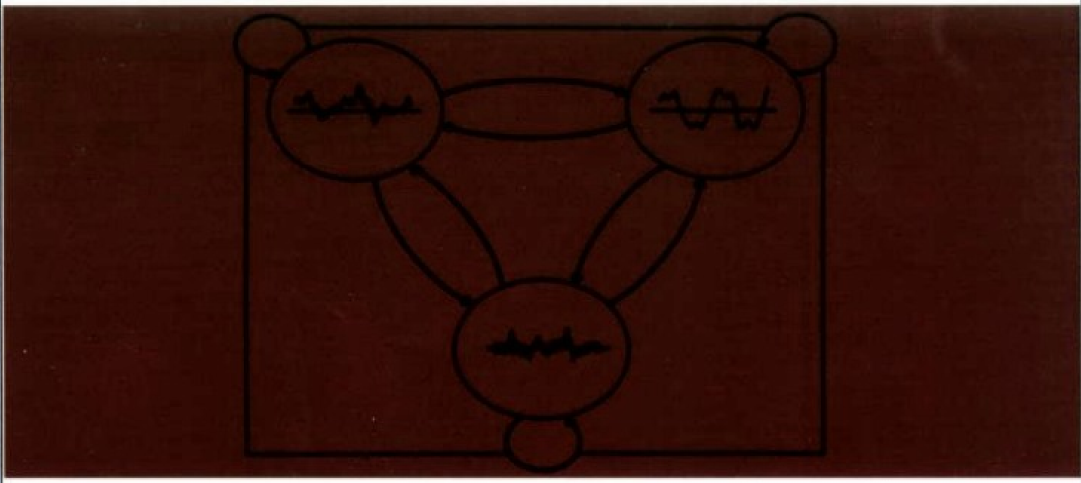
大学计算机教育丛书（影印版）

FUNDAMENTALS OF SPEECH RECOGNITION



语音识别基本原理

LAWRENCE RABINER
BIING-HWANG JUANG

清华大学出版社 · PRENTICE HALL
http://www.tup.tsinghua.edu.cn

FUNDAMENTALS OF SPEECH RECOGNITION

语音识别基本原理

Lawrence Rabiner
Biing-Hwang Juang

清华大学出版社
Prentice-Hall International, Inc.

(京)新登字 158 号

Fundamentals of speech recognition/Lawrence Rabiner, Biing-Hwang Juang
Copyright © 1993 by AT & T
Original English Language Edition Published by Prentice Hall PTR
All Rights Reserved.
For sale in Mainland China only.

本书影印版由 Prentice Hall 出版公司授权清华大学出版社在中国境内(不包括中国香港特别行政区、澳门和台湾地区)独家出版、发行。
未经出版者书面许可,不得以任何方式复制或抄袭本书的任何部分。

本书封面贴有清华大学出版社激光防伪标签,无标签者不得销售。
北京市版权局著作权合同登记号: 01-1999-2524

图书在版编目(CIP)数据

语音识别基本原理:英文/罗宾纳(Rabiner, L.)著;阮平望译. -影印版. -北京:清华大学出版社,1999.8
(大学计算机教育丛书)
ISBN 7-302-03640-3

Ⅰ.语… Ⅱ.①罗… ②阮… Ⅲ.语音识别-基本知识-英文 Ⅳ.H012

中国版本图书馆 CIP 数据核字(1999)第 30669 号

出版者: 清华大学出版社(北京清华大学校内,邮编 100084)
http:// www.tup.tsinghua.edu.cn
印刷者: 清华大学印刷厂
发行者: 新华书店总店北京发行所
开　本: 787×960 1/16　印张: 34
版　次: 1999 年 9 月第 1 版　1999 年 9 月第 1 次印刷
书　号: ISBN 7-302-03640-3/TP·2024
印　数: 0001~5000
定　价: 41.00 元

