Digital Processing of Speech Signals

L.R. Rabiner / R.W. Schafer

Digital Processing of Speech Signals

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

0-13-213603-1

PRENTICE-HALL SIGNAL PROCESSING SERIES

Alan V. Oppenheim, Editor

ANDREWS and HUNT *Digital Image Restoration*
BRIGHAM *The Fast Fourier Transform*
BURDIC *Underwater Acoustic System Analysis*
CASTLEMAN *Digital Image Processing*
CROCHIERE and RABINER *Multirate Digital Signal Processing*
DUDGEON and MERSEREAU *Multidimensional Digital Signal Processing*
HAMMING *Digital Filters, 2e*
HAYKIN, ED. *Array Signal Processing*
LEA, ED. *Trends in Speech Recognition*
LIM, ED. *Speech Enhancement*
MCCLELLAN and RADER *Number Theory in Digital Signal Processing*
OPPENHEIM, ED. *Applications of Digital Signal Processing*
OPPENHEIM, WILLSKY, with YOUNG *Signals and Systems*
OPPENHEIM and SCHAFER *Digital Signal Processing*
RABINER and GOLD *Theory and Applications of Digital Signal Processing*
RABINER and SCHAFER *Digital Processing of Speech Signals*
ROBINSON and TREITEL *Geophysical Signal Analysis*
TRIBOLET *Seismic Applications of Homomorphic Signal Processing*

AT&T

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

19 18 17 16 15 14

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.

Contents

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

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

(rather than message content) may require from 500 to upwards of 1 million bits per second. In the design and implementation of these representations, the methods of signal processing play a fundamental role.

1.2 Signal Processing

The general problem of information manipulation and processing is depicted in Figure 1.1. In the case of speech signals the human speaker is the information source. The measurement or observation is generally the acoustic waveform.

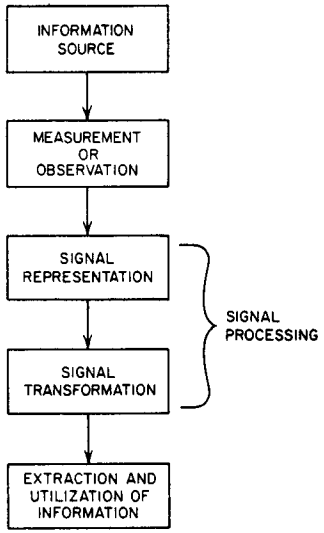


Fig. 1.1 General view of information manipulation and processing.

Signal processing involves first obtaining a representation of the signal based on a given model and then the application of some higher level transformation in order to put the signal into a more convenient form. The last step in the process is the extraction and utilization of the message information. This step may be performed either by human listeners or automatically by machines. By way of example, a system whose function is to automatically identify a speaker from a given set of speakers might use a time-dependent spectral representation of the speech signal. One possible signal transformation would be to average spectra across an entire sentence, compare the average spectrum to a stored averaged spectrum template for each possible speaker, and then based on a spectral similarity measurement choose the identity of the speaker. For this example the "information" in the signal is the identity of the speaker.

Thus, processing of speech signals generally involves two tasks. First, it is a vehicle for obtaining a general representation of a speech signal in either

waveform or parametric form. Second, signal processing serves the function of aiding in the process of transforming the signal representation into alternate forms which are less general in nature, but more appropriate to specific applications. Throughout this book we will see numerous specific examples of the importance of signal processing in the area of speech communication.

1.3 Digital Signal Processing

The focus of this book is to explore the role of digital techniques in processing speech signals. Digital signal processing is concerned both with obtaining discrete representations of signals, and with the theory, design, and implementation of numerical procedures for processing the discrete representation. The objectives in digital signal processing are identical to those in analog signal processing. Therefore, it is reasonable to ask why digital signal processing techniques should be singled out for special consideration in the context of speech communication. A number of very good reasons can be cited. First, and probably most important, is the fact that extremely sophisticated signal processing functions can be implemented using digital techniques. The algorithms that we shall describe in this book are intrinsically discrete-time, signal processing systems. For the most part, it is not appropriate to view these systems as approximations to analog systems. Indeed in many cases there is no realizable counterpart available with analog implementation.

Digital signal processing techniques were first applied in speech processing problems, as simulations of complex analog systems. The point of view initially was that analog systems could be simulated on a computer to avoid the necessity of building the system in order to experiment with choices of parameters and other design considerations. When digital simulations of analog systems were first applied, the computations required a great deal of time. For example, as much as an hour might have been required to process only a few seconds of speech. In the mid 1960's a revolution in digital signal processing occurred. The major catalysts were the development of faster computers and rapid advances in the theory of digital signal processing techniques. Thus, it became clear that digital signal processing systems had virtues far beyond their ability to simulate analog systems. Indeed the present attitude toward laboratory computer implementations of speech processing systems is to view them as exact simulations of a digital system that could be implemented either with special purpose digital hardware or with a dedicated computer system.

In addition to theoretical developments, concomitant developments in the area of digital hardware have led to further strengthening of the advantage of digital processing techniques over analog systems. Digital systems are reliable and very compact. Integrated circuit technology has advanced to a state where extremely complex systems can be implemented on a single chip. Logic speeds are fast enough so that the tremendous number of computations required in many signal processing functions can be implemented in real-time at speech sampling rates.

There are many other reasons for using digital techniques in speech communication systems. For example, if suitable coding is used, speech in digital form can be reliably transmitted over very noisy channels. Also, if the speech signal is in digital form it is identical to data of other forms. Thus a communications network can be used to transmit both speech and data with no need to distinguish between them except in the decoding. Also, with regard to transmission of voice signals requiring security, the digital representation has a distinct advantage over analog systems. For secrecy, the information bits can be scrambled in a manner which can ultimately be unscrambled at the receiver. For these and numerous other reasons digital techniques are being increasingly applied in speech communication problems [3].

1.4 Digital Speech Processing

In considering the application of digital signal processing techniques to speech communication problems, it is helpful to focus on three main topics: the representation of speech signals in digital form, the implementation of sophisticated processing techniques, and the classes of applications which rely heavily on digital processing.

The representation of speech signals in digital form is, of course, of fundamental concern. In this regard we are guided by the well-known sampling theorem [4] which states that a bandlimited signal can be represented by samples taken periodically in time – provided that the samples are taken at a high enough rate. Thus, the process of sampling underlies all of the theory and application of digital speech processing. There are many possibilities for discrete representations of speech signals. As shown in Figure 1.2, these representations can be classified into two broad groups, namely waveform representations and parametric representations. Waveform representations, as

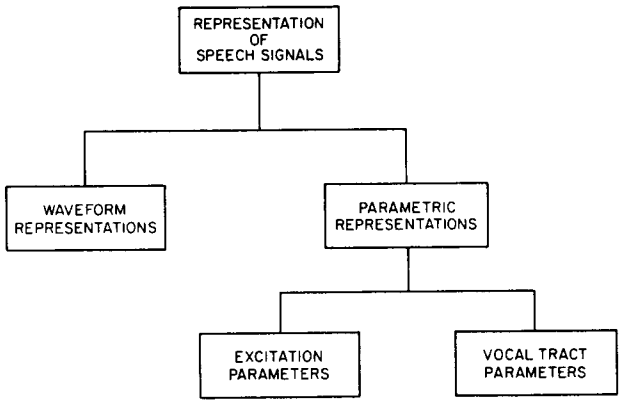


Fig. 1.2 Representations of speech signals.

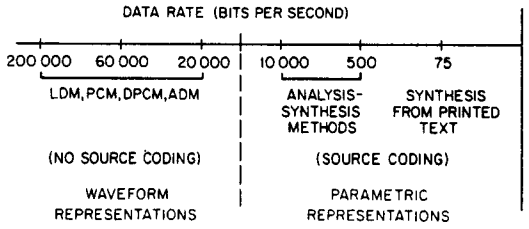


Fig. 1.3 Range of bit rates for various types of speech representations. (After Flanagan [3]).

the name implies, are concerned with simply preserving the "wave shape" of the analog speech signal through a sampling and quantization process. Parametric representations, on the other hand, are concerned with representing the speech signal as the output of a model for speech production. The first step in obtaining a parametric representation is often a digital waveform representation; that is, the speech signal is sampled and quantized and then further processed to obtain the parameters of the model for speech production. The parameters of this model are conveniently classified as either excitation parameters (i.e., related to the source of speech sounds) or vocal tract response parameters (i.e., related to the individual speech sounds).[1]

Figure 1.3 shows a comparison of a number of different representations of speech signals according to the data rate required. The dotted line, at a data rate of about 15,000 bits per second, separates the high data rate waveform representations at the left from the lower data rate parametric representations at the right. This figure shows variations in data rate from 75 bits per second (approximately the basic message information of the text) to data rates upward of 200,000 bits per second for simple waveform representations. This represents about a 3000 to 1 variation in data rates depending on the signal representation. Of course the data rate is not the only consideration in choosing a speech representation. Other considerations are cost, flexibility of the representation, quality of the speech, etc. We defer a discussion of such issues to the remaining chapters of this book.

