

**THE ELECTRICAL ENGINEERING
AND APPLIED SIGNAL PROCESSING SERIES**

Edited by Alexander Poularikas

*The Advanced Signal Processing Handbook:
Theory and Implementation for Radar, Sonar,
and Medical Imaging Real-Time Systems*

Stergios Stergiopoulos

The Transform and Data Compression Handbook

K.R. Rao and P.C. Yip

Handbook of Multisensor Data Fusion

David Hall and James Llinas

Handbook of Neural Network Signal Processing

Yu Hen Hu and Jenq-Neng Hwang

Handbook of Antennas in Wireless Communications

Lal Chand Godara

Noise Reduction in Speech Applications

Gillian M. Davis

Signal Processing Noise

Vyacheslav P. Tuzlukov

Digital Signal Processing with Examples in MATLAB®

Samuel Stearns

Applications in Time-Frequency Signal Processing

Antonia Papandreou-Suppappola

The Digital Color Imaging Handbook

Gaurav Sharma

Pattern Recognition in Speech and Language Processing

Wu Chou and Bing Huang Juang

Forthcoming Titles

Propagation Data Handbook for Wireless Communication System Design

Robert Crane

Smart Antennas

Lal Chand Godara

Nonlinear Signal and Image Processing: Theory, Methods, and Applications

Kenneth Barner and Gonzalo R. Arce

Forthcoming Titles (*continued*)

Soft Computing with MATLAB[®]

Ali Zilouchian

Signal and Image Processing Navigational Systems

Vyacheslav P. Tuzlukov

Wireless Internet: Technologies and Applications

Apostolis K. Salkintzis and Alexander Poularikas

PATTERN
RECOGNITION in
SPEECH and
LANGUAGE
PROCESSING

Edited by
WU CHOU
Avaya Labs Research

BIING HWANG JUANG
Georgia Institute of Technology



CRC PRESS

Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Pattern recognition in speech and language processing / edited by Wu Chou and Biing-Hwang Juang.

p. cm.

Includes bibliographical references and index.

ISBN 0-8493-1232-9 (alk. paper)

1. Automatic speech recognition. 2. Pattern recognition systems. I. Chou, Wu. II. Juang, B. H. (Biing-Hwang)

TK7882.S65 P39 2003

006.4'54—dc21

2002191163

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the authors and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the publisher.

All rights reserved. Authorization to photocopy items for internal or personal use, or the personal or internal use of specific clients, may be granted by CRC Press LLC, provided that \$1.50 per page photocopied is paid directly to Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA. The fee code for users of the Transactional Reporting Service is ISBN 0-8493-1232-9/03/\$0.00+\$1.50. The fee is subject to change without notice. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

The consent of CRC Press LLC does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press LLC for such copying.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

Visit the CRC Press Web site at www.crcpress.com

© 2003 by CRC Press LLC

No claim to original U.S. Government works

International Standard Book Number 0-8493-1232-9

Library of Congress Card Number 2002191163

Printed in the United States of America 1 2 3 4 5 6 7 8 9 0

Printed on acid-free paper

Preface

Approaches to the problems of designing speech and language processing algorithms for human machine communication used to be taken from the perspectives of linguistics and speech science, until the late 1970s. Due to the advances in computing and statistical modeling, data driven pattern recognition methods have become a fast moving research area during the past two decades and contributed much to the progress in this field. As the era of information age continues to develop, we witness an ever increasing need in intelligent human-machine communications, as well as the creation of machine understandable metadata for Web content and other information sources. This handbook is to fill the need of a systematic and up-to-date presentation of new pattern recognition approaches in speech and language processing.

The book starts with fundamentals and recent theoretical advances in pattern recognition with an emphasis on classifier design criteria and optimization procedures. It covers several recent research advances in this area, such as the minimum error rate (MCE) method, the minimum Bayes risk approach, adaptive system design and decision rules, neural networks, distributed recognizers, and decision fusion. These methods depart from the conventional paradigm which links a classifier design to the classical problem of distribution estimation. Instead, more meaningful criteria are introduced which significantly improve the discrimination power of a classifier, particularly when applied to speech problems in which the notion of data distribution is difficult to realize.

