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 Biing 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

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

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

CONTENTS

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 classier design, has been studied in many areas. In particular, Tsypkin [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, $i = 1, 2, \ldots, M$. 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 $P(C_j \mid x)$ 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]

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

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

$$\mathcal{L} = \int R(C(x) \mid x) dP(x) \tag{1.2}$$

where $C(x)$ represents the classifier's decision (assuming one of the M "values," C_1, $C_2, \ldots, C_M$), 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

$$R(C(x) \mid x) = \min_i \; R(C_i \mid x) \; , \tag{1.3}$$

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

$$e_{ij} = \begin{cases} 0, \; i = j \\ 1, \; i \neq j \end{cases} \quad i, j = 1, 2, \ldots, M \tag{1.4}$$

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

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

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

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

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) $\{x_1, x_2, \ldots, x_K\}$ with known class labels, estimate the a posterior probabilities $P(C_i \mid x)$, $i = 1, 2, \ldots, M$ for any x to implement the maximum a posterior decision for minimum Bayes risk. The a posterior probability $P(C_i \mid x)$ can be rewritten as

$$P(C_i \mid x) = P(x \mid C_i)P(C_i)/P(x) \; . \tag{1.7}$$

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 $P(x \mid C_i)$.

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.