The ultimate application is perhaps the most important consideration in the choice of a signal representation and the methods of digital signal processing subsequently applied. Figure 1.4 shows just a few of the many applications areas in speech communications. Although we have already referred to several of these areas, it is worthwhile giving a brief discussion of each of these areas as a means for motivating the techniques to be discussed in subsequent chapters.

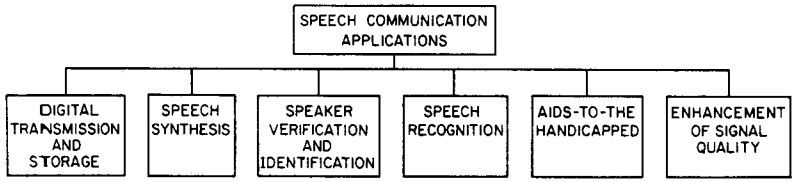


Fig. 1.4 Some typical speech communications applications.

[1]Chapter 3 provides a detailed discussion of parametric speech models.

1.4.1 Digital transmission and storage of speech [3]

One of the earliest and most important applications of speech processing was the vocoder, or *voice coder,* invented by Homer Dudley in the 1930's [5]. The purpose of the vocoder was to reduce the bandwidth required to transmit the speech signal. The need to conserve bandwidth remains, in many situations, in spite of the increased bandwidth provided by sattelite, microwave, and optical communications systems. Furthermore, a need has arisen for systems which digitize speech at as low a bit rate as possible, consistent with low terminal cost for future applications in the all-digital telephone plant. Also, the possibility of extremely sophisticated encryption of the speech signal is sufficient motivation for the use of digital transmission in many applications.

1.4.2 Speech synthesis systems

Much of the interest in speech synthesis systems is stimulated by the need for economical digital storage of speech for computer voice response systems [6]. A computer voice response system is basically an all-digital, automatic information service which can be queried by a person from a keyboard or terminal, and which responds with the desired information by voice. Since an ordinary Touch-Tone® telephone can be the keyboard for such a system, the capabilities of such automatic information services can be made universally available over the switched telephone facilities without the need for any additional specialized equipment [3]. Speech synthesis systems also play a fundamental role in learning about the process of human speech production [7].

1.4.3 Speaker verification and identification systems [8]

The techniques of speaker verification and identification involve the authentication or identification of a speaker from a large ensemble of possible speakers. A speaker verification system must decide if a speaker is the person he claims to be. Such a system is potentially applicable to situations requiring control of access to information or restricted areas and to various kinds of automated credit transactions. A speaker identification system must decide which speaker among an ensemble of speakers produced a given speech utterance. Such systems have potential forensic applications.

1.4.4 Speech recognition systems [9]

Speech recognition is, in its most general form, a conversion from an acoustic waveform to a written equivalent of the message information. The

nature of the speech recognition problem is heavily dependent upon the constraints placed on speaker, speaking situation and message context. The potential applications of speech recognition systems are many and varied; e.g. a voice operated typewriter and voice communication with computers. Also, a speech recognizing system combined with a speech synthesizing system comprises the ultimate low bit rate communication system.

1.4.5 Aids-to-the-handicapped

This application concerns processing of a speech signal to make the information available in a form which is better matched to a handicapped person than is normally available. For example variable rate playback of prerecorded tapes provides an opportunity for a blind "reader" to proceed at any desired pace through given speech material. Also a variety of signal processing techniques have been applied to design sensory aids and visual displays of speech information as aids in teaching deaf persons to speak [10].

1.4.6 Enhancement of signal quality

In many situations, speech signals are degraded in ways that limit their effectiveness for communication. In such cases digital signal processing techniques can be applied to improve the speech quality. Examples include such applications as the removal of reverberation (or echos) from speech, or the removal of noise from speech, or the restoration of speech recorded in a helium-oxygen mixture as used by divers.

1.5 Summary

In this chapter we have introduced the ways in which digital signal processing techniques are applied in speech communication. It is clear that we have selected a very wide range of topics, and to cover them in complete depth would be extremely difficult. There are a number of ways in which a book of this type could be organized. For example, it could be organized with respect to the signal representations of Figure 1.2. Alternatively, a book could be written that would emphasize applications areas. Indeed, a book could be written about each area shown in Figure 1.4. A third possibility, which we have chosen, is to organize the book with respect to signal processing methods. We feel that this approach offers the greatest opportunity to focus on topics that will be of continued importance. As such, the remaining chapters of this book provide a review of digital signal processing methods (Chapter 2), an introduction to the digital speech model (Chapter 3), discussions of time domain representations of speech (Chapter 4), waveform representations (Chapter 5), short-time spectral representations (Chapter 6), homomorphic representations (Chapter 7), and linear predictive representations (Chapter 8). These chapters detail the basic theory of digital speech processing. This theory is widely applicable in many applications areas. To illustrate such applications, the final chapter (Chapter 9) discusses several examples of man-machine communications systems which involve extensive use of the digital signal processing methods discussed in this book.

REFERENCES

1. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, Vol. 27, pp. 623-656, October 1968.
2. J. L. Flanagan, *Speech Analysis, Synthesis, and Perception,* 2nd Edition, Springer Verlag, New York, 1972.
3. J. L. Flanagan, "Computers That Talk and Listen: Man-Machine Communication by Voice," *Proc. IEEE,* Vol. 64, No. 4, pp. 416-432, April 1976.
4. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Trans. AIEE,* Vol. 47, pp. 617-644, February 1928.
5. H. Dudley, "Remaking Speech," *J. Acoust. Soc. Am.*, Vol. 11, pp. 169-177, 1939.
6. L. R. Rabiner and R. W. Schafer, "Digital Techniques for Computer Voice Response: Implementations and Applications," *Proc. IEEE,* Vol. 64, pp. 416-433, April 1976.
7. C. H. Coker, "A Model of Articulatory Dynamics and Control," *Proc. IEEE,* Vol. 64, No. 4, pp. 452-460, April 1976.
8. B. S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proc. IEEE,* Vol. 64, No. 4, pp. 460-475, April 1976.
9. D. R. Reddy, "Speech Recognition by Machine: A Review," *Proc. IEEE,* Vol. 64, No. 4, pp. 501-531, April 1976.
10. H. Levitt, "Speech Processing Aids for the Deaf: An Overview," *IEEE Trans. on Audio and Electroacoustics,* Vol. AU-21, pp. 269-273, June 1973.

2

Fundamentals of Digital Signal Processing

2.0 Introduction

Since the speech processing schemes and techniques that we shall discuss in this book are intrinsically discrete-time signal processing systems, it is essential that a reader have a good understanding of the basic techniques of digital signal processing. In this chapter we present a brief review of the important concepts. This review is intended to serve as a convenient reference for later chapters and to establish the notation that will be used throughout the book. Those readers who are completely unfamiliar with techniques for representation and analysis of discrete-time signals and systems may find it worthwhile to consult a textbook on digital signal processing [1-3] when this chapter does not provide sufficient detail.