The second part of the book is, therefore, specially focused on the approaches and methods applied to speech processing. It covers topics such as Bayes minimum risk approach to speech recognition, large vocabulary speech recognition based on statistical methods, recognition of spontaneous speech in dialogue interaction, speech and speaker verification, and audio information retrieval and indexing. These chapters provide a comprehensive coverage of recent advances in applying pattern recognition to real systems in speech and audio processing.

The third part of the book is devoted to topics of pattern recognition in language processing. It contains chapters in language modeling based on latent semantic indexing, salient information representation and processing in natural language dialogue system, statistical machine translation, methods in topic detection, tracking, and name identity identification. These topics are new trends in language processing, and significant progress has been made in recent years. It has a direct impact to the practice and implementation of information processing systems for Web content, broadcast news, and other content-rich information resources.

This book is a collective effort, motivated by the excitement of the new advances in

this field and the urgent need to bring these advances to a general audience. The contributing authors of this book are leading experts in the field of speech and language processing. Attempts are made to make each chapter self-contained and comprehensible for readers with general background in pattern recognition and information processing. It is intended to be a handbook or reference textbook for researchers, graduate students, and advanced undergraduate students who want to follow the new advances in pattern recognition. Sufficient references are provided at the end of each chapter to serve as an entry point for an interested reader to pursue further.

We would like to thank all contributors of this book. Without their commitment and quality of work, this book would not be possible. We appreciate the support and encouragement from our colleagues at Avaya Labs Research during the preparation of this book. It was a pleasant working experience with CRC Press - their technical support was very helpful to us.

Wu Chou
Biing-Hwang Juang

*Basking Ridge, New Jersey
September, 2002*

Contributors

A. Abella

Speech Research
AT&T Laboratories
Florham Park, NJ

James Allan

Department of Computer Science
University of Massachusetts Amherst
Amherst, MA

T. Alonso

Speech Research
AT&T Laboratories
Florham Park, NJ

Jerome R. Bellegarda

Spoken Language Group
Apple Computer, Inc.
Cupertino, CA

William Byrne

Center for Language and Speech
Processing
Johns Hopkins University
Baltimore, MD

Wu Chou

Avaya Labs Research
Basking Ridge, NJ

Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
Tokyo, Japan

Jean-Luc Gauvain

LIMSI-CNRS

Université de Paris Sud
Orsay Cedex, France

Vaibhava Goel

T.J. Watson Research Center
IBM
Yorktown Heights, NY

Allen L. Gorin

Speech Research
AT&T Laboratories
Florham Park, NJ

Qiang Huo

Department of Computer Science and
Information Systems
The University of Hong Kong
Hong Kong, China

Biing-Hwang Juang

Avaya Labs Research
Basking Ridge, NJ

Shigeru Katagiri

Intelligent Communication Laboratory and
Speech Open Laboratory
Nippon Telegraph and Telephone
Corporation
Tokyo, Japan

Lori Lamel

LIMSI-CNRS
Université de Paris Sud
Orsay Cedex, France

Qi (Peter) Li

Bell Laboratories

Lucent Technologies
Murray Hill, NJ

John Makhoul
BBN Technologies
Cambridge, MA

Hermann Ney
Lehrstuhl fuer Informatik VI
Human Language Technology and
Pattern Recognition
Computer Science Department
University of Technology
Aachen, Germany

F. J. Och
Lehrstuhl fuer Informatik VI
Human Language Technology and
Pattern Recognition
Computer Science Department
University of Technology
Aachen, Germany

