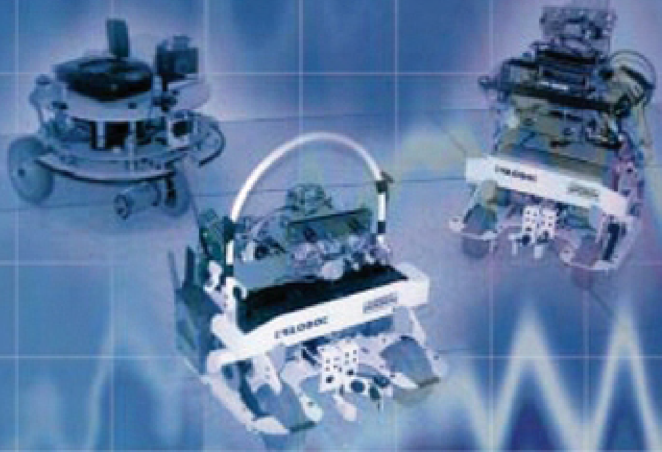

Mathematical Models for Speech Technology

Stephen E. Levinson

WILEY

Mathematical Models for Speech Technology

Stephen E. Levinson
University of Illinois at Urbana-Champaign, USA

John Wiley & Sons, Ltd

Mathematical Models for Speech Technology

Mathematical Models for Speech Technology

Stephen E. Levinson
University of Illinois at Urbana-Champaign, USA

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Levinson, Stephen C.
Mathematical models for speech technology / Stephen Levinson.
p. cm.
Includes bibliographical references and index.
ISBN 0-470-84407-8 (cloth)
1. Speech processing systems. 2. Computational linguistics. 3. Applied linguistics–Mathematics. 4. Stochastic processes. 5. Knowledge, Theory of. I. Title.
TK7882.S65L48 2005
006.4′54′015118–dc22
2004026215

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-84407-8

Typeset in 10/12 Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

To my parents
Doris R. Levinson
and
Benjamin A. Levinson

Contents

Preface

Evolution

This monograph was written over the past four years to serve as a text for advanced graduate students in electrical engineering interested in the techniques of automatic speech recognition and text to speech synthesis. However, the book evolved over a considerably longer period for a significantly broader purpose. Since 1972, I have sought to demonstrate how mathematical analysis captures and illuminates the phenomena of language and mind.

The first draft was written in 1975 during my tenure as a J. Willard Gibbs instructor at Yale University. The manuscript grew out of my lecture notes for a graduate course in pattern recognition, the main component of which was a statistical approach to the recognition of acoustic patterns in speech. The connection to language and mind was the result of both incorporating syntactic and semantic information into the statistical decision-theoretic process and observing that the detection and identification of patterns is fundamental to perception and intelligence.

The incomplete manuscript was set aside until 1983, at which time an opportunity to resurrect it appeared in the guise of a visiting fellowship in the Engineering Department of Cambridge University. A revised draft was written from lecture notes prepared for another course in pattern recognition for third-year engineering students. This time, topics of syntax and semantics were augmented with several other aspects of linguistic structure and were encompassed by the notion of composite pattern recognition as the classification of complicated patterns composed of a multi-leveled hierarchy of smaller and simpler ones. This second draft also included a brief intellectual history of the collection of ideas designated by what I will later argue is an unfortunate name, artificial intelligence (AI), and a recognition of its role in speech.

Once again the manuscript was set aside until the occasion of my appointment to the Department of Electrical and Computer Engineering at the University of Illinois. In 1997 I began organizing a program of graduate study in speech signal processing that would include both instruction in the existing technology and research to advance it. In its present form, the program comprises three tightly integrated parts: a course devoted to speech as an acoustic signal, another course on the linguistic structure of the acoustic signal, and research directed at automatic language acquisition. The first course required little innovation as there are several texts that provide standard and comprehensive coverage of the material. This book is a modification of my long-dormant manuscript and is now the basis for both the second course covering mathematical models of linguistic structure and the research project studying automatic language acquisition.

