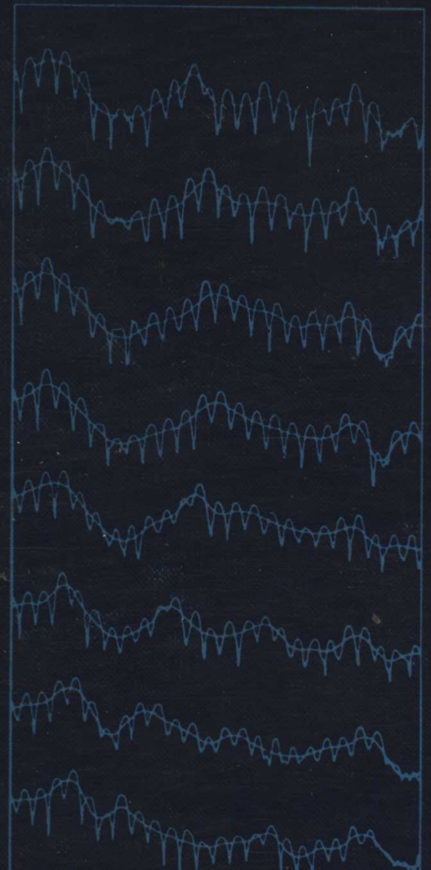

Digital Processing of Speech Signals

L.R. Rabiner / R.W. Schafer

PRENTICE-HALL
SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM,
SERIES EDITOR

DIGITAL PROCESSING OF SPEECH SIGNALS

Lawrence R. Rabiner

Acoustics Research Laboratory
Bell Telephone Laboratories
Murray Hill, New Jersey

Ronald W. Schafer

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia

Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging in Publication Data

Rabiner, Lawrence R (date)
Digital processing of speech signals.

(Prentice-Hall signal processing series)
Includes bibliographies and index.
1. Speech processing systems. 2. Digital electronics. I. Schafer, Ronald W., joint author.
II. Title.
TK7882.S65R3 621.38'0412 78-8555
ISBN 0-13-213603-1

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Prentice-Hall International, Inc., *London*
Prentice-Hall of Australia Pty., Limited, *Sydney*
Prentice-Hall of Canada, Ltd., *Toronto*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Prentice-Hall of Southeast Asia Pte. Ltd., *Singapore*
Whitehall Books Limited, *Wellington, New Zealand*

To our parents,

Dr. and Mrs. Nathan Rabiner

and

Mr. and Mrs. William Schafer,

for instilling within us the thirst for knowledge
and the quest for excellence;

and to our families,

Suzanne, Sheri, and Wendi Rabiner

and

Dorothy, Bill, John, and Kate Schafer,

for their love, encouragement, and support.

Preface

This book is an outgrowth of an association between the authors which started as fellow graduate students at MIT, was nurtured by a close collaboration at Bell Laboratories for slightly over 6 years, and has continued ever since as colleagues and close friends. The spark which ignited formal work on this book was a tutorial paper on digital representations of speech signals which we prepared for an IEEE Proceedings special issue on Digital Signal Processing, edited by Professor Alan Oppenheim of MIT. At the time we wrote that paper we realized that the field of digital speech processing had matured sufficiently that a book was warranted on the subject.

Once we convinced ourselves that we were both capable of and ready to write such a text, a fundamental question concerning organization had to be resolved. We considered at least 3 distinct ways of organizing such a text and the problem was deciding which, if any, would provide the most cohesive treatment of this field. The 3 organizations considered were

1. According to digital representations
2. According to parameter estimation problems
3. According to individual applications areas.

After much discussion it was felt that the most fundamental notions were those related to digital speech representations and that a sound understanding of such representations would allow the reader both to understand and to advance the methods and techniques for parameter estimation and for designing speech processing systems. Therefore, we have chosen to organize this book around several basic approaches to digital representations of speech signals, with discussions of specific parameter estimation techniques and applications serving as examples of the utility of each representation.