出版者的话

今天,我们的大学生、研究生和教学、科研工作者,面临的是一个国际化的信息时代。他们将需要随时查阅大量的外文资料;会有更多的机会参加国际性学术交流活动;接待外国学者;走上国际会议的讲坛。作为科技工作者,他们不仅应有与国外同行进行口头和书面交流的能力,更为重要的是,他们必须具备极强的查阅外文资料获取信息的能力。有鉴于此,在国家教委所颁布的“大学英语教学大纲”中有一条规定:专业阅读应作为必修课程开设。同时,在大纲中还规定了这门课程的学时和教学要求。有些高校除开设“专业阅读”课之外,还在某些专业课拟进行英语授课。但教、学双方都苦于没有一定数量的合适的英文原版教材作为教学参考书。为满足这方面的需要,我们陆续精选了一批国外计算机科学方面最新版本的著名教材,进行影印出版。我社获得国外著名出版公司和原著作者的授权将国际先进水平的教材引入我国高等学校,为师生们提供了教学用书,相信会对高校教材改革产生积极的影响。

我们欢迎高校师生将使用影印版教材的效果,意见反馈给我们,更欢迎国内专家、教授积极向我社推荐国外优秀计算机教育教材,以利我们将《大学计算机教育丛书(影印版)》做得更好,更适合高校师生的需要。

清华大学出版社
《大学计算机教育丛书(影印版)》项目组
1999.6

016877

PREFACE

This book is an outgrowth of an association between the authors which started over 10 years ago when one of us (BHJ) was a graduate student at the University of California at Santa Barbara and the other (LRR) was a supervisor at AT&T Bell Laboratories. We began our relationship with a mutual interest in the problem of designing and implementing vector quantization for speech processing. This association turned into a full technical partnership and strong friendship when Fred Juang joined Bell Laboratories, initially in the development area and subsequently in research. The spark that ignited formal work on this book was a series of short courses taught by one of us (LRR) on speech recognition. After several iterations of teaching, it became clear that the area of speech recognition, although still changing and growing, had matured to the point where a book that covered its theoretical underpinnings was warranted.

Once we had decided to write this book, there were several key issues that had to be resolved, including how deep to go into areas like linguistics, natural language processing, and the practical side of the problem; whether to discuss individual systems proposed by various research labs around the world; and how extensively to cover applications. Although there were no simple answers to these questions, it rapidly became obvious to us that the fundamental goal of the book would be to provide a theoretically sound, technically accurate, and reasonably complete description of the basic knowledge and ideas that constitute a modern system for speech recognition by machine. With these basic guiding principles in mind, we were able to decide consistently (and hopefully reasonably) what material had to be included, and what material would be presented in only a cursory manner. We leave it up to you, the reader, to decide if our choices have been wise ones.

The formal organization of the book is as follows. Chapter 1, called "Fundamentals of Speech Recognition," provides an overview of the entire field with a discussion of the breadth and depth of the various disciplines that are required for a deep understanding of all aspects of speech recognition. The concept of a task-oriented, speech-recognition system is introduced and it is shown that "base level" speech or sound recognition is only one step in a much larger process where higher-level task information, in the form of syntax, semantics, and pragmatics, can often play a major role. After a formal description of the material to be covered in each of the chapters, we give a brief history of speech recognition research in order to put the material presented in this book in its proper perspective.

Chapter 2, entitled the "The Speech Signal: Production, Perception, and Acoustic-Phonetic Characterization," provides a review of the theory of acoustic-phonetics in which we try to characterize basic speech sounds according to both their linguistic properties and the associated acoustic measurements. We show that although there is a solid basis for the linguistic description of sounds and a good understanding of the associated acoustics of sound production, there is, at best, a tenuous relationship between a given linguistic sound and a repeatable, reliable, measureable set of acoustic parameters. As such a wide variety of approaches to speech recognition have been proposed, including those based on the ideas of acoustic-phonetics, statistical pattern-recognition methods, and artificial intelligence (so-called expert system) ideas. We discuss the relative advantages and disadvantages of each of these approaches and show why, on balance, the pattern-recognition approach has become the method of choice for most modern systems.