G. Riccardi
Speech Research
AT&T Laboratories
Florham Park, NJ

Richard M. Schwartz
BBN Technologies
Cambridge, MA

J. H. Wright
Speech Research
AT&T Laboratories
Florham Park, NJ

Contents

1 Minimum Classification Error (MCE) Approach in Pattern Recognition

Wu Chou Avaya Labs Research, Avaya Inc., USA

- 1.1 Introduction
- 1.2 Optimal Classifier from Bayes Decision Theory
- 1.3 Discriminant Function Approach to Classifier Design
- 1.4 Speech Recognition and Hidden Markov Modeling
 - 1.4.1 Hidden Markov Modeling of Speech
- 1.5 MCE Classifier Design Using Discriminant Functions
 - 1.5.1 MCE Classifier Design Strategy
 - 1.5.2 Optimization Methods
 - 1.5.3 Other Optimization Methods
 - 1.5.4 HMM as a Discriminant Function
 - 1.5.5 Relation between MCE and MMI
 - 1.5.6 Discussions and Comments
- 1.6 Embedded String Model Based MCE Training
 - 1.6.1 String Model Based MCE Approach
 - 1.6.2 Combined String Model Based MCE Approach
 - 1.6.3 Discriminative Feature Extraction
- 1.7 Verification and Identification
 - 1.7.1 Speaker Verification and Identification
 - 1.7.2 Utterance Verification
- 1.8 Summary

2 Minimum Bayes-Risk Methods in Automatic Speech Recognition

Vaibhava Goel^{*} and *William Byrne*[†] ^{*}IBM; [†]Johns Hopkins University

- 2.1 Minimum Bayes-Risk Classification Framework
 - 2.1.1 Likelihood Ratio Based Hypothesis Testing
 - 2.1.2 Maximum A-Posteriori Probability Classification
 - 2.1.3 Previous Studies of Application Sensitive ASR
- 2.2 Practical MBR Procedures for ASR
 - 2.2.1 Summation over Hidden State Sequences
 - 2.2.2 MBR Recognition with N-best Lists
 - 2.2.3 MBR Recognition with Lattices
- 2.3 Segmental MBR Procedures
 - 2.3.1 Segmental Voting
 - 2.3.2 ROVER

- 2.3.3 e-ROVER
- 2.4 Experimental Results
 - 2.4.1 Parameter Tuning within the MBR Classification Rule
 - 2.4.2 Utterance Level MBR Word and Keyword Recognition
 - 2.4.3 ROVER and e-ROVER for Multilingual ASR
- 2.5 Summary
- 2.6 Acknowledgements

3 A Decision Theoretic Formulation for Robust Automatic Speech Recognition

Qiang Huo The University of Hong Kong, Hong Kong, China

- 3.1 Introduction
- 3.2 Optimal Bayes' Decision Rule for ASR
- 3.3 Adaptive Decision Rules Constructed from Training Samples
 - 3.3.1 Plug-in Bayes' Decision Rules with Maximum-likelihood Density Estimate
 - 3.3.2 Maximum-Discriminant Decision Rules Minimizing the Empirical Classification Error
 - 3.3.3 Discussion
- 3.4 Violations of Modeling Assumptions in ASR
 - 3.4.1 Types of Distortions
 - 3.4.2 Towards Adaptive and Robust ASR
- 3.5 Improving Adaptive Decision Rules via Decision Parameter Adaptation
 - 3.5.1 Decision Parameter Adaptation for Stationary Operating Conditions
 - 3.5.2 Decision Parameter Adaptation for Slowly Changing Operating Conditions
 - 3.5.3 Decision Parameter Adaptation for Switching Operating Conditions
 - 3.5.4 Discussion
- 3.6 Robust Decision Rules
 - 3.6.1 Decision Rule Robustness
 - 3.6.2 Minimax Classification Rule
 - 3.6.3 Bayesian Predictive Classification Rule
 - 3.6.4 Discussion
- 3.7 Summary

4 Speech Pattern Recognition using Neural Networks

Shigeru Katagiri NTT Communication Science Laboratories

- 4.1 Introduction
- 4.2 Bayes Decision Theory
 - 4.2.1 Preparations
 - 4.2.2 Decision Rule
 - 4.2.3 Minimum Error-rate Classification