The formal organization of this book is as follows. Chapter 1 provides an introduction to the area of speech processing, and gives a brief discussion of application areas which are directly related to topics discussed throughout the book. Chapter 2 provides a brief review of the fundamentals of digital signal processing. It is expected that the reader has a solid understanding of linear systems and Fourier transforms and has taken, at least, an introductory course in digital signal processing. Chapter 2 is not meant to provide such background, but rather to establish a notation for discussing digital speech processing, and to provide the reader with handy access to the key equations of digital signal processing. In addition, this chapter provides an extensive discussion of sampling, and decimation and interpolation, key processes that are fundamental to most speech processing systems. Chapter 3 deals with digital models for the speech signal. This chapter discusses the physical basis for sound production in the vocal tract, and this leads to various types of digital models to approximate this process. In addition this chapter gives a brief introduction to acoustic phonetics; that is, a discussion of the sounds of speech and some of their physical properties.

Chapter 4 deals with time domain methods in speech processing. Included in this chapter are discussions of some fundamental ideas of digital speech processing—e.g., short-time energy, average magnitude, short-time average zero-crossing rate, and short-time autocorrelation. The chapter concludes with a section on a nonlinear smoothing technique which is especially appropriate for smoothing the time-domain measurements discussed in this chapter. Chapter 5 deals with the topic of direct digital representations of the speech waveform—i.e., waveform coders. In this chapter the ideas of instantaneous quantization (both uniform and nonuniform), adaptive quantization, differential quantization, and predictive coding (both fixed and adaptive) are discussed and are shown to form the basis of a variety of coders from simple pulse code modulation (PCM) to adaptive differential PCM (ADPCM) coding.

Chapter 6 is the first of two chapters that deal with spectral representations of speech. This chapter concerns the ideas behind short-time Fourier analysis and synthesis of speech. This area has traditionally been the one which has received most attention by speech researchers since some of the key speech processing systems, such as the sound spectograph and the channel vocoder, are directly related to the concepts discussed in this chapter. Here it is shown how a fairly general approach to speech spectral analysis and synthesis provides a framework for discussing a wide variety of speech processing systems, including those mentioned above. Chapter 7, the second chapter on spectral representations of speech, deals with the area of homomorphic speech processing. The idea behind homomorphic processing of speech is to transform the speech waveform (which is naturally represented as a convolution) to the frequency domain as a sum of terms which can be separated by ordinary linear filtering techniques. Techniques for carrying out this procedure are discussed in this chapter, as are several examples of applications of homomorphic speech processing.

Chapter 8 deals with the topic of linear predictive coding of speech. This representation is based upon a minimum mean-squared error approximation to the time-varying speech waveform, subject to an assumed linear system model of the speech signal. This method has been found to be a robust, reliable, and accurate method for representing speech signals for a wide variety of conditions.

The final chapter, Chapter 9, provides a discussion of several speech processing systems in the area of man-machine communication by voice. The purpose of this chapter is twofold: first, to give concrete examples of specific speech processing systems which are used in real world applications, and second, to show how the ideas developed throughout the book are applied in representative speech processing systems. The systems discussed in this chapter deal with computer voice response, speaker verification and identification, and speech recognition.

The material in this book is intended as a one-semester course in speech processing. To aid the teaching process, each chapter (from Chapter 2 to Chapter 8) contains a set of representative homework problems which are intended to reinforce the ideas discussed in each chapter. Successful completion of a reasonable percentage of these homework problems is essential for a good understanding of the mathematical and theoretical concepts of speech processing. However, as the reader will see, much of speech processing is, by its very nature, empirical. Thus, some "hands on" experience is essential to learning about digital speech processing. In teaching courses based on this book, we have found that a first order approximation to this experience can be obtained by assigning students a term project in one of the following three broad categories:

1. A literature survey and report
2. A hardware design project
3. A computer project

Some guidelines and lists of suggested topics for the three types of projects are given at the end of Chapter 9. We have found that these projects, although demanding, have been popular with our students. We strongly encourage other instructors to incorporate such projects into courses using this book.

Acknowledgements

Several people have had a significant impact, both directly and indirectly, on the material presented in this book. Our biggest debt of gratitude goes to Dr. James L. Flanagan, head of the Acoustics Research Department at Bell Laboratories. Jim has had the dual roles of supervisor and mentor for both authors of this book. For a number of years he has provided us with a model both for how to conduct research and how to report on the research in a meaningful way. His influence on both this book, and our respective careers, has been profound.

Other people with whom we have had the good fortune to collaborate and learn from include Dr. Ben Gold of MIT Lincoln Laboratory, Professor Alan Oppenheim of MIT, and Professor Kenneth Stevens of MIT. These men have served as our teachers and our colleagues and we are grateful to them for their guidance.