2.1 Discrete-Time Signals and Systems

In almost every situation involving information processing or communication, it is natural to begin with a representation of the signal as a continuously varying pattern. The acoustic wave produced in human speech is most certainly of this nature. It is mathematically convenient to represent such continuously varying patterns as functions of a continuous variable t, which represents time. In this book we shall use notation of the form $x_a(t)$ to denote continuously varying (or analog) time waveforms. As we shall see, it is also possible to represent

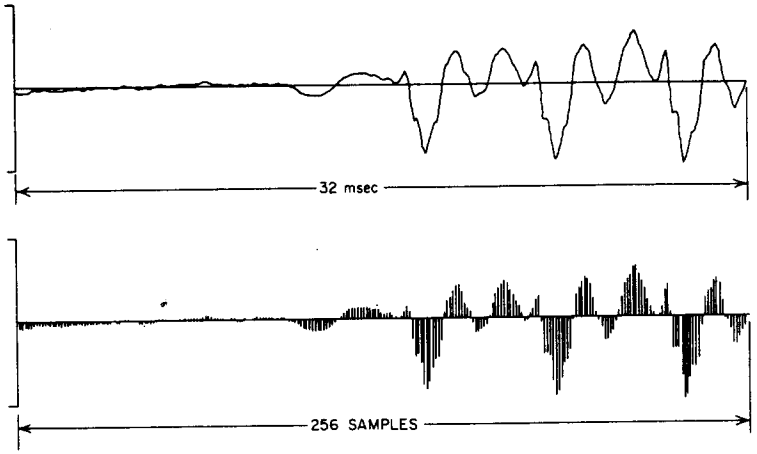


Fig. 2.1 Representations of a speech signal.

the speech signal as a sequence of numbers; indeed, that is what this book is all about. In general we shall use notation of the form, $x(n)$, to denote sequences. If, as is the case for sampled speech signals, a sequence can be thought of as a sequence of samples of an analog signal taken periodically with sampling period, T, then we may find it useful to explicitly indicate this by using the notation, $x_a(nT)$. Figure 2.1 shows an example of a speech signal represented both as an analog signal and as a sequence of samples at a sampling rate of 8 kHz. In subsequent figures, convenience in plotting may dictate the use of the analog representation (i.e., continuous functions) even when the discrete representation is being considered. In such cases, the continuous curve can simply be viewed as the envelope of the sequence of samples.

In our study of digital speech processing systems we will find a number of special sequences repeatedly arising. Several of these sequences are depicted in Fig. 2.2. The unit sample or unit impulse sequence is defined as

$$\begin{aligned} \delta(n) &= 1 \quad n = 0 \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{2.1}$$

The unit step sequence is

$$\begin{aligned} u(n) &= 1 \quad n \geqslant 0 \\ &= 0 \quad n < 0 \end{aligned} \tag{2.2}$$

An exponential sequence is of the form

$$x(n) = a^n \tag{2.3}$$

If a is complex, i.e., $a = re^{j\omega_0}$, then

$$x(n) = r^n e^{j\omega_0 n} = r^n(\cos \omega_0 n + j \sin \omega_0 n) \tag{2.4}$$

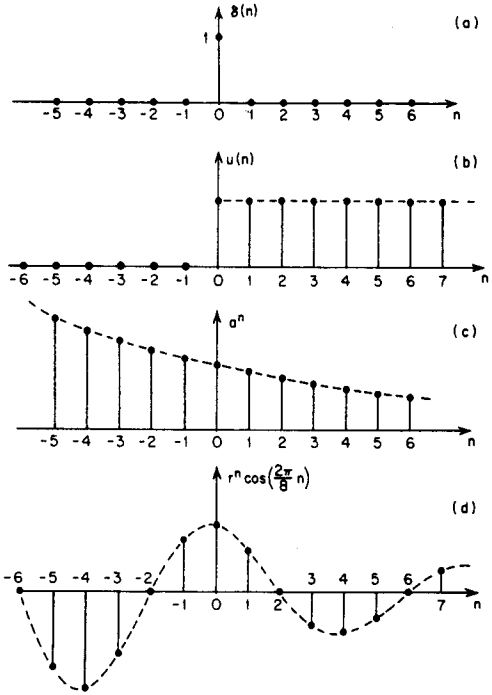


Fig. 2.2 (a) Unit sample; (b) unit step; (c) real exponential; and (d) damped cosine.

If $r = 1$ and $\omega_0 \neq 0$, $x(n)$ is a complex sinusoid; if $\omega_0 = 0$, $x(n)$ is real; and if $r < 1$ and $\omega_0 \neq 0$, then $x(n)$ is an exponentially decaying oscillatory sequence. Sequences of this type arise especially in the representation of linear systems and in modelling the speech waveform.

Signal processing involves the transformation of a signal into a form which is in some sense more desirable. Thus we are concerned with discrete systems, or equivalently, transformations of an input sequence into an output sequence. We shall depict such transformations by block diagrams such as Fig. 2.3a. Many speech analysis systems are designed to estimate several time-varying parameters from samples of the speech wave. Such systems therefore

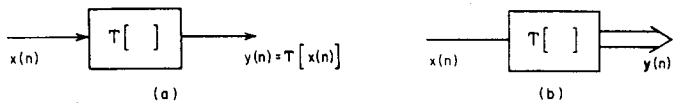


Fig. 2.3 Block diagram representations of: (a) single input/single output system; (b) single input/multiple output system.

have a multiplicity of outputs; i.e., a single input sequence representing the speech signal is transformed into a vector of output sequences as depicted in Fig. 2.3b. In this book, we shall discuss both single output and multiple output speech processing systems.

The special class of linear shift-invariant systems is especially useful in speech processing. Such systems are completely characterized by their response to a unit sample input. For such systems, the output can be computed from the input, $x(n)$, and the unit sample response, $h(n)$, using the convolution sum expression

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) = x(n) * h(n) \tag{2.5a}$$

where the symbol * stands for discrete convolution. An equivalent expression is

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = h(n) * x(n) \tag{2.5b}$$

Linear shift invariant systems are useful for performing filtering operations on speech signals and, perhaps more importantly, they are useful as models for speech production.

2.2 Transform Representation of Signals and Systems

The analysis and design of linear systems are greatly facilitated by frequency-domain representations of both signals and systems. Thus, it is useful to review Fourier and z-transform representations of discrete-time signals and systems.

2.2.1 The z-transform

The *z-transform* representation of a sequence is defined by the pair of equations

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \tag{2.6a}$$

$$x(n) = \frac{1}{2\pi j} \oint_C X(z)\, z^{n-1} dz \tag{2.6b}$$

The "z-transform" or "direct transform" of $x(n)$ is defined by Eq. (2.6a). It can be seen that in general $X(z)$ is an infinite power series in the variable z^{-1}, where the sequence values, $x(n)$, play the role of coefficients in the power series. In general such a power series will converge (add up) to a finite value only for certain values of z. A sufficient condition for convergence is

$$\sum_{n=-\infty}^{\infty} |x(n)||z^{-n}| < \infty \tag{2.7}$$

The set of values for which the series converges defines a region in the complex z-plane known as the *region of convergence*. In general this region is of the form

$$R_1 < |z| < R_2 \tag{2.8}$$

To see the relationship of the region of convergence to the nature of the sequence, let us consider some examples.

2.2.1a Example 1

Let $x(n) = \delta(n-n_0)$. Then by substitution into Eq. (2.6a)

$$X(z) = z^{-n_0}$$

2.2.1b Example 2

Let $x(n) = u(n) - u(n-N)$. Then

$$X(z) = \sum_{n=0}^{N-1} (1)z^{-n} = \frac{1 - z^{-N}}{1 - z^{-1}}$$

In both of these cases, $x(n)$ is of finite duration. Therefore $X(z)$ is simply a polynomial in the variable z^{-1}, and the region of convergence is everywhere but $z = 0$. All finite length sequences have a region of convergence that is at least the region $0 < |z| < \infty$.

2.2.1c Example 3

Let $x(n) = a^n u(n)$. Then

$$X(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \frac{1}{1 - az^{-1}}, \quad |a| < |z|$$

In this case the power series is recognized as a geometric series for which a convenient closed form expression exists for the sum. This result is typical of infinite duration sequences which are nonzero for $n > 0$. In this general case, the region of convergence is of the form $|z| > R_1$.

2.2.1d Example 4

Let $x(n) = -b^n u(-n-1)$. Then

$$X(z) = \sum_{n=-\infty}^{-1} b^n z^{-n} = \frac{1}{1 - bz^{-1}}, \quad |z| < |b|$$

This is typical of infinite duration sequences that are nonzero for $n < 0$, where the region of convergence is, in general, $|z| < R_2$. The most general case in which $x(n)$ is nonzero for $-\infty < n < \infty$ can be viewed as a combination of the cases illustrated by Examples 3 and 4. Thus for this case, the region of convergence is of the form $R_1 < |z| < R_2$.