- 4.2.4 Probability Function Estimation
- 4.2.5 Discriminative Training
- 4.3 Speech Recognizers Based on Neural Networks
 - 4.3.1 Preparations
 - 4.3.2 Classification Error Minimization
 - 4.3.3 Squared Error Minimization
 - 4.3.4 Cross Entropy Minimization
- 4.4 Fusion of Multiple Classification Decisions
 - 4.4.1 Principles
 - 4.4.2 Examples of Embodiment
- 4.5 Concluding Remarks
- 4.6 Appendix: Maximizing Mutual Information

5 Large Vocabulary Speech Recognition Based on Statistical Methods

Jean-Luc Gauvain and Lori Lamel LIMSI, France

- 5.1 Introduction
- 5.2 Overview
- 5.3 Language Modeling
 - 5.3.1 Text Preparation
 - 5.3.2 Vocabulary Selection
 - 5.3.3 N-gram Estimation
 - 5.3.4 LM Adaptation
- 5.4 Pronunciation Modeling
- 5.5 Acoustic Modeling
 - 5.5.1 Acoustic Front-end
 - 5.5.2 Modeling Allophones
 - 5.5.3 HMM Parameter Estimation
 - 5.5.4 HMM Adaptation
- 5.6 Decoding
 - 5.6.1 Speech/Non-speech Detection
 - 5.6.2 Decoding Strategies
 - 5.6.3 Efficiency
 - 5.6.4 Confidence Measures
- 5.7 Indicative Performance Levels
 - 5.7.1 Dictation
 - 5.7.2 Speech Recognition for Dialog Systems
 - 5.7.3 Transcription for Audio Indexation
- 5.8 Portability and Language Dependencies

6 Toward Spontaneous Speech Recognition and Understanding

Sadaoki Furui Tokyo Institute of Technology

- 6.1 Introduction
- 6.2 Four Categories of Speech Recognition Tasks
- 6.3 Spontaneous Speech Recognition and Understanding - Review
 - 6.3.1 Category I (human-to-human dialogue)

- 6.3.2 Category II (human-to-human monologue)
- 6.3.3 Category III (human-to-machine dialogue)
- 6.4 Japanese National Project on Spontaneous Speech Corpus and Processing Technology
 - 6.4.1 Project Overview
 - 6.4.2 Corpus
- 6.5 Automatic Transcription of Spontaneous Presentation
 - 6.5.1 Recognition Task
 - 6.5.2 Language and Acoustic Modeling
 - 6.5.3 Recognition Results
 - 6.5.4 Analysis on Individual Differences
 - 6.5.5 Discussion
- 6.6 Automatic Speech Summarization and Evaluation
 - 6.6.1 Summarization of Each Sentence Utterance
 - 6.6.2 Summarization of Multiple Utterances
 - 6.6.3 Evaluation
 - 6.6.4 Discussion
- 6.7 Spontaneous Speech Recognition and Understanding Research Issues
 - 6.7.1 Language Models and Corpora
 - 6.7.2 Message-driven Speech Recognition and Understanding
 - 6.7.3 Statistical Approaches and Speech Science
 - 6.7.4 Research on the Human Brain
 - 6.7.5 Dynamic Spectral Features
- 6.8 Conclusion

7 Speaker Authentication

*Qi Li** and *Biing-Hwang Juang*[†] *Bell Labs; [†]Avaya Labs Research

- 7.1 Introduction
 - 7.1.1 Speaker Recognition and Verification
 - 7.1.2 Verbal Information Verification
- 7.2 Pattern Recognition in Speaker Authentication
 - 7.2.1 Bayesian Decision Theory
 - 7.2.2 Stochastic Models for Stationary Process
 - 7.2.3 Stochastic Models for Non-Stationary Process
 - 7.2.4 Speech Segmentation
 - 7.2.5 Statistical Verification
- 7.3 Speaker Verification System
- 7.4 Verbal Information Verification
 - 7.4.1 Utterance Segmentation
 - 7.4.2 Subword Hypothesis Testing
 - 7.4.3 Confidence Measure Calculation
 - 7.4.4 Sequential Utterance Verification
 - 7.4.5 VIV Experimental Results
- 7.5 Speaker Authentication by Combining SV and VIV