Goals and Methods

Linguists, electrical engineers, and psychologists have collectively contributed to our knowledge of speech communication. In recognition of the interdisciplinary nature of the subject, this book is written so that it may be construed as either a mathematical theory of language or an introduction to the technologies of speech recognition and synthesis. This is appropriate since the speech technologies rest on psycholinguistic concepts of the modularity of the human language engine. On the other hand, the models and techniques developed by electrical engineers can quite properly be regarded as the single most comprehensive collection of linguistic knowledge ever assembled. Moreover, linguistic theories can only be applied and tested by embedding them in a mathematically rational and computationally tractable framework. However, mathematical and computational models are useful only to the extent that they capture the essential structure and function of the language engine.

To the best of my knowledge, no single text previously existed that both covers all of the relevant material in a coherent framework and presents it in such a multidisciplinary spirit. A course of this nature could, heretofore, have been taught only by using several texts and a collection of old scholarly papers published in a variety of journals. Moreover, when significant portions of the material have been included in books on speech processing, they have been, without exception, presented as immutable canon of the subject. The unpleasant fact is that while modern speech technology is a triumph of engineering, it falls far short of constructing a machine that is able to use natural spoken language in a manner even approaching normal human facility. There is, at present, only an incomplete science of speech communication supporting a correspondingly limited technology. Based on the assumption that the shortcomings of our technology are the consequence of gaps in our knowledge rather than pervasive error, it does not seem unreasonable to examine our current knowledge with an eye toward extracting some general principles, thereby providing students with the background required to read the existing literature critically and to forge a strategy for research in the field that includes both incremental improvements and revolutionary ideas. Sadly, the recent literature is almost exclusively about technical refinements.

There are several specific pedagogic techniques I have adopted to foster this perspective. Discussions of all of the mathematical models of linguistic structure include their historical contexts, their underlying early intuitions and the mechanisms by which they capture the essential features of the phenomena they are intended to represent. Wherever possible, it is shown how these models draw upon results from related disciplines. Since topics as diverse as acoustics and semantics are included, careful attention has been paid to reconciling the perspectives of the different disciplines, to unifying the formalisms, and to using coherent nomenclature.

Another guiding principle of this presentation is to emphasize the meaningful similarities and relationships among the mathematical models in preference to their obvious but superficially common features. For example, not all models that have state spaces or use dynamic programming to explore them serve identical purposes, even if they admit of identical formal descriptions. Conversely, there are some obscure but significant similarities amongst seemingly disparate models. For example, hidden Markov models and stochastic formal grammars are quite different formally yet are similar in the important sense that they both have an observable process to account for measurements and an underlying but hidden process to account for structure.

Finally, students should know what the important open questions in the field are. The orientation of this book makes it possible to discuss explicitly some of the current theoretical debates. In particular, most current research is aimed at transcribing speech into text without any regard for comprehension of the message. At the very least, this distorts the process by placing undue emphasis on word recognition accuracy and ignoring the more fundamental roles of syntax and semantics in message comprehension. At worst, it may not even be possible to obtain an accurate transcription without understanding the message. Another mystery concerns the relative importance of perceptual and cognitive processes. Informed opinion has vacillated from one extreme to the other and back again. There is still no agreement, as different arguments are often organized along disciplinary boundaries.

When this book is used as a text for a graduate course on speech technology, Chapters 1 and 2 should be considered a review of a prerequisite course on speech signal processing. Chapters 3 through 8 contain the technical core of the course and Chapters 9 and 10 place the material in its scientific and philosophical context. These last two chapters are also intended as guidance and motivation for independent study by advanced students.

Whereas a technical synopsis of the contents of this book is given in Chapter 1, here I shall analyze it in a more didactic manner. The prerequisite material covered in Chapter 2 comprises succinct if standard presentations of the physics of speech generation by the vocal apparatus, methods of spectral analysis, methods of statistical pattern recognition for acoustic/phonetic perception, and a traditional taxonomy of linguistic structure. From these discussions we extract a few themes that will appear frequently in the succeeding chapters. First, the speech signal is a non-stationary time–frequency distribution of energy. This both motivates the importance of the short-duration amplitude spectrum for encoding the intelligence carried by the signal and justifies the use of the spectrogram which is shown to be an optimal representation in a well-defined sense. Linear prediction is seen as a particularly useful spectral parameterization because of its close relationship to the geometry and physics of the vocal tract.