The "inverse transform" is given by the contour integral in Eq. (2.6b), where C is a closed contour that encircles the origin of the z-plane and lies inside the region of convergence of $X(z)$. For the special case of rational transforms, a partial fraction expansion provides a convenient means for finding inverse transforms [1].

There are many theorems and properties of the z-transform representation that are useful in the study of discrete-time systems. A working familiarity with these theorems and properties is essential for complete understanding of the material in subsequent chapters. A list of important theorems is given in Table 2.1. These theorems can be seen to be similar in form to corresponding theorems for Laplace transforms of continuous time functions. However, this similarity should not be construed to mean that the z-transform is in any sense an approximation to the Laplace transform. The Laplace transform is an *exact* representation of a continuous-time function, and the z-transform is an *exact* representation of a sequence of numbers. The appropriate way to relate the continuous and discrete representations of a signal is through the sampling theorem as discussed in Section 2.4.

Table 2.1 Sequences and Their Corresponding z-Transforms

	Sequence	*z-Transform*
1. Linearity	$ax_1(n) + bx_2(n)$	$aX_1(z) + bX_2(z)$
2. Shift	$x(n+n_0)$	$z^{n_0}X(z)$
3. Exponential Weighting	$a^n x(n)$	$X(a^{-1}z)$
4. Linear Weighting	$nx(n)$	$-z\frac{dX(z)}{dz}$
5. Time Reversal	$x(-n)$	$X(z^{-1})$
6. Convolution	$x(n)*h(n)$	$X(z)H(z)$
7. Multiplication of Sequences	$x(n)w(n)$	$\frac{1}{2\pi j}\oint_C X(v)W(z/v)v^{-1}dv$

2.2.2 The Fourier transform

The Fourier transform representation of a discrete-time signal is given by the equations

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \tag{2.9a}$$

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega \tag{2.9b}$$

These equations can easily be seen to be a special case of Eqs. (2.6). Specifically the Fourier representation is obtained by restricting the z-transform to the unit circle of the z-plane; i.e., by setting $z = e^{j\omega}$. As depicted in Fig. 2.4, the digital frequency variable, ω, also has the interpretation as angle in the

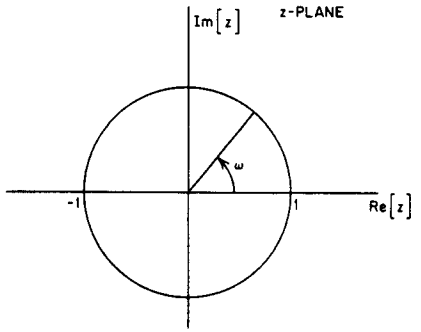


Fig. 2.4 The unit circle of the z-plane.

z-plane. A sufficient condition for the existence of a Fourier transform representation can be obtained by setting $|z| = 1$ in Eq. (2.7), thus obtaining

$$\sum_{n=-\infty}^{\infty} |x(n)| < \infty \tag{2.10}$$

As examples of typical Fourier transforms, we can return to the examples of Section 2.2.1. The Fourier transform is obtained simply by setting $z = e^{j\omega}$ in the given expression. In the first two examples, the result is clearly the Fourier transform since the region of convergence of $X(z)$ includes the unit circle. However, in Examples 3 and 4, the Fourier transform will exist only if $|a| < 1$ and $|b| > 1$ respectively. These conditions, of course, correspond to decaying sequences for which Eq. (2.10) holds.

An important feature of the Fourier transform of a sequence is that $X(e^{j\omega})$ is a periodic function of ω, with period 2π. This follows easily by substituting $\omega + 2\pi$ into Eq. (2.9a). Alternatively, since $X(e^{j\omega})$ is the evaluation of $X(z)$ on the unit circle, we can see that $X(e^{j\omega})$ must repeat each time we go completely around the unit circle; i.e., ω has gone through 2π radians.

By setting $z = e^{j\omega}$ in each of the theorems in Table 2.1, we obtain a corresponding set of theorems for the Fourier transform. Of course, these results are valid only if the Fourier transforms that are involved do indeed exist.

2.2.3 The discrete Fourier transform

As in the case of analog signals, if a sequence is periodic with period N; i.e.,

$$\tilde{x}(n) = \tilde{x}(n+N) \quad -\infty < n < \infty \tag{2.11}$$

then $\tilde{x}(n)$ can be represented by a discrete sum of sinusoids rather than an integral as in Eq. (2.9b). The Fourier series representation for a periodic sequence is

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n) e^{-j\frac{2\pi}{N}kn} \tag{2.12a}$$

$$\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) e^{j\frac{2\pi}{N}kn} \tag{2.12b}$$

This is an *exact* representation of a periodic sequence. However, the great utility of this representation lies in imposing a different interpretation upon Eqs. (2.12). Let us consider a finite length sequence, $x(n)$, that is zero outside the interval $0 \leqslant n \leqslant N-1$. Then the z-transform is

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} \tag{2.13}$$

If we evaluate $X(z)$ at N equally spaced points on the unit circle, i.e., $z_k = e^{j2\pi k/N}$, $k = 0,1,\ldots,N-1$, then we obtain

$$X(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \quad k = 0,1,\ldots,N-1 \tag{2.14}$$

If we construct a periodic sequence as an infinite sequence of replicas of $x(n)$,

$$\tilde{x}(n) = \sum_{r=-\infty}^{\infty} x(n+rN) \tag{2.15}$$

then, the samples $X(e^{j2\pi k/N})$ are easily seen from Eqs. (2.12a) and (2.14) to be the Fourier coefficients of the periodic sequence $\tilde{x}(n)$ in Eq. (2.15). Thus a sequence of length N can be exactly represented by a discrete Fourier transform (DFT) representation of the form

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \qquad k = 0,1,\ldots,N-1 \tag{2.16a}$$

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j\frac{2\pi}{N}kn} \qquad n = 0,1,\ldots,N-1 \tag{2.16b}$$

Clearly the only difference between Eqs. (2.16) and (2.12) is a slight modification of notation (removing the $\sim$ symbols which indicate periodicity) and the explicit restriction to the finite intervals $0 \leqslant k \leqslant N-1$ and $0 \leqslant n \leqslant N-1$. It is extremely important, however, to bear in mind when using the DFT representation that all sequences behave as if they were periodic when represented by a DFT representation. That is, the DFT is really a representation of the periodic sequence given in Eq. (2.15). An alternative point of view is that when DFT representations are used, sequence indices must be interpreted modulo N. This follows from the fact that if $x(n)$ is of length N

$$\tilde{x}(n) = \sum_{k=-\infty}^{\infty} x(n+rN) = x(n \text{ modulo } N)$$

$$= x((n))_N. \tag{2.17}$$

The double parenthesis notation provides a convenient expression of the inherent periodicity of the DFT representation. This built-in periodicity has a significant effect on the properties of the DFT representation. Some of the more important theorems are listed in Table 2.2. The most obvious feature is that shifted sequences are shifted modulo N. This leads, for example, to significant differences in the discrete convolution.

The DFT representation, with all its peculiarities, is important for a number of reasons:

1. The DFT, $X(k)$, can be viewed as a sampled version of the z-transform (or Fourier transform) of a finite length sequence
2. The DFT has properties very similar (with modifications due to the inherent periodicity) to many of the useful properties of z-transforms and Fourier transforms.
3. The N values of $X(k)$ can be computed very efficiently (with time proportional to $N \log N$) by a set of computational algorithms known collectively as the fast Fourier transform (FFT) [1-4].

The DFT is widely used for computing spectrum estimates, correlation functions and for implementing digital filters [5-6]. We shall have frequency occasion to apply DFT representations in speech processing.

