X.D. HUANG, Y. ARIKI, M.A. JACK

HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

EDINBURGH UNIVERSITY PRESS

Edinburgh University Press
22 George Square, Edinburgh

Set by the University of Edinburgh
and printed in Great Britain by
Redwood Press Limited
Melksham, Wilts

British Library Cataloguing
in Publication Data

Huang, X. D.
Hidden Markov models for
speech recognition.—(Edinburgh
information technology series; 7)
1. Speech. Recognition by
computer systems
I. Title II. Ariki, Y.
III. Jack, Mervyn A. IV. Series
006.454

ISBN 0 7486 0162 7

Edinburgh
Information Technology Series

HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

S. Michaelson and M. Steedman
Series Editors

PREFACE

The theory of hidden Markov models was first developed in the mid 1960s by Baum and Welch. Applications of hidden Markov models to automatic speech recognition became a research topic in the 1970s in the pioneering work of Baker, Jelinek and others. The theory has subsequently been successfully applied in many state-of-the-art speech recognition systems. In the 1980s, there has been a dramatic increase in the application of hidden Markov models, not only for speech recognition, but also for many other areas. We have been involved for several years in the development of hidden Markov models for speech recognition and we believe that an appropriate textbook on hidden Markov modelling will greatly help postgraduate students in Electrical Engineering or Computer Science.

This book is primarily concerned with basic theories in hidden Markov modelling. However, some essential results of general pattern recognition and speech processing are included to help readers understand and apply the hidden Markov model for speech recognition. The main body of this book is devoted to the unified treatment of conventional vector quantisation, discrete hidden Markov models, and continuous hidden Markov models. We have presented an extensive discussion of Q-functions that are crucial in using and understanding the theory. We have also devoted many pages to practical issues in hidden Markov modelling. Finally, experimental examples are included to demonstrate how the theory is applied in practice. We hope such a treatment will be useful to both beginners as an introductory book and experts as a reference book.

In writing this book, the authors have had the benefit of advice from many people. We gratefully acknowledge the help of Ditang Fang, Hsiao-Wuen Hon, John Laver, Kai-Fu Lee, Fergus McInnes and our wives and families.

Xuedong D. Huang
Yasuo Ariki
Mervyn A. Jack

CONTENTS

CHAPTER ONE

INTRODUCTION

Speech has evolved over many centuries to achieve today's rich and elaborate form. Human communications are today dominated by spoken language, whether face to face, over the telephone, or through television and radio. In direct contrast, human–machine (computer) interaction is largely dependent on keyboard strokes, or other mechanical means. As such, this interaction mode demands skill development by individuals, and presents a barrier to widespread use of computer systems. Consequently, the goal of overcoming this barrier by building machines that understand spoken language has attracted the attention of scientists over the past 50 years. A spoken language understanding interface would be invaluable since speech communication is a natural and efficient mode for the human user. Example applications include automatic dictation (especially for Chinese and Japanese), database query (such as airline reservations), command and control, and computer-assisted instruction. Achievement of this spoken language understanding demands integration of both speech processing and natural language processing. One of the key problems is automatic speech recognition. The task of a speech recognition system is to take, as input, the acoustic waveform produced by the speaker and to produce, as output, a sequence of linguistic words corresponding to the input utterance.

Many uncertainties exist in speech recognition. The uncertainty associated with words that have been spoken to a speech recognition system is compounded by the acoustic uncertainty of the different accents and speaking styles of

individual speakers; by the lexical and grammatical uncertainty of the language the speaker uses; and by the semantic uncertainty of the subject the speaker wants to talk about. The speaker may inquire about flights to Beijing, or may reserve a ticket to Edinburgh, or may even be dictating an article in Chinese. Acoustic uncertainty has many components, such as the general quality of a speaker's voice; speaking speed and loudness; accent; and unusual speaking conditions such as illness or stress. In addition, acoustic contaminants such as room noise or competing speakers constitute a problem. A successful speech recognition system must take into account all of these acoustic uncertainties. Lexical, syntactic, and semantic knowledge must then be applied in a manner that permits cooperative interaction among the various levels of acoustic, phonetic and linguistic knowledge in minimising the uncertainty. However, when compared with human performance, only very restricted speech can currently be used in existing speech recognition systems. The principle constraints include:

(1) speaker dependence rather than speaker independence,

(2) isolated word input rather than continuous speech operation,

(3) limited rather than extensive vocabulary, and

(4) artificial grammar rather than natural language.