Second, speech is literate. Thus, the spectral information must encode a small finite alphabet of symbols, the sequencing of which is governed by a hierarchy of linguistic rules. It follows, then, that any useful analysis of the speech signal must account for the representation of structured sequences of discrete symbols by continuous, noisy measurements of a multivariate, non-stationary function of time. This is best accomplished using non-parametric methods of statistical pattern recognition that employ a topological metric as a measure of perceptual dissimilarity. These techniques not only are optimal in the sense of minimum error, but also provide a justification for the direct normalization of time scales to define a metric that is invariant with respect to changes of time scale in signal.

The next six chapters are devoted to a detailed examination of techniques that address precisely these unique properties of the speech signal and, in so doing, capture linguistic structure. We begin with a study of probabilistic functions of a Markov process. Often referred to in the literature as hidden Markov models (HMMs), they have become a ubiquitous yet often seriously misunderstood mathematical object. The HMM owes its widespread application to the existence of a class of techniques for robust estimation of its parameters from large collections of data. The true value of the HMM, however, lies not in its computational simplicity but rather in its representational power. Not only does it intrinsically capture non-stationarity and the transformation of continuous measurements into discrete symbols, it also provides a natural way to represent acoustic phonetics, phonology, phonotactics, and even prosody.

In this book we develop the mathematical theory incrementally, beginning with the simple quantized observation case. We include a standard proof of Baum's algorithm for this case. The proof rests on the convexity of the log-likelihood function and is somewhat opaque, providing little insight into the reestimation formulas. However, by relating the parameter estimation problem for HMMs to the classical theory of constrained optimization, we are able to give a novel, short, and intuitively appealing geometric proof showing that the reestimation formulas work by computing a finite step in a direction that has a positive projection on the gradient of the likelihood function. We then progress to models of increasing complexity, including the little-known cases of non-stationary observation distributions and semi-Markov processes with continuous probability density functions for state duration.

We end the presentation of Chapter 3 with an account of two seminal but often overlooked experiments demonstrating the remarkable power of the HMM to discover and represent linguistic structure in both text and speech. The Cave–Neuwirth and Poritz experiments are then contrasted with the common formulation based on the special case of the non-ergodic HMM as a means of treating piecewise stationarity.

As powerful and versatile as it is, the HMM is not the only nor necessarily the best way to capture linguistic structure. We continue, therefore, with a treatment of formal grammars in the Chomsky hierarchy and their stochastic counterparts. The latter are seen to be probabilistic functions of an unobservable stochastic process with some similarities to the HMM. For example, we observe that the right linear grammar is equivalent to the discrete symbol HMM. However, the more complex grammars provide greater parsimony for fixed representational power. In particular, they provide a natural way to model the phonology and syntax of natural language.

Based on these formalisms, Chapter 4 approaches the problem of parsing, that is, determining the syntactic structure of a sentence with respect to a given grammar. Despite its central role in linguistics, this problem is usually ignored in the speech processing literature because it is usually assumed that word order constraints are sufficient for transcription of an utterance and the underlying grammatical structure is superfluous. We prefer the position that transcription is only an intermediate goal along the way to extracting the meaning of the message, of which syntactic structure is a prerequisite. Later we advance the idea that, in fact, transcription without meaning is a highly error-prone process. Parsing a spoken utterance is beset by two sources of uncertainty, variability of the acoustic signal and ambiguity in the production rules of the grammar. Here we show that these uncertainties can be accounted for probabilistically in two complementary ways, assigning likelihoods to the words conditioned on the acoustic signal and placing fixed probabilities on the rules of the grammar. Both of these ideas can be efficiently utilized at the first two levels of the Chomsky hierarchy and, in fact, they may be combined. We develop probabilistic parsing algorithms based on the Dijkstra and Cocke–Kasami–Younger algorithms for the right linear and context-free cases, respectively.