Table. 2.2 Sequences and Their Corresponding Discrete Fourier Transforms

	Sequence	*N-point DFT*
1. Linearity	$ax_1(n) + bx_2(n)$	$aX_1(k) + bX_2(k)$
2. Shift	$x((n+n_0))_N$	$e^{j\frac{2\pi}{N}kn_0}X(k)$
3. Time Reversal	$x((-n))_N$	$X^*(k)$
4. Convolution	$\sum_{m=0}^{N-1} x(m)h((n-m))_N$	$X(k)H(k)$
5. Multiplication of Sequences	$x(n)w(n)$	$\frac{1}{N}\sum_{r=0}^{N-1} X(r)W((k-r))_N$

2.3 Fundamentals of Digital Filters

A digital filter is a discrete-time linear shift-invariant system. Recall that for such a system the input and output are related by the convolution sum expression of Eqs. (2.5). The corresponding relation between the z-transform of the sequences involved is as given in Table 2.1,

$$Y(z) = H(z)X(z) \tag{2.18}$$

The z-transform of the unit sample response, $H(z)$, is called the *system function* of the system. The Fourier transform of the unit impulse response, $H(e^{j\omega})$, is called the *frequency response*. $H(e^{j\omega})$ is in general a complex function of ω,

which can be expressed in terms of real and imaginary parts as

$$H(e^{j\omega}) = H_r(e^{j\omega}) + jH_i(e^{j\omega}) \tag{2.19}$$

or in terms of magnitude and phase angle as

$$H(e^{j\omega}) = |H(e^{j\omega})|e^{j\arg[H(e^{j\omega})]} \tag{2.20}$$

A *causal* linear shift invariant-system is one for which $h(n) = 0$ for $n < 0$. A *stable* system is one for which every bounded input produces a bounded output. A necessary and sufficient condition for a linear shift-invariant system to be stable is

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty \tag{2.21}$$

This condition is identical to Eq. (2.10) and thus is sufficient for the existence of $H(e^{j\omega})$.

In addition to the convolution sum expression of Eq. (2.5), all linear shift invariant systems of interest for implementation as filters have the property that the input and output satisfy a linear difference equation of the form

$$y(n) - \sum_{k=1}^{N} a_k y(n-k) = \sum_{r=0}^{M} b_r x(n-r) \tag{2.22}$$

By evaluating the z-transform of both sides of this equation we can show that

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{r=0}^{M} b_r z^{-r}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2.23}$$

A useful observation results from comparing Eq. (2.22) to Eq. (2.23). That is, given a difference equation in the form of Eq. (2.22) we can obtain $H(z)$ directly by simply identifying the coefficients of the delayed input in Eq. (2.22) with corresponding powers of z^{-1} in the numerator and coefficients of the delayed output with corresponding powers of z^{-1} in the denominator.

The system function, $H(z)$, is in general a rational function of z^{-1}. As such it is characterized by the locations of its poles and zeros in the z-plane. Specifically $H(z)$ can be expressed as

$$H(z) = \frac{A\prod_{r=1}^{M}(1-c_r z^{-1})}{\prod_{k=1}^{N}(1-d_k z^{-1})} \tag{2.24}$$

From our discussion of z-transforms, we recall that a causal system will have a region of convergence of the form $|z| > R_1$. If the system is also stable, then R_1 must be less than unity so that the region of convergence contains the unit

circle. Therefore the poles of $H(z)$ must all be inside the unit circle for a stable and causal system.

It is convenient to define two classes of linear shift invariant systems. These are the class of finite duration impulse response (FIR) systems and the class of infinite duration impulse response (IIR) systems. These classes have distinct properties which we shall summarize below.

2.3.1 FIR systems

If all the coefficients, a_k, in Eq. (2.22) are zero, the difference equation becomes

$$y(n) = \sum_{r=0}^{M} b_r x(n-r) \tag{2.25}$$

Comparing Eq. (2.25) to Eq. (2.5b) we observe that

$$\begin{aligned} h(n) &= b_n \quad 0 \leqslant n \leqslant M \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{2.26}$$

FIR systems have a number of important properties. First, we note that $H(z)$ is a polynomial in z^{-1}, and thus $H(z)$ has no nonzero poles, only zeros. Also, FIR systems can have exactly linear phase. If $h(n)$ satisfies the relation

$$h(n) = \pm h(M-n) \tag{2.27}$$

then $H(e^{j\omega})$ has the form

$$H(e^{j\omega}) = A(e^{j\omega})e^{-j\omega(M/2)} \tag{2.28}$$

where $A(e^{j\omega})$ is either purely real or imaginary depending upon whether Eq. (2.27) is satisfied with + or − respectively.

The possibility of *exactly* linear phase is often very useful in speech processing applications where precise time alignment is essential. This property of FIR filters also can greatly simplify the approximation problem since it is only necessary to be concerned with approximating a desired magnitude response. The penalty that is paid for being able to design filters with an exact linear phase response is that a large impulse response duration is required to adequately approximate sharp cutoff filters.

Based on the properties associated with linear phase FIR filters there have developed three well known design methods for approximating an arbitrary set of specifications with an FIR filter. These three methods are:

1. Window design [1,2,5,7]
2. Frequency sampling design [1,2,8]
3. Optimal (minimax error) design [1,2,9-11]

Only the first of these techniques is an analytical design technique, i.e., a closed

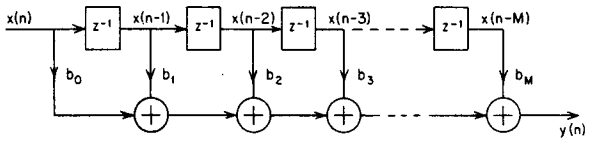


Fig. 2.5 Digital network for FIR system.

form set of equations can be solved to obtain the filter coefficients. The second and third design methods are optimization methods which use iterative (rather than closed form) approaches to obtain the desired filter. Although the window method is simple to apply, the third method is also widely used. This is in part due to a series of intensive investigations into the properties of the optimal FIR filters, and in part due to the general availability of a well-documented design program which enables the user to approximate any desired set of specifications [2,10].

In considering the implementation of digital filters, it is often useful to represent the filter in block diagram form. The difference equation of Eq. (2.25) is depicted in Fig. 2.5. Such a diagram, often called a digital filter structure, graphically depicts the operations required to compute each value of the output sequence from values of the input sequence. The basic elements of the diagram depict means for addition, multiplication of sequence values by constants (constants indicated on branches imply multiplication), and storage of past values of the input sequence. Thus the block diagram gives a clear indication of the complexity of the system. When the system has linear phase, further significant simplifications can be incorporated into the implementation. (See Problem 2.7).

2.3.2 IIR systems

If the system function of Eq. (2.24) has poles as well as zeros, then the difference equation of Eq. (2.22) can be written as

$$y(n) = \sum_{k=1}^{N} a_k y(n-k) + \sum_{r=0}^{M} b_r x(n-r) \tag{2.29}$$

This equation is a recurrence formula that can be used sequentially to compute the values of the output sequence from past values of the output and present and past values of the input sequence. If $M < N$ in Eq. (2.24), $H(z)$ can be expanded in a partial fraction expansion as in

$$H(z) = \sum_{k=1}^{N} \frac{A_k}{1 - d_k z^{-1}} \tag{2.30}$$

For a causal system, it is easily shown (See Problem 2.9) that

$$h(n) = \sum_{k=1}^{N} A_k (d_k)^n u(n) \tag{2.31}$$

Thus, we see that $h(n)$ has infinite duration. However, because of the recurrence formula of Eq. (2.29), it is often possible to implement an IIR filter that approximates a given set of specifications more efficiently (i.e., using fewer computations) than is possible with an FIR system. This is particularly true for sharp cutoff frequency selective filters.

A wide variety of design methods are available for IIR filters. Design methods for frequency selective filters (lowpass, bandpass, etc.) are generally based on transformations of classical analog design procedures that are straightforward to implement. Included in this class are

1. Butterworth designs - (maximally flat amplitude)
2. Bessel designs - (maximally flat group delay)
3. Chebyshev designs (equiripple in either passband or stopband)
4. Elliptic designs - (equiripple in both passband and stopband)

All the above methods are analytical in nature and have been widely applied to the design of IIR digital filters [1,2]. In addition a variety of IIR optimization

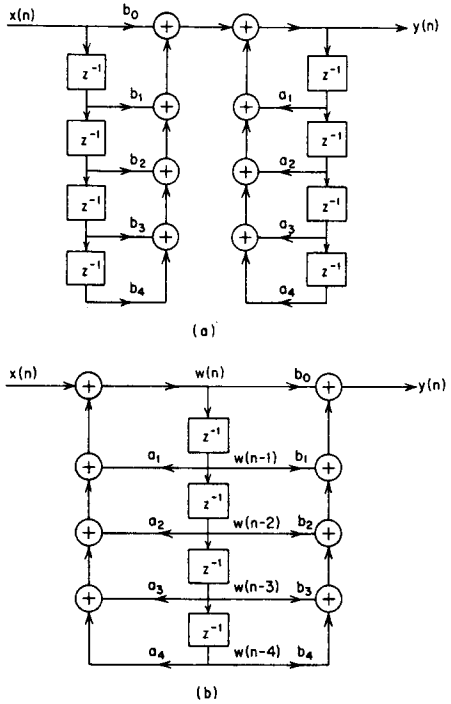


Fig. 2.6 (a) Direct form IIR structure; (b) direct form structure with minimum storage.

methods have been developed for approximating design specifications which ar not easily adapted to one of the above approximation methods [12].