7.6 Summary

8 HMMs for Language Processing Problems

Richard M. Schwartz and John Makhoul BBN Technologies, Verizon

8.1 Introduction

8.2 Use of Probabilities

8.2.1 Hidden Markov Models

8.3 Name Spotting

8.4 Topic Classification

8.4.1 The Model

8.4.2 Estimating HMM Parameters

8.4.3 Classification

8.4.4 Experiments

8.5 Information Retrieval

8.5.1 A Bayesian Model for IR

8.5.2 Training the IR HMM

8.5.3 Performance

8.6 Event Tracking

8.7 Unsupervised Topic Detection

8.8 Summary

9 Statistical Language Models With Embedded Latent Semantic Knowledge

Jerome R. Bellegarda Apple Computer, Inc.

9.1 Introduction

9.1.1 Scope Locality

9.1.2 Syntactically-Driven Span Extension

9.1.3 Semantically-Driven Span Extension

9.1.4 Organization

9.2 Latent Semantic Analysis

9.2.1 Feature Extraction

9.2.2 Singular Value Decomposition

9.2.3 General Behavior

9.3 LSA Feature Space

9.3.1 Word Clustering

9.3.2 Word Cluster Example

9.3.3 Document Clustering

9.3.4 Document Cluster Example

9.4 Semantic Classification

9.4.1 Framework Extension

9.4.2 Semantic Inference

9.4.3 Caveats

9.5 N-gram+LSA Language Modeling

9.5.1 LSA Component

9.5.2 Integration with N-grams

- 9.5.3 Context Scope Selection
- 9.6 Smoothing
 - 9.6.1 Word Smoothing
 - 9.6.2 Document Smoothing
 - 9.6.3 Joint Smoothing
- 9.7 Experiments
 - 9.7.1 Experimental Conditions
 - 9.7.2 Experimental Results
 - 9.7.3 Context Scope Selection
- 9.8 Inherent Trade-Offs
 - 9.8.1 Cross-Domain Training
 - 9.8.2 Discussion
- 9.9 Conclusion

10 Semantic Information Processing of Spoken Language – How May I Help You?sm

A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright, AT&T Laboratories

- 10.1 Introduction
- 10.2 Call-Classification
- 10.3 Language Modeling for Recognition and Understanding
- 10.4 Dialog
- 10.5 Conclusions

11 Machine Translation Using Statistical Modeling

Herman Ney, and F. J. Och Aachen University of Technology, Germany

- 11.1 Introduction
- 11.2 Statistical Decision Theory and Linguistics
 - 11.2.1 The Statistical Approach
 - 11.2.2 Bayes Decision Rule for Written Language Translation
 - 11.2.3 Related Approaches
- 11.3 Alignment and Lexicon Models
 - 11.3.1 Concept of Alignment Modelling
 - 11.3.2 Hidden Markov Models
 - 11.3.3 Models IBM 1–5
 - 11.3.4 Training
 - 11.3.5 Search
 - 11.3.6 Algorithmic Differences between Speech Recognition and Language Translation
- 11.4 Alignment Templates: From Single Words to Word Groups
 - 11.4.1 Concept
 - 11.4.2 Training
 - 11.4.3 Search
- 11.5 Experimental Results
 - 11.5.1 The Task and the Corpus

- 11.5.2 Offline Results
- 11.5.3 Integration into the VERBMOBIL Prototype System
- 11.5.4 Final Evaluation
- 11.6 Speech Translation: The Integrated Approach
 - 11.6.1 Principle
 - 11.6.2 Practical Implementation
- 11.7 Summary
- 11.8 References