In Chapter 5, we address the inverse of the parsing problem, that of grammatical inference. This is the problem of determining a grammar from a set of possibly well-formed sentences, the syntactic structure of which is not provided. This is a classical problem and is usually ignored by linguists as too difficult. In fact, the difficulty of this problem is regarded by strict Chomskians as proof that the human language engine is innate. We, however, treat the problem of grammatical inference as one simply of parameter estimation. We show that the reestimation formulas for the discrete symbol HMM and the little-known Baker algorithm for stochastic context-free grammars are actually grammatical inference algorithms. Once the stochastic grammars are estimated, their deterministic counterparts are easily constructed. Finally, we show how parsing algorithms can be used to provide the sufficient statistics required by the EM algorithm so that it may be applied to the inference problem.

Chapter 6 is a divertimento in which we reflect on some of the implications of our mathematical models of phonology, phonotactics, and syntax. We begin by recalling an instructive experiment of Miller *et al.* demonstrating quantitatively that human listeners use linguistic structure to disambiguate corrupted utterances. This phenomenon is widely interpreted in the speech literature to mean that the purpose of grammar is to impose constraints on word order and thereby reduce recognition error rates in the presence of noise or other naturally occurring variability in the speech signal. Moreover, this analysis of Miller is the unstated justification for ignoring the grammatical structure itself and using only word order for transcription.

The information-theoretic concept of entropy is correctly used in the literature on speech recognition as a measure of the uncertainty inherent in word order, leading to the intuition that recognition error rate rises with increasing entropy. Entropy is typically estimated by playing the Shannon game of sequential prediction of words from a statistical analysis of large corpora of text or phonetic transcriptions thereof. Here we take a unique approach showing how the entropy of a language can be directly calculated from a formal specification of its grammar. Of course, entropy is a statistical property most easily obtained if the grammar is stochastic. However, we show that entropy can be obtained from a deterministic grammar simply by making some weak assumptions about the distributions of sentences in the language. Taking this surprising result one step further, we derive from the Fano bound a quantitative relationship among the entropy of a language, the variability intrinsic to speech, and the recognition error rate. This result may be used to explain how grammar serves as the error-correcting code of natural language.

All of the foregoing material is unified in Chapter 7 into a constructive theory of language or, from the engineer's perspective, the design of a speech recognition machine. We discuss two basic architectures, one integrated, the other modular. The latter approach is inspired by psycholinguistic models of human language processing and depends crucially on the Cave–Neuwirth and Poritz experiments featured in Chapter 3. We note the use of the semi-Markov model to represent aspects of prosody, phonotactics, and phonology. We also demonstrate the ability of the modular system to cope with words not contained in its lexicon.

In evaluating the performance of these systems, we observe that their ability to transcribe speech into text without regard for the meaning of the message arguably exceeds human performance on similar tasks such as recognizing fluent speech in an unknown language. And yet, this remarkable achievement does not provide speech recognition machines with anything remotely like human linguistic competence.

It seems quite natural, then, to try to improve the performance of our machines by providing them with some method for extracting the meaning of an utterance. On the rare occasions when this idea is discussed in the literature, it is often inverted so that the purpose of semantic analysis becomes simply that of improving word recognition accuracy. Of course, this is a very narrow view of human linguistic behavior. Humans use language to convey meaningful messages to each other. Linguistic competence consists in the ability to express meaning reliably, not to simply obtain faithful lexical transcriptions. It is in this ability to communicate that our machines fail. Chapter 8, therefore, is devoted to augmenting the grammatical model with a semantic one and linking them in a cooperative way.

We begin with a description of a laboratory prototype for a speech understanding system. Such a system should not simply be a transcription engine followed by a text processing semantic module. We note that such a system would require two separate syntax analyzers. Whereas, if the parsing algorithms described in Chapter 4 are used, the requisite syntactic structure is derived at the same time that the word order constraints are applied to reduce the lexical transcription error rate.

The most straightforward approach is to base the understanding system on the simplified semantics of a carefully circumscribed subset of natural language. Such formal artificial languages bear a strong resemblance to programming languages and can be analyzed using compiler techniques. Such systems may be made to carry out dialogs of limited scope with humans. However, the communication process is quite restricted and brittle. Extension of the technique to another domain of discourse is time-consuming because little if any data can be reused.