In Chapter 3, entitled "Signal Processing and Analysis Methods for Speech Recognition," we discuss the fundamental techniques used to provide the speech features used in all recognition systems. In particular we discuss two well-known and widely used methods of spectrum analysis, namely the filter bank approach and the linear prediction method. We also show how the method of vector quantization can be used to code a spectral vector into one of a fixed number of discrete symbols in order to reduce the computation required in a practical system. Finally we discuss an advanced spectral analysis method that is based on processing within the human auditory system—an ear model. The ultimate goal of such a system is to increase the robustness of the signal representation and make the system relatively insensitive to noise and reverberation, in much the same way as the human ear.

Chapter 4, entitled "Pattern-Comparison Techniques," deals with the fundamental problems of defining speech feature vector patterns (from spoken input), and comparing pairs of feature vector patterns both locally (i.e., at some point in time), and globally (i.e., over the entire pattern) so as to derive a measure of similarity between patterns. To solve this pattern-comparison problem requires three types of algorithms, namely a speech-detection method (which essentially separates the speech signal from the background), a spectral vector comparison method (which compares two individual spectral vectors), and a global pattern comparison method which aligns the two patterns locally in time and compares the aligned patterns over the entire duration of the patterns. It is shown that a key issue is the way in which time alignment between patterns is achieved.

Chapter 5, entitled "Speech Recognition System Design and Implementation Issues," discusses the key issues of training a speech recognizer and adapting the recognizer parameters to different speakers, transmission conditions, and speaking environments. A key concept in most modern systems is that of learning, namely improving recognizer performance over time based on additional training provided by the user of the system. Adaptation methods provide a formalism for such learning.

In Chapter 6, "Theory and Implementation of Hidden Markov Models," we discuss a basic set of statistical modeling techniques for characterizing speech. The collection of methods, popularly called Hidden Markov Models, is a powerful set of tools for providing a statistical model of both the static properties of sounds and the dynamical changes that occur across sounds. Methods for time aligning patterns with models are discussed along with different ways of building the statistical models based on the type of representation, the sound being modeled, the class of talkers, and so forth.

Chapters 7 and 8, entitled "Speech Recognition Based on Connected Word Models" and "Large Vocabulary Continuous Speech Recognition," extend the speech-recognition problem from single word sequences to fluent speech. Modeling techniques based on whole word models are discussed in Chapter 7 where we assume that we are interested in recognizing sequences of digits, alphanumerics, and so forth. For this type of system whole-word models are most reasonable since the vocabulary is typically small and highly constrained. Hence the statistical properties of the word models, in all word contexts, can be learned from a reasonably sized training set. Modeling techniques based on subword units are discussed in Chapter 8 where we assume unlimited size vocabulary. Hence a key issue is what units are used, how context dependent the units should be, how unit models are trained reliably (and robustly to different vocabularies and tasks), and how large vocabulary recognition systems based on such units are efficiently implemented.

Finally, in Chapter 9, entitled "Task-Oriented Applications of Automatic Speech Recognition," we come full circle and return to the concept of a task-oriented system. We discuss the basic principles that make some tasks successful while others fail. By way of example we discuss, in fairly general terms, a couple of task-oriented recognizers and show how they perform in practice.

The material in this book is primarily intended for the practicing engineer, scientist, linguist, programmer, and so forth, who wants to learn more about this fascinating field. We assume a basic knowledge of signal processing and linear system theory as provided in an entry level course in digital signal processing. Although not intended as a formal university course, the material in this book is indeed suitable for a one-semester course at the graduate or high undergraduate level. Within almost every chapter we have provided "exercises" for the student to assess how well he or she understands the material. Solutions to the exercises are provided immediately following the exercise. Hence, for maximum effectiveness, each student must exercise self-discipline to work through an answer before comparing it with the published solution.

In order to truly understand the fundamentals of speech recognition, a person needs hands-on experience with the software, hardware, and platforms. Hence we strongly encourage all serious readers of this book to program the algorithms, implement the systems, and literally build applications. Without such practical experience the words in this book will not come alive for most people.