12 Modeling Topics for Detection and Tracking

James Allan University of Massachusetts Amherst

- 12.1 Topic Detection and Tracking
 - 12.1.1 Topic and Events
 - 12.1.2 TDT Tasks
 - 12.1.3 Corpora
 - 12.1.4 Evaluation
- 12.2 Basic Topic Models
 - 12.2.1 Vector Space
 - 12.2.2 Language Models
- 12.3 Implementing the Models
 - 12.3.1 Named Entities
 - 12.3.2 Document Expansion
 - 12.3.3 Clustering
 - 12.3.4 Time Decay
- 12.4 Comparing Models
 - 12.4.1 Nearest Neighbors
 - 12.4.2 Decision Trees
 - 12.4.3 Model-to-Model
- 12.5 Miscellaneous Issues
 - 12.5.1 Deferral
 - 12.5.2 Multi-modal Issues
 - 12.5.3 Multi-lingual Issues
- 12.6 Using TDT Interactively
 - 12.6.1 Demonstrations
 - 12.6.2 Timelines
- 12.7 Modeling Events
- 12.8 Conclusion

1

Minimum Classification Error (MCE) Approach in Pattern Recognition

Wu Chou

Avaya Labs Research, Avaya Inc., USA

CONTENTS

- 1.1 Introduction
- 1.2 Optimal Classifier from Bayes Decision Theory
- 1.3 Discriminant Function Approach to Classifier Design
- 1.4 Speech Recognition and Hidden Markov Modeling
- 1.5 MCE Classifier Design Using Discriminant Functions
- 1.6 Embedded String Model Based MCE Training
- 1.7 Verification and Identification
- 1.8 Summary
- Acknowledgement
- References

1.1 Introduction

Pattern recognition is a fast moving research area. The advent of powerful computing devices and the success of statistical approaches, such as hidden Markov model for speech and language processing, triggered a renewed pursuit for more powerful statistical methods to further reduce the pattern recognition error rate and improve the robustness of the pattern classifier across various adverse conditions. Among this new pursuit, the use of discriminant function methods in pattern recognition has emerged as a promising approach, and it is applied successfully to speech and language processing. This chapter is intended to provide a revisit to the statistical formulation of the minimum classification error (MCE) based discriminative methods in speech and language processing, take a critical view of the approach, provide a comprehensive overview of the field, and hopefully inspire other innovations that would potentially lead to new discriminative methods in pattern recognition.

Although the statistical formulation of MCE based discriminative methods has its root in the classical Bayes decision theory, it departs from the conventional paradigm

This chapter is developed based on "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," by Wu Chou, appeared in *Proceedings of The IEEE*, Vol. 88, No. 8, ©2000 IEEE.

which links a recognition task to the problem of distribution estimation. Instead, it takes a discriminant function based statistical pattern classification approach, and for a given family of discriminant function, optimal classifier/recognizer design involves finding a set of parameters which minimize the empirical pattern recognition error rate. The use of discriminant function in pattern recognition was started many years ago. One classical example of using discriminant function for classifier design in statistical literature is the two class classification problem using linear discriminant functions [28, 31]. In particular, a window based method was described in [28] for the two class classification problem using linear discriminant functions that minimize the probability of classification error rate. The focus of this chapter is on the recent development of the general MCE based discriminative methods. The discriminant functions that we encounter are usually non-linear and often related to the structure of the statistical framework used in speech and language processing such as hidden Markov models.

The reason of taking a discriminant function based approach to classifier design, as will be further elaborated, is due mainly to the fact that we lack complete knowledge of the form of the data distribution and training data are inadequate, particularly in dealing with speech and language problems. The performance of a recognizer is normally defined by its expected recognition error rate, and an optimal recognizer should be the one that achieves the least expected rate of recognition error. The difference between the distribution estimation based approach and the discriminant function based MCE approach lies in the way the recognition error is expressed and in the computational steps that would lead to the minimization of such error functions. A key to the development of the MCE method is a new error function which incorporates the recognition operation and performance in a functional form, from which the performance of the classifier can be directly evaluated and optimized. Classifier design without assuming the knowledge of class posterior probabilities, which are the basis of the distribution estimation based classifier design, has been studied in many areas. In particular, Tsytkin [112] and Amari [5, 2] pioneered this approach for self-learning and self-organizing nets. They formulated the problem of self-learning into a classification problem which consists of optimal partitioning of the observation space into regions, X_k , for which the expected risk, R , is minimized. In addition, a mathematical minimization procedure, generalized probabilistic descent (GPD) algorithm or stochastic approximation, was proposed as a means for classifier design under this framework. Since then, various loss functions have been used in designing classifiers, including those popular mean-square error based loss functions. However, many tractable loss functions do not have a direct relation to the recognition error rate minimization, and therefore, albeit based on discriminant functions, they are not directly related to recognition error rate which should be the most sensible choice for classifier design.