The major difference between FIR and IIR filters is that IIR filters cann be designed to have exact linear phase, whereas FIR filters can have this pr perty. In exchange, the IIR filter is often orders of magnitude more efficient realizing sharp cutoff filters than FIR filters [13].

There is considerable flexibility in the implementation of IIR system The network implied by Eq. (2.29) is depicted in Fig. 2.6a, for the ca $M = N = 4$. This is often called the direct form implementation. The gener ization to arbitrary M and N is obvious. The difference equation Eq. (2.29) c be transformed into many equivalent forms. Particularly useful among these the set of equations

$$w(n) = \sum_{k=1}^{N} a_k w(n-k) + x(n)$$

$$y(n) = \sum_{r=0}^{M} b_r w(n-r) \qquad (2.3$$

(See Problem 2.10). This set of equations can be implemented as shown Fig. 2.6b, with a significant saving of memory required to store the delay sequence values.

Equation (2.24) shows that $H(z)$ can be expressed as a product of pol and zeros. These poles and zeros occur in complex conjugate pairs since t coefficients a_k and b_r are real. By grouping the complex conjugate poles a zeros into complex conjugate pairs it is possible to express $H(z)$ as a product elementary second-order system functions, of the form

$$H(z) = A\prod_{k=1}^{K}\left(\frac{1 + b_{1k}z^{-1} + b_{2k}z^{-2}}{1 - a_{1k}z^{-1} - a_{2k}z^{-2}}\right) \qquad (2.3$$

where K is the integer part of $(N+1)/2$. Each second order system can implemented as in Fig. 2.6 and the systems cascaded to implement $H(z)$. T is depicted in Fig. 2.7a for $N = M = 4$. Again the generalization to higher o ers is obvious. The partial fraction expansion of Eq. (2.30) suggests s another approach to implementation. By combining terms involving comp conjugate poles, $H(z)$ can be expressed as

$$H(z) = \sum_{k=1}^{K} \frac{c_{0k} + c_{1k}z^{-1}}{1 - a_{1k}z^{-1} - a_{2k}z^{-2}} \qquad (2.$$

This suggests a parallel form implementation as depicted in Fig. 2.7b for $N =$

All of the implementations discussed are used in speech processing. linear filtering applications, the cascade form generally exhibits superior perf mance with respect to roundoff noise, coefficient inaccuracies, and stabi [1,2]. All of the above forms have been used in speech synthesis applicatic with the direct form being particularly important in synthesis from linear pr iction parameters (See Chapter 8).

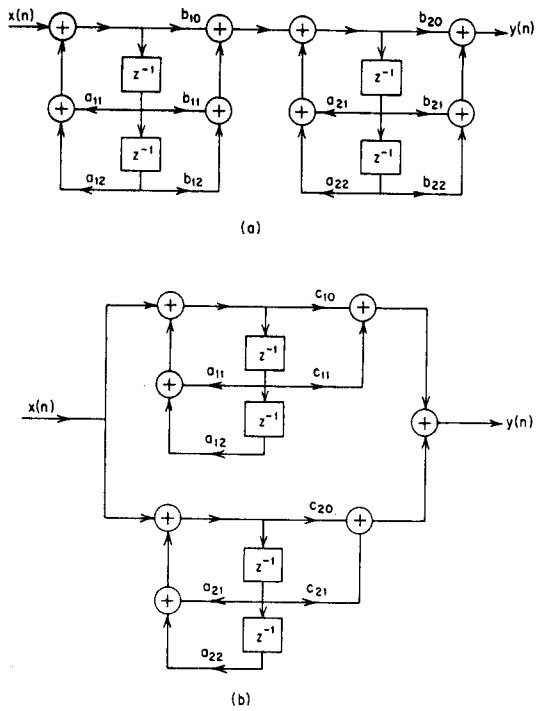


Fig. 2.7 (a) Cascade form; (b) parallel form.

2.4 Sampling

To use digital signal processing methods on an analog signal such as speech, it is necessary to represent the signal as a sequence of numbers. This is commonly done by sampling the analog signal, denoted $x_a(t)$, periodically to produce the sequence

$$x(n) = x_a(nT) \quad -\infty < n < \infty \tag{2.35}$$

where n, of course, takes on only integer values. Figure 2.1 shows a speech waveform and the corresponding set of samples with period $T = 1/8000$ sec.

2.4.1 The sampling theorem

The conditions under which the sequence of samples in Eq. (2.35) is a unique representation of the original analog signal are well known and are often summarized as follows:

> *The Sampling Theorem:* If a signal $x_a(t)$ has a bandlimited Fourier transform $X_a(j\Omega)$, such that $X_a(j\Omega) = 0$ for $\Omega \geqslant 2\pi F_N$, then

> $x_a(t)$ can be uniquely reconstructed from equally spaced samp
> $x_a(nT)$, $-\infty < n < \infty$, if $1/T > 2F_N$.

The above theorem follows from the fact that if the Fourier transform $x_a(t)$ is defined as

$$X_a(j\Omega) = \int_{-\infty}^{\infty} x_a(t)e^{-j\Omega t}dt \tag{2.}$$

and the Fourier transform of the sequence $x(n)$ is defined as in Eq. (2.) then if $X(e^{j\omega})$ is evaluated for frequencies $\omega = \Omega T$, then $X(e^{j\Omega T})$ is relate $X_a(j\Omega)$ by [1,2]

$$X(e^{j\Omega T}) = \frac{1}{T}\sum_{k=-\infty}^{\infty} X_a\left(j\Omega + j\frac{2\pi}{T}k\right) \tag{2.}$$

To see the implications of Eq. (2.37), let us assume that $X_a(j\Omega)$ is as show Fig. 2.8a; i.e., assume that $X_a(j\Omega) = 0$ for $|\Omega| > \Omega_N = 2\pi F_N$. The quency F_N is called the *Nyquist frequency*. Now according to Eq. (2. $X(e^{j\Omega T})$ is the sum of an infinite number of replicas of $X_a(j\Omega)$, each cente at integer multiples of $2\pi/T$. Fig. 2.8b depicts the case when $1/T > 2F_N$ that the images of the Fourier transform do not overlap into the base b $|\Omega| < 2\pi F_N$. Figure 2.8c, on the other hand, shows the case $1/T < 2F_N$.

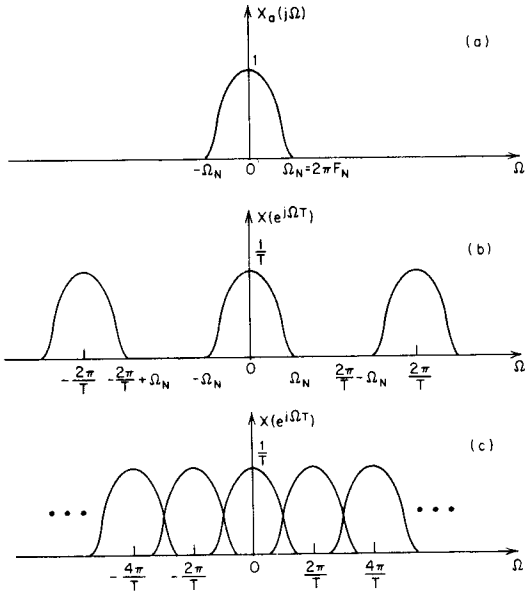


Fig. 2.8 Illustration of sampling.

this case, the image centered at $2\pi/T$ overlaps into the base band. This condition, where a high frequency seemingly takes on the identity of a lower frequency, is called *aliasing*. Clearly, aliasing can be avoided only if the Fourier transform is bandlimited and if the sampling frequency $(1/T)$ is equal to at least twice the Nyquist frequency $(1/T > 2F_N)$.