ACKNOWLEDGMENTS

Although the authors take full responsibility for the material presented in this book, we owe a great debt to our colleagues, both with AT&T Bell Laboratories and outside, for their many technical contributions which underly the material presented. In particular the authors owe a debt of gratitude to Dr. James L. Flanagan (currently director of the CAIP Institute at Rutgers University) for his roles in guiding and shaping both our careers and the field of speech processing. Without Jim's understanding and inspiration, this book would never have existed.

The number of people who have made substantial contributions to speech recognition are too numerous to mention. However there are three individuals who have had a profound influence on the field and they deserve special mention. The first is Professor Raj Reddy of Carnegie-Mellon University who was essentially the first person to realize the vast potential of speech recognition and has devoted over 25 years as a leader, innovator, and educator in this field. The second individual of note is Dr. Jack Ferguson (retired from Institute for Defense Analyses in Princeton) who is the person most responsible for development of the theory of the Hidden Markov Model as applied to speech recognition. Dr. Ferguson, as editor of the key textbook in this area and lecturer, par excellence, has spread the word on Hidden Markov Models so that this technology has rapidly risen from technical obscurity to become the preeminent method of speech recognition today. Finally, the third individual of note is Dr. Fred Jelinek of IBM, who has led the world's largest speech-recognition research group for almost two decades and has been responsible for a large number of major innovations in large vocabulary speech recognition. These three individuals have played major roles in nurturing the technology and enabling it to reach the state of maturity it has achieved today.

Within Bell Laboratories the authors have drawn freely from the research of our former and current colleagues. We would like to acknowledge the direct support and contributions of the following individuals: Prof. Ronald Schafer (currently at Georgia Tech.), Dr. Steve Levinson, Dr. Bishnu Atal, Dr. Esther Levin, Dr. Tali Tishby (currently at the Hebrew University in Jerusalem), Dr. Oded Ghitza, Jay Wilpon, Dr. Frank Soong, Dr. Mohan Sondhi, Dr. Yariv Ephraim, Dr. Cory Myers (currently at Atlantic Aerospace), Dr. Aaron Rosenberg, Dr. Chin Lee and Dr. Roberto Pieraccini. We thank these colleagues for their research contributions and for their friendship and guidance over the years.

Several individuals provided technical comments on a preliminary version of the manuscript. These comments were invariably insightful and provided valuable feedback on ways to improve the book. We would like to acknowledge and thank the following individuals for their help: Dr. Mohan Sondhi, Dr. Yariv Ephraim, Dr. Esther Levin, Dr. Oded Ghitza, Dr. Roberto Pieraccini, Dr. Chin Lee, Dr. Wu Chou and Dr. Sadaoki Furui (NTT Corporation, JAPAN).

The production of a book is an essentially infinite task which is seemingly an endless one. However all good things do come to an end and this one was no exception. The authors owe a great deal to Ms. Martina Sharp of the Bell Labs Word Processing group, who entered all the text material for the book using the LaTeX system. Tina worked on three

generations of the manuscript with a degree of skill that is essentially unmatched among her peers. It was a pleasure working with Tina and we look forward to future projects with her. Most of the art work for the book was produced by Ms. Danuta Sowinska-Khan of the Bell Labs art department. Danuta was a pleasure to work with and is a professional in her trade. The material she produced always met the highest professional standards and did honor to the manuscript in which it appears. A great deal of assistance in the preparation of drafts of the book, getting figures ready for production, and so forth was provided by Irene Morrongiello. We thank her and wish to express our appreciation for a job well done.

Finally, to the crew at Prentice Hall we owe a debt of gratitude for the professional way the book was handled from inception to final production. Karen Gettman provided the incentive to the authors to actually produce the manuscript and make the book a reality. We thank Lisa Garboski, the production editor, for help in all phases of the book production cycle.

Lawrence R. Rabiner
Biing-Hwang Juang

CONTENTS

LIST OF FIGURES