Colleagues who have been involved directly with the preparation of this book include Professor Peter Noll of Bremen University, who provided critical comments on Chapter 5, Dr. Ronald Crochiere of Bell Laboratories, who reviewed the entire first draft of this book, and Professor Tom Barnwell of Georgia Tech, who provided valuable comments and criticisms of the entire text. Mr. Gary Shaw carefully worked all the homework problems. His solutions have provided the basis for a solutions manual that is available to instructors. Messrs. J. M. Tribolet, D. Dlugos, P. Papamichalis, S. Gaglio, M. Richards, and L. Kizer provided valuable comments on the final draft of the book. Finally, we wish to thank the Bell Laboratories book review board for overseeing the production of the book, and Ms. Carmela Patuto who was primarily responsible for a superb job of typing of the text of this book throughout its many revisions. In addition we acknowledge the assistance of Ms. Penny Blaine, Jeanette Reinbold, Janie Evans, Nancy Kennell, and Chris Tillery in preparing early drafts of some of the chapters. The generous support of the John and Mary Franklin Foundation to one of us (RWS) is gratefully acknowledged. The authors also wish to acknowledge use of the phototypesetting services at Bell Laboratories on which the entire text of this book was set.

LAWRENCE R. RABINER and RONALD W. SCHAFER

Contents

1

Introduction

1.0 Purpose of This Book

The purpose of this book is to show how digital signal processing techniques can be applied in problems related to speech communication. Therefore, this introductory chapter is devoted to a general discussion of questions such as: what is the nature of the speech signal, how can digital signal processing techniques play a role in learning about the speech signal, and what are some of the important application areas of speech communication in which digital signal processing techniques have been used?

1.1 The Speech Signal

The purpose of speech is communication. There are several ways of characterizing the communications potential of speech. One highly quantitative approach is in terms of information theory ideas as introduced by Shannon [1]. According to information theory, speech can be represented in terms of its *message content,* or *information.* An alternative way of characterizing speech is in terms of the *signal* carrying the message information, i.e., the acoustic waveform. Although information theoretic ideas have played a major role in sophisticated communications systems, we shall see throughout this book that it is the speech representation based on the waveform, or some parametric model, which has been most useful in practical applications.

In considering the process of speech communication, it is helpful to begin by thinking of a message represented in some abstract form in the brain of the speaker. Through the complex process of producing speech, the information in that message is ultimately converted to an acoustic signal. The message information can be thought of as being represented in a number of different ways in the process of speech production. For example, the message information is first converted into a set of neural signals which control the articulatory mechanism (that is, the motions of the tongue, lips, vocal cords, etc.). The articulators move in response to these neural signals to perform a sequence of gestures, the end result of which is an acoustic waveform which contains the information in the original message.

The information that is communicated through speech is intrinsically of a discrete nature; i.e., it can be represented by a concatenation of elements from a finite set of symbols. The symbols from which every sound can be classified are called *phonemes.* Each language has its own distinctive set of phonemes, typically numbering between 30 and 50. For example, English can be represented by a set of around 42 phonemes. (See Chapter 3.)

A central concern of information theory is the rate at which information is conveyed. For speech a crude estimate of the information rate can be obtained by noting that physical limitations on the rate of motion of the articulators require that humans produce speech at an average rate of about 10 phonemes per second. If each phoneme is represented by a binary number, then a six-bit numerical code is more than sufficient to represent all of the phonemes of English. Assuming an average rate of 10 phonemes per second and neglecting any correlation between pairs of adjacent phonemes we get an estimate of 60 bits/sec for the average information rate of speech. In other words, the *written* equivalent of speech contains information equivalent to 60 bits/sec at normal speaking rates. Of course a lower bound on the "true" information content of speech is considerably higher than this rate. The above estimate does not take into account factors such as the identity and emotional state of the speaker, the rate of speaking, the loudness of the speech, etc.

In speech communication systems, the speech signal is transmitted, stored, and processed in many ways. Technical concerns lead to a wide variety of representations of the speech signal. In general, there are two major concerns in any system:

1. Preservation of the message content in the speech signal.
2. Representation of the speech signal in a form that is convenient for transmission or storage, or in a form that is flexible so that modifications may be made to the speech signal without seriously degrading the message content.

The representation of the speech signal must be such that the information content can easily be extracted by human listeners, or automatically by machine. Throughout this book we shall see that representations of the speech signal