Under the condition $1/T > 2F_N$, it is clear that the Fourier transform of the sequence of samples is proportional to the Fourier transform of the analog signal in the base band; i.e.,

$$X(e^{j\Omega T}) = \frac{1}{T} X_a(j\Omega) \quad |\Omega| < \frac{\pi}{T} \tag{2.38}$$

Using this result, it can be shown that [1,2] the original signal can be related to the sequence of samples by the interpolation formula

$$x_a(t) = \sum_{n=-\infty}^{\infty} x_a(nT) \left[\frac{\sin[\pi(t-nT)/T]}{\pi(t-nT)/T} \right] \tag{2.39}$$

Thus, given samples of a bandlimited analog signal taken at a rate at least twice the Nyquist frequency, it is possible to reconstruct the original analog signal using Eq.(2.39). Practical digital-to-analog converters seek to approximate Eq. (2.39).

Sampling is implicit in many speech processing algorithms that seek to estimate basic parameters of speech production such as pitch and formant frequencies. In such cases, an analog function is not available to be sampled directly, as in the case of sampling the speech waveform itself. However, such parameters change very slowly with time, and thus it is possible to estimate (sample) them at rates on the order of 100 samples/sec. Given samples of a speech parameter, a bandlimited analog function for that parameter can of course be constructed using Eq. (2.39).

2.4.2 Decimation and interpolation of sampled waveforms

In many examples that we shall discuss in this book, there arises the need to change the sampling rate of a discrete time signal. One example occurs when speech is sampled using one-bit differential quantization at a high sampling rate (delta modulation), and then converted to a multi-bit PCM representation at a lower sampling rate. Another example is when some parameter of the speech signal is sampled at a low rate for efficient coding, and then a higher rate is required for reconstruction of the speech signal. The sampling rate must be reduced in the first case and increased in the second case. The processes of sampling rate reduction and increase will henceforth be called decimation and interpolation.

In discussing both cases, let us assume that we have a sequence of samples $x(n) = x_a(nT)$, where the analog function $x_a(t)$ has a bandlimited Fourier transform such that $X_a(j\Omega) = 0$ for $|\Omega| > 2\pi F_N$. Then we have just seen that if $1/T > 2F_N$, the Fourier transform of $x(n)$ will satisfy

$$X(e^{j\Omega T}) = \frac{1}{T} X_a(j\Omega) \quad |\Omega| < \frac{\pi}{T} \tag{2.40}$$

2.4.2a Decimation

Let us suppose that we wish to reduce the sampling rate by a factor M; i.e., we wish to compute a new sequence corresponding to samples of $x_a(t)$ taken with period $T' = MT$, i.e.,

$$y(n) = x_a(nT') = x_a(nTM) \tag{2.41}$$

It is easily seen that

$$y(n) = x(Mn) \quad -\infty < n < \infty. \tag{2.42}$$

That is, $y(n)$ is obtained simply by periodically retaining only one out of every M samples. From our previous discussion of the sampling theorem we note that if $1/T' > 2F_N$, then the samples $y(n)$ will also be adequate to uniquely represent the original analog signal. The Fourier transforms of $x(n)$ and $y(n)$ are related by the expression [14]

$$Y(e^{j\Omega T}) = \frac{1}{M} \sum_{k=0}^{M-1} X(e^{\frac{j(\Omega T - 2\pi k)}{M}}) \tag{2.43}$$

From Eq. (2.43) it can be seen that in order that there be no overlap between the images of $X(e^{j\Omega T})$, we must have $1/T' > 2F_N$. If this condition holds, then we see that

$$\begin{aligned} Y(e^{j\Omega T}) &= \frac{1}{M} X(e^{\frac{j\Omega T'}{M}}) \\ &= \frac{1}{M} \frac{1}{T} X_a(j\Omega) \\ &= \frac{1}{T'} X_a(j\Omega) \qquad -\frac{\pi}{T'} < \Omega < \frac{\pi}{T'} \end{aligned} \tag{2.44}$$

Figure 2.9 shows an example of sampling rate reduction. Figure 2.9a shows the Fourier transform of the original analog signal. Figure 2.9b shows the Fourier transform of $x(n) = x_a(nT)$ where the sampling rate $(1/T)$ is somewhat greater than the Nyquist rate $(2F_N)$. Figure 2.9c shows the case of sampling rate reduction by a factor of 3; i.e., $T' = 3T$. For this case aliasing occurs because $1/T' < 2F_N$. However, suppose $x(n)$ is filtered with a digital lowpass filter with cutoff frequency $\pi/T' = \pi/(3T)$ producing a sequence $w(n)$. For our example, the Fourier transform of the output of the lowpass filter is shown in Figure 2.9d. Aliasing does not occur when the sampling rate of the filtered signal is reduced by a factor of 3 as depicted in Figure 2.9e; however, the samples $y(n)$ no longer represent $x_a(t)$ but rather a new signal $y_a(t)$ which is a lowpass filtered version of $x_a(t)$. A block diagram of a general decimation system is given in Figure 2.10.

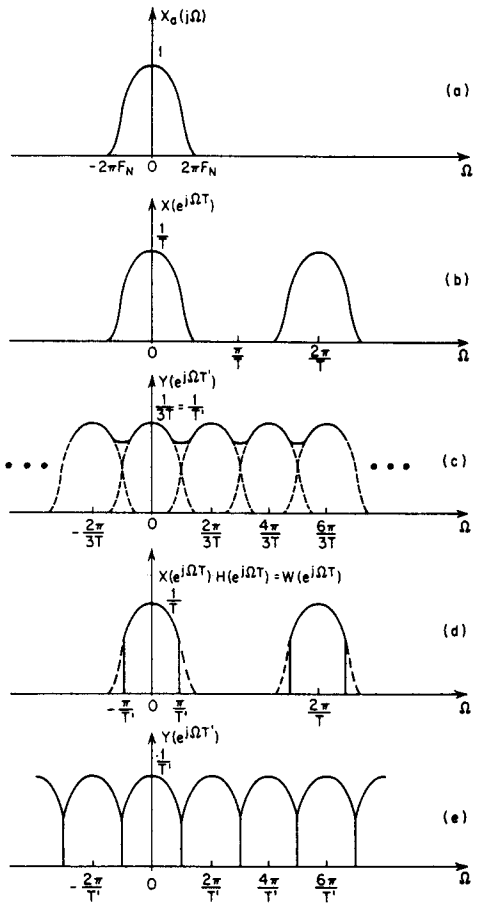


Fig. 2.9 Illustration of decimation.

2.4.2b Interpolation

Now suppose that we have samples of an analog waveform $x(n) = x_a(nT)$. If we wish to increase the sampling rate by an integer factor L, we must compute a new sequence corresponding to samples of $x_a(t)$ taken with period $T' = T/L$; i.e.,

$$y(n) = x_a(nT') = x_a(nT/L) \tag{2.45}$$

Clearly, $y(n) = x(n/L)$ for $n = 0, \pm L, \pm 2L, \ldots$ but we must fill in the unknown samples for all other values of n by an interpolation process [14]. To see how this can be done using a digital filter, consider the sequence

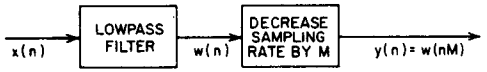


Fig. 2.10 Block diagram representation of decimation.

$$\begin{aligned} v(n) &= x\left(\frac{n}{L}\right) \qquad n = 0, \pm L, \pm 2L, \ldots \\ &= 0 \qquad\qquad \textit{otherwise} \end{aligned} \tag{2.46}$$

The Fourier transform of $v(n)$ is easily shown to be [14]

$$\begin{aligned} V(e^{j\Omega T'}) &= X(e^{j\Omega T'L}) \\ &= X(e^{j\Omega T}). \end{aligned} \tag{2.47}$$