Over the past decade, the MCE based approach has been developed to overcome the fundamental limitations of the traditional approach and to directly link the classifier design problem to classification error rate minimization. In order to alleviate the dependency on the class posterior distributions, a discriminant function based MCE approach was proposed by Juang et al. [50] as an alternative to optimal classifier de-

sign. Although this approach applies to the pattern recognition problem in general, it finds various applications in speech and language processing. It was first applied to dynamic time warping based pattern recognition systems [16, 56]. Application to hidden Markov model based continuous speech recognition systems was formulated as a segmental and string model based MCE approach [18, 19], and successful applications of this approach were reported in [20, 27, 35, 39, 80, 81]. This approach was further extended to form a combined string model, in which training of other model components in speech and language processing can be achieved under a unified MCE framework [22, 41]. It was applied to discriminative model combination [13, 79] and to applications in speaker identification and verification [74, 36, 62]. The basic idea of the MCE approach was further developed for applications in utterance verification problems [101, 111, 77]. A general framework of combining detection and verification in speech recognition and understanding was also proposed, in which the discriminant function based pattern recognition approach was applied in both detection and verification processes [54, 60].

We begin in the next section with a brief review of the Bayes decision theory and its application to the formulation of statistical pattern recognition problem. We introduce the discriminant function based statistical pattern recognition approach in Section 3. In Section 4, we provide a brief introduction to speech recognition and hidden Markov modeling. The discriminant function based MCE pattern recognition approach and its application to HMM based speech recognition systems are introduced in Section 5. Comparisons are made to other criteria in speech recognition and in particular, we study the relation between MCE and MMI (maximum mutual information) criteria in classifier design in the second half of Section 5. In Section 6, we study the embedded string model based MCE approach and its extension to the higher level combined string model. We discuss issues and applications in discriminative model combination, discriminative language model estimation, and discriminative feature extraction under the general theoretical framework of the combined string model. Section 7 is devoted to applications of discriminant function based pattern recognition approach in verification and identification. The discriminant function approach is studied for various applications in speech and language processing, such as speaker identification and verification, utterance verification, recognition based on generalized confidence measures, detection and verification based approach in speech recognition and understanding. The chapter is summarized with discussions in Section 8.

1.2 Optimal Classifier from Bayes Decision Theory

For an M class classification problem, a classifier is to classify each random sample x into one of the M classes. We denote these classes by C_i , $i = 1, 2, \dots, M$. The classifier $C(x)$ defines a mapping from the sample space $x \in X$ to the discrete

categorical set $C_i \in Y$. Let $P(x, C_i)$ be the joint probability distribution of x and C_i , a quantity which is assumed to be known to the designer of the classifier. In other words, the designer has full knowledge of the random nature of the source. From the set of joint probability distributions, the marginal and the conditional probability distributions can be easily calculated.

