will bridge the gap between laboratory research and real application in the form of task-specific systems and complete market products. I think the evaluation standards and methods guide the developments, which is another motivation of speech recognition researches. "Practical Fundamentals of Speech Recognition" is completed in such considerations. It is called fundamentals because it certainly includes some fundamental theories and concepts, technologies and methods. It is called practical speech recognition because it attempts to put the speech recognition out of laboratory and arranges the contents according to application. In such a framework, this book puts emphasis on introducing new practical theories, and more attention has been paid to mathematical preciseness and rigorous logic. In the engineering introduction, we briefly introduce our research ideas and achievements and pay attention to the maneuverability based on theoretical fundamentals of college students in engineering.

The contents of this book are arranged into four parts of seventeen chapters. Chapter 1 briefly introduces the development process, research status and intended trends of speech technologies. The first part (Chapters 2 - 5) presents fundamentals of speech recognition. (Chapter 2 gives the perceptual mechanism and the basis of the mandarin speech. Chapters 3 to 5 elaborately present speech processing methods including time processing, time-frequency analysis, cepstral homeostasis processing.) The second part (chapters 6 - 10), speech recognition system, introduces in detail the construction process of the practical speech recognition system comprising the foundation principles of speech corpus, pre-processing of speech signals, feature extraction, feature transformation and recognition models. (Chapter 10 puts emphasis on five common models: dynamic time warping, hidden markov model, support vector machine, artificial neural network and mixture Gaussian model.) In Part 3 (chapters 11 - 13), practical key processing technologies in speech recognition and key technologies to improve the performance of speech recognition systems are presented. (Chapter 11 gives the technologies of the speaker adaptation and speaker normalization. Chapter 12 gives the methods of noise suppression. In chapter 13, for non-matching of speech recognition systems in various environments, methods of signal compensation are provided. All the technologies in part 3 are the important factors of the application of speech recognition system.) Part 4, application of speech recognition (chapters 14 - 17), introduce four primary applications of speech recognition systems. They are speaker recognition, keyword spotting, language identification and continuous speech recognition. Each of the four parts can be used independently in teaching. Each chapter is attached with references for further study. An English-Chinese translation of special terminology is attached at the end of the book. Chapters 1, 6, 7, 11, 12, 16 are compiled by Qu Dan, chapters 2, 3, 5, 9, 15 by Peng Xuan, and chapters 10 and 12 by both of Wang Bo and Peng Xuan. The remaining chapters 4, 13, 14, 17 by Ma Zhanwu, Wang Wei, Hou Fenglei and Xu Wang respectively. Qu Dan revised all of the material. The whole book was planned, supervised and finally checked by Professor Wang Bingxi.

This book is intended to present most popular theories, state-of-the-art technologies, and prospective development trends. The chapters are arranged step by step from theory to application with regional knowledge structure, close relationship between chapters, and continuity between parts of the book. The book is based on our achievements and also provides a large amount of parameters, figures and tables, which are experience from practical projects having good maneuverability and practicality. Through this book, readers can learn the most popular recognition technologies, master the basic knowledge and systematic theories, and also obtain the basic skills of applying the speech recognition technologies. In addition, through the heuristic process described, the readers not only improve their capacities of self-study, but also improve their innovative thinking.

In the course of compiling the book, we have obtained supports and helps from the domestic academic fellows, and also have some beneficial communications and discussions with some famous experts, which have benefited the authors a lot. Their superexcellent opinions improve the theoretical values and practicality and great thanks are to them. My doctorial candidates, postgraduates have combined their own work to conduct practical studies, which forms a solid foundation of this book. Their effective work is an important step to practicability and makes the book richer in contents and more credible. I appreciate their hard work. The great amount of reference for this book is arduous labors of their original authors, to whom I am very grateful. Finally, I will show my great thanks to the leaders and colleagues of PLA Information Engineering University, Information Engineering Institute for their strong support and help, which gives me confidence to overcome difficulties and challenges of various kinds to make this book available.

Finally, especial thanks are to National Nature Science Foundation of China (NSFC) for their supports to *Project of Language Identification Of The Spontaneous Speech Over Telephone Channel*. I hope genuinely that the readers can criticize and correct the improper remarks in the book.

In the new century, speech recognition will keep developing with related disciplines. In the garden of information science, it will give out more brilliant color. Just as a saying goes: when the mountain flowers are in full bloom, it will smile mingling in their midst, I intend this book to be a pavement stone, looking forward to more and better such books in print.

Bingxi Wang

April, 2004

In Information Engineering University

内 容 简 介

本书从语音识别的基本理论出发，以“从理论到实用”为主线，讲解了国际上最新、最前沿的语音识别领域的关键技术，从语料库建立、语音信号预处理、特征提取、特征变换、模型建立等方面详细介绍了语音识别系统建立的过程，并针对语音识别系统实用化的问题，给出了一些改善语音识别系统性能的关键技术，力求语音识别能走出实验室，向实用发展。

全书共分四个部分(17章)，第一部分介绍语音识别的基本理论；第二部分介绍实用语音识别系统建立的过程；第三部分列举了语音识别系统工程化所需的关键技术；第四部分对语音识别的4个主要应用领域进行了详尽的、深入浅出的讲解，并根据最新的研究与实验结果提供了大量的实际参数、图表，与实际工作联系紧密，具有很强的可操作性与实用性。章节之间紧密配合、前后呼应，具有很强的系统性。同时，通过书中的研究过程和研究方法，读者能够在以后的研究工作中得到很大的启发。

本书可作为高等院校理工科通信和信息处理及相关专业的高年级本科生和(硕士、博士)研究生的教材或参考书，也可供从事信息处理、通信工程等专业的研究人员参考。

目　　录

第一部分　基本理论

第二部分 语音识别系统