Scientists with backgrounds in signal processing, pattern recognition, artificial intelligence, linguistics, statistics, information theory, and psychology have been attacking the many problems of speech recognition. Their efforts can be broadly classified into the following.

(1) Modelling of the speech signal and its variabilities to facilitate efficient information extraction. These variabilities include phonetic and linguistic effects, inter-speaker and intra-speaker variabilities, and environmental acoustic variabilities.

(2) Automatic acquisition and modelling of linguistic events (lexicon, syntax, semantics, discourse, pragmatics, and task structure).

(3) Developing human factors methods for the design of an effective user interface.

Research in speech recognition has followed two primary routes: those adopting a knowledge-based approach, and those adopting a statistically data-based approach. Knowledge-based approaches to speech recognition and understanding [10] have attempted to express human knowledge of speech in terms of acoustic-phonetic rules based on specified *features* of the acoustic waveform. For these approaches, the acoustic signal is usually first segmented and labelled into phoneme-like units, and the resulting phoneme string is used for lexical and syntactic analysis. Words in the lexicon are represented in terms of phonemic spellings, and syntax is usually described by conventional linguistic means. Knowledge is represented in computer programs created by linguistic and phonetic experts [5,15]. It is known that human experts can be trained to *read* speech spectra, which supports the proposition that distinct features exist in the speech spectrum [15]. Machine realisation of this human ability is however currently far poorer than the well-trained human expert. In addition to the absence of good understanding of the human auditory mechanism, this knowledge-based approach is also constrained by the inability of human experts to formalise completely their knowledge. Totally reliable features are required to represent speech signals, before acoustic segmentation, phonetic labelling and lexical decoding can be carried out with any degree of accuracy. Formants are considered to be one of the most important features in speech recognition, and various methods have been developed to track formants from speech signals. None of this work to date has achieved the required accuracy for speech recognition and it can be argued that, even if an excellent formant tracker were available, *a priori* knowledge

would still be needed to indicate phonetic context for formant tracking. However, without good feature representation (of formants etc.), it is extremely difficult to obtain the necessary *a priori* phonetic knowledge based on these features. Thus some sophisticated and interactive formant tracking algorithms are necessary in order to obtain reliable formant estimation. This remains an unsolved problem for the knowledge-based approach. It should be noted, however, that the knowledge-based approach remains an important research area [2,7].

In contrast to the knowledge-based approach, alternative *data-based* statistical approaches have achieved considerable success. These are usually based on modelling the speech signal itself by some well-defined statistical algorithms that can automatically extract *knowledge* from speech data. This book will focus on the alternative statistical approaches, and the knowledge-based approach will not be considered further in this work. The predominant class of these algorithms is the hidden Markov model (HMM) [1,4,9,11]. An HMM-based speech system depends on three key factors:

(1) a detailed modelling scheme which is capable of accommodating various uncertainties,

(2) access to sufficient speech training data, and

(3) an automatic learning algorithm to improve the recognition accuracy.

By using HMMs, the speech signal variability in parameter space and time can be modelled effectively. Unlike the knowledge-based approach, the HMM learning procedure is achieved by presenting speech data to HMMs and automatically improving the models by data as opposed to some heuristic rules presented by human experts. In general, the more data presented to the model, the higher the recognition accuracy achieved. Motivated by neural network research, improvements can also be obtained by incorporating classification into parameter estimation [3,6].

HMM methods have presented speech recognition with a solid theoretical basis, and have resulted in significant advances in large-vocabulary continuous speaker-independent speech recognition [11].