In order to characterize the performance of the classifier, every class pair (j, i) can be associated with a cost or loss function e_{ji} which signifies the cost of classifying (or recognizing) a class i observation into a class j event. The loss function is generally non-negative with $e_{ii} = 0$ representing correct classification. The loss function is a function from $X \times Y \rightarrow R$ where R is the set of real numbers. In classification, we make a decision $C(x)$ for observing a random sample x . Since $P(C_j | x)$ is the class posterior probability that the random input x is from C_j , the average loss associated with making a decision $C(x) = C_i$ can be defined as [31]

$$R(C_i | x) = \sum_{j=1}^M e_{ji} P(C_j | x). \quad (1.1)$$

This leads to a reasonable performance measure for the classifier, i.e., the expected loss, defined as

$$\mathcal{L} = \int R(C(x) | x) dP(x) \quad (1.2)$$

where $C(x)$ represents the classifier's decision (assuming one of the M "values," C_1, C_2, \dots, C_M), based on a random observation x drawn from a probability distribution $P(x)$. The decision function, $C(x)$, depends on the classifier design. Obviously, if the classifier is so designed that for every x

$$R(C(x) | x) = \min_i R(C_i | x), \quad (1.3)$$

the expected loss in equation (1.2) will be minimized.

For many applications, including speech recognition, the loss function e_{ij} is usually chosen to be the zero-one loss function defined by

$$e_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \quad i, j = 1, 2, \dots, M \quad (1.4)$$

which assigns no loss to correct classification and a unit loss to any error, regardless of the class. With this type of loss function, the expected loss \mathcal{L} is thus the error probability of classification or recognition. The conditional loss becomes

$$R(C_i | x) = \sum_{i \neq j} P(C_j | x) = 1 - P(C_i | X). \quad (1.5)$$

The optimal classifier that achieves minimum \mathcal{L} is thus the one that implements the following:

$$C(x) = C_i \quad \text{if} \quad P(C_i | x) = \max_j P(C_j | x). \quad (1.6)$$

For minimum error rate classification, the classifier employs the decision rule of (1.6) which is called the “maximum a posterior” (MAP) decision. The minimum error rate achieved by MAP decision is called “Bayes risk”. When all posterior probabilities are known, the classifier based on MAP rule is an optimal classifier based on the Bayes decision theory. However, if these probabilities are not known or the decision rule is not based on the class posterior probability, then we cannot use this result directly.

In practice, these probabilities have to be estimated from a training data set with known class labels. The classical Bayes decision theory thus effectively transforms the classifier design problem into a distribution estimation problem. This is the basis of the Bayesian statistical approach to pattern recognition which can be stated as: given (or collect) a set of training data (observations) $\{x_1, x_2, \dots, x_K\}$ with known class labels, estimate the a posterior probabilities $P(C_i | x)$, $i = 1, 2, \dots, M$ for any x to implement the maximum a posterior decision for minimum Bayes risk. The a posterior probability $P(C_i | x)$ can be rewritten as

$$P(C_i | x) = P(x | C_i)P(C_i)/P(x) . \quad (1.7)$$

Since $P(x)$ is not a function of the class index and thus has no effect in the MAP decision, the needed probabilistic knowledge can be represented by the class prior $P(C_i)$ and the conditional probability $P(x | C_i)$.

There are several issues associated with this classical approach. First, the distributions usually have to be parameterized in order for them to be practically useful for the implementation of the MAP rule of (1.6). The classifier designer therefore has to determine the right parametric form of the distributions. For most of the real world problems, this is a difficult task. Our choice of the distribution form is often limited by the mathematical tractability of the particular distribution functions and is very likely to be inconsistent with the actual distribution. This means that the true MAP decision can rarely be implemented and the minimum Bayes risk generally remains an unachievable lower bound. Second, given a parameterized distribution form, the unknown parameters defining the distribution have to be estimated from a finite amount of labeled training data, requiring that the estimation method has to be able to produce consistent parameter values when the size of the training samples varies. Third, it requires a training data set of sufficient size in order to have reliable parameter estimates. But in practice and for speech and language processing in particular, training data are always sparse compared to all possible realizations and variations in human speech and language. These three basic issues point out a fundamental fact; that is, despite the conceptual optimality of the Bayes decision theory and its applications to pattern recognition, it cannot always be accomplished in practice, because most practical “MAP” decisions in speech and language processing are not true MAP decisions. This understanding is critical for the discussion that follows.