What is required to enable the machine to converse in colloquial discourse is a generalized model of unrestricted semantics. There are many such models, but they all reduce to mathematical logic or searching labeled, directed graphs. The former rests on the intuition that the extraction of meaning is equivalent to the derivation of logically true statements about reality, said statements being expressed formally in first-order logic. The latter model rests on the intuition that meaning emerges out of the properties of and relationships among objects and actions and can be extracted by finding suitable paths in an abstract graph. Such ideas have yet to be applied to speech processing. Thus, Chapter 8 concludes in an unsatisfying manner in that it provides neither theoretical nor empirical validation of a model of semantic analysis.

Up to this juncture, the exposition is presented in the customary, turgid scientific style. The mathematics and its application are objective and factual. No personal opinions regarding their significance are advanced. For Chapters 9 and 10, that conservatism is largely discarded as the unfinished work of the first eight chapters deposits us directly on the threshold of some of the very deepest and most vociferously debated ideas in the Western philosophical tradition. We are forced to confront the question of what sort of theory would support the construction of a machine with a human language faculty and we are obliged to assess the role of our present knowledge in such a theory. This profound shift of purpose must be emphasized. In the two concluding chapters, then, the mathematics nearly vanishes and, to preserve some semblance of intellectual responsibility, I employ the first person singular verb form.

It is my strongly held belief that a simulation of the human language engine requires nothing less than the construction of a complete human mind. Although this goal has proved to be utterly elusive, I insist that there is no inherent reason why it cannot be accomplished. There is, however, a cogent reason for our quandary revealed by a critical review of the intellectual history of AI.

In a remarkable work entitled *Fin-de-Siècle Vienna: Politics and Culture*, Carl E. Schorske gives a highly instructive explanation for the unkept promises of AI. He convincingly argues that cultural endeavors stagnate and fail when they become ahistorical by losing contact with both their diachronic history (i.e. their intellectual antecedents) and their synchronic history (i.e. their connections to independently developed but related ideas), and become fixated in the technical details of contemporary thought. Although Schorske did not include science in his analysis, his thesis seems highly appropriate there, too, with AI as a striking instance. Specifically, the loss of history in rapid response to an overwhelming but narrow discovery is made manifest by comparing the work of Norbert Wiener and Alan Turing.

The first edition of Wiener's *Cybernetics* was published in 1948, the very year that ENIAC, the first electronic, stored program, digital computer became operational. From this very early vantage point, Wiener has a fully diachronic perspective and recognizes that from ancient times to the present, metaphors for mind have always been expressed in the high technology of the day. Yet he clearly sees that the emerging computer offers a powerful tool with which to study and simulate, information and control in machines and organisms alike.

By 1950, Turing, on the other hand, had developed a deep understanding of the implications of his prior work in the foundations of mathematics for theories of mind. Since the Universal Turing Machine, and, hence, its reification in the form of the digital computer, is capable of performing almost any symbolic manipulation process, it is assumed sufficient for creating a mental model of the real world of our everyday experience. This intuition has evolved into what we today refer to as the "strong theory of AI". It is an almost exclusively contemporary view and was, in fact, Turing's preferred interpretation of thought as a purely abstract symbolic process. There is, however, a historical aspect to the remarkable 1950 paper. This is not surprising since the ideas it expresses date from the mid-1930s, at which time the metaphors for mind derived from classical electromechanical devices. In the penultimate paragraph of the paper, Turing offers an astounding and often overlooked alternative to the technical model of thought as symbolic logic. He suggests that the symbols and the relations among them could be inferred from real-world sensory data, a cybernetic and hence, historical view.

Unfortunately, the next generation of thinkers following Wiener and Turing fully endorsed the mind–software identity and en route lost all semblance of the historical trajectory. Based on my interpretation of Schorske, I submit that there have been no conceptual advances since then in the AI tradition. There has been some technical progress but no enlightenment. This is a rather frustrating conclusion in light of the elegance of Turing's theory which seemed to promise the immediate construction of an indisputably mechanical mind.

The key to revitalizing research on the theory of mind lies in synthesizing the synchronic and diachronic histories in what I call the cybernetic paradigm. This presently unfashionable mode of interdisciplinary thought unifies Turing's and Wiener's work and