The HMM can be based on either discrete output probability distributions or continuous output probability density functions, which are very important to acoustic modelling. Both the discrete HMM and the continuous HMM are widely used in state-of-the-art speech recognition systems [1,6,11,12,14]. For the discrete HMM, vector quantisation (VQ) makes it possible to use a non-parametric, discrete output probability distribution to model the observed speech signals. The objective of VQ is to find the set of reproduction vectors, or codebook, that represents an information source with minimum expected distortion. Under the discrete HMM framework, VQ is first used to obtain discrete output symbols. The discrete HMM then models observed discrete symbols. In contrast, the continuous mixture HMM [13] uses continuous mixture probability density functions to model speech parameters directly without using VQ, and usually needs extensive training data and computation times.

On the other hand, the semi-continuous HMM [8], which is a very general model including both discrete and continuous mixture HMMs as its special forms, unifies VQ, the discrete HMM, and the continuous mixture HMM. Based on the assumption that each VQ codeword can be represented by a continuous probability density function, the semi-continuous output probability is then a combination of *discrete* model-dependent weighting coefficients with these *continuous* codebook probability density functions. In comparison with the conventional continuous mixture HMM, the semi-continuous HMM can offer the modelling ability of large-mixture probability density functions. In addition, the number of free parameters and the computational complexity can be reduced because all of the probability density functions are tied together in the codebook. The semi-continuous hidden Markov model thus provides a good

solution to the conflict between detailed acoustic modelling and insufficient training data. In comparison with the conventional discrete HMM, robustness can be enhanced by using multiple codewords in deriving the semi-continuous output probability; and the VQ codebook itself can be optimised together with the HMM parameters in terms of the maximum likelihood criterion. Unified modelling can substantially minimise the information lost in conventional VQ and therefore leads to better performance than both the discrete HMM and the continuous mixture HMM.

This book will introduce the necessary mathematical background to understand the theory of HMMs; present a complete theory of hidden Markov modelling in depth and scope; and offer practical guidance for the use of both fundamental and advanced HMM technologies in speech recognition problems; in particular, acoustic modelling problems.

1.1. Book Organisation

Throughout the book, unless explicitly noted otherwise, the discrete probability of finite symbols O will be denoted by $Pr(O)$; and the continuous probability density function for the continuous observations $\mathbf{x}$ will be denoted by $f(\mathbf{x})$. Fundamentals of probability and pattern recognition theories involved in speech recognition will be reviewed and discussed in Chapters 2 and 3.

Chapter 4 describes VQ as a special pattern recognition technique that has been widely used in speech processing, coding and recognition. The modelling method can be viewed as a problem of estimating parameters for a family of continuous mixture probability density functions, which pave the way for the unified modelling approach of the VQ and HMM.

Mathematical principles of the HMM and related techniques for speech recognition are described in Chapter 5. This chapter is the foundation of the statistical modelling tool, the HMM, which will be discussed throughout the book. Chapter 6 describes continuous HMMs, which parallel discrete HMMs. The continuous mixture HMM is discussed in detail, since it is strongly related to the semi-continuous HMM. Chapters 2–6 represent the theoretic foundation to the semi-continuous HMM.

The semi-continuous HMM is presented in Chapter 7. It offers modelling power similar to the continuous mixture HMM with a large number of mixture density functions, while demanding much lower computational complexity than the continuous mixture HMM. In addition, the semi-continuous output probability density function can be well smoothed in comparison with the discrete HMM. From the discrete HMM point of view, the semi-continuous HMM can minimise the information lost in VQ. From the continuous mixture HMM point of view, the semi-continuous HMM can reduce the number of free parameters and computational complexity by tying continuous density functions. The unified theory of VQ and hidden Markov modelling, which are heavily relevant to the discussion in Chapters 5 and 6, are highlighted.

Chapter 8 discusses issues for designing a speech recognition system using HMMs. Topics such as choice of modelling unit, use of smoothing techniques, re-estimation criteria, and multiple features are included.

Chapter 9 presents experimental examples in several typical speech recognition systems. Implementational issues are discussed. C programs are included as examples. Relationships among continuous HMMs, discrete HMMs, and semi-continuous HMMs are highlighted.