Thus $V(e^{j\Omega T'})$ is periodic with period $2\pi/T = 2\pi/(LT')$, as well as with period $2\pi/T'$ as is the case in general for sequences associated with a sampling period T'. Figure 2.11a shows $V(e^{j\Omega T'})$ [and $X(e^{j\Omega T})$] for the case $T' = T/3$. In order to obtain the sequence

$$y(n) = x_a(nT')$$

from the sequence $v(n)$, we must ensure that

$$Y(e^{j\Omega T'}) = \frac{1}{T'} X_a(j\Omega) \quad -\frac{\pi}{T'} \leqslant \Omega \leqslant \frac{\pi}{T'} \tag{2.48}$$

Assuming that

$$X(e^{j\Omega T}) = \frac{1}{T} X_a(j\Omega) \quad -\frac{\pi}{T} \leqslant \Omega \leqslant \frac{\pi}{T} \tag{2.49}$$

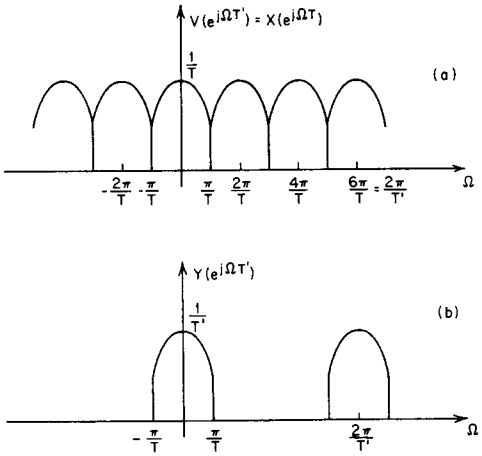


Fig. 2.11 Illustration of interpolation.

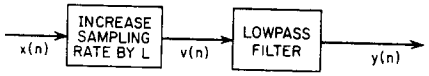


Fig. 2.12 Block diagram representation of interpolation.

then it is clear from Figure 2.11a that what is required is that the images of $X_a(j\Omega)$ in $V(e^{j\Omega T})$, that are centered at $\Omega = 2\pi/T$ and $\Omega = 4\pi/T$, must be removed by a digital lowpass filter that rejects all frequency components in the range $\pi/T \leqslant \Omega \leqslant \pi/T'$. Moreover, to ensure that the amplitude is correct for sampling interval T', the gain of the filter must be $L = T/T'$. That is

$$Y(e^{j\Omega T'}) = H(e^{j\Omega T'})\,V(e^{j\Omega T'}) = H(e^{j\Omega T'})\,X(e^{j\Omega T})$$

$$= H(e^{j\Omega T'})\,\frac{1}{T}\,X_a(j\Omega) \tag{2.50}$$

Thus, in order that $Y(e^{j\Omega T'}) = (1/T')\,X_a(j\Omega)$ for $\Omega \leqslant \pi/T'$ we require that

$$H(e^{j\Omega T'}) = L \quad |\Omega| \leqslant \frac{\pi}{T}$$

$$= 0 \quad \textit{otherwise} \tag{2.51}$$

The general interpolation system is depicted in Fig. 2.12.

2.4.2c Non-Integer Sampling Rate Changes

It is readily seen that samples corresponding to a sampling period $T' = MT/L$ can be obtained by a combination of interpolation by a factor L followed by decimation by a factor M. By suitable choice of the integers M and L, we can approach arbitrarily close to any desired ratio of sampling rates. By combining Figures 2.9 and 2.11, we observe that a single lowpass filter suffices for both the interpolation and decimation filter. This is depicted in Fig. 2.13.

2.4.2d Advantages of FIR Filters

An extremely important consideration in the implementation of decimators and interpolators is the choice of the type of lowpass filter. For these systems, a significant savings in computation over alternative filter types can be obtained by using finite impulse response (FIR) filters in a standard direct form implementation. The savings in computations for FIR filters is due to the observation that for decimators only one of each M output samples needs to be calculated, while for interpolators, $L - 1$ out of every L samples of the input are zero valued, and therefore do not affect the computation. These facts cannot be fully exploited using IIR filters [14].

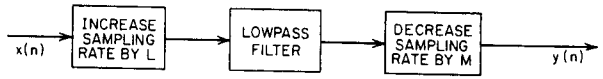


Fig. 2.13 Block diagram representation of sampling rate increase by a factor of L/M.

Assuming that the required filtering is being performed using FIR filters, then for large changes in the sampling rate (i.e, large M for decimators, or large L for interpolators) it has been shown that it is more efficient to reduce (or increase) the sampling rate with a series of decimation stages than to make the entire rate reduction with one stage. In this way the sampling rate is reduced gradually resulting in much less severe filtering requirements on the lowpass filters at each stage. The details of multistage implementation of decimation, interpolation and narrowband filtering are given in Refs. [15-18].

2.5 Summary

In this chapter we have presented a review of the fundamentals of discrete-time signal processing. The notions of discrete convolution, difference equations, and frequency domain representations of signals and systems will be used extensively in this book. Also the concepts of sampling of analog signals and digital alteration of the sampling rate discussed in Section 2.4 are extremely important in all types of digital speech processing systems.

REFERENCES

1. A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.

2. L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.

3. A. Peled and B. Liu, *Digital Signal Processing, Theory, Design and Implementation,* John Wiley and Sons, New York, 1976.

4. J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Computation of Complex Fourier Series," *Math Computation,* Vol. 19, pp. 297-381, April 1965.

5. H. D. Helms, "Fast Fourier Transform Method of Computing Difference Equations and Simulating Filters," *IEEE Trans. Audio and Electroacoustics,* Vol. 15, No. 2, pp. 85-90, 1967.

6. T. G. Stockham, "High-Speed Convolution and Correlation," *1966 Spring Joint Computer Conference,* AFIPS Proc., Vol. 28, pp. 229-233, 1966.

7. J. F. Kaiser, "Nonrecursive Digital Filter Design Using the I_o-Sinh Window Function," *Proc. 1974 IEEE Int. Symp. on Circuits and Systems,* San Francisco, pp. 20-23, April 1974.

8. L. R. Rabiner, B. Gold, and C. A. McGonegal, "An Approach to the Approximation Problem for Nonrecursive Digital Filters," *IEEE Trans. Audio and Electroacoustics,* Vol. 19, No. 3, pp. 200-207, September 1971.

9. T. W. Parks and J. H. McClellan, "Chebyshev Approximation for Nonre-

cursive Digital Filter with Linear Phase," *IEEE Trans. Circuit Theory,* Vol. CT-19, pp. 189-194, March 1972.

10. J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," *IEEE Trans. Audio and Electroacoustics,* Vol. AU-21, pp. 506-526, December 1973.
11. L. R. Rabiner, J. H. McClellan, and T. W. Parks, "FIR Digital Filter Design Techniques Using Weighted Chebyshev Approximation," *Proc. IEEE,* Vol. 63, No. 4, pp. 595-609, April 1975.
12. A. G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum p-Error Criterion," *IEEE Trans. Audio and Electroacoustics,* Vol. AU-20, No. 5, pp. 257-263, October 1972.
13. L. R. Rabiner, J. F. Kaiser, O. Herrmann, and M. T. Dolan, "Some Comparisons Between FIR and IIR Digital Filters," *Bell Syst. Tech. J.,* Vol. 53, No. 2, pp. 305-331, February 1974.
14. R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," *Proc. IEEE,* Vol. 61, No. 6, pp. 692-702, June 1973.
15. L. R. Rabiner and R. E. Crochiere, "A Novel Implementation for FIR Digital Filters," *IEEE Trans. Acoustics, Speech, and Signal Proc.,* Vol. ASSP-23, pp. 457-464, October 1975.
16. R. E. Crochiere and L. R. Rabiner, "Optimum FIR Digital Filter Implementation for Decimation, Interpolation and Narrowband Filters," *IEEE Trans. Acoust. Speech, and Signal Proc.,* Vol. ASSP-23, pp. 444-456, October 1975.
17. R. E. Crochiere and L. R. Rabiner, "Further Considerations in the Design of Decimators and Interpolators," *IEEE Trans. Acoustics, Speech, and Signal Processing,* Vol. ASSP-24, No. 4, pp. 269-311, August 1976.
18. D. J. Goodman, "Digital Filters for Code Format Conversion," *Electronics Letters,* Vol. 11, February 1975.

PROBLEMS

2.1 Consider the sequence

$$\begin{aligned} x(n) &= a^n \quad n \geq n_0 \\ &= 0 \quad n < n_0 \end{aligned}$$

(a) Find the z-transform of $x(n)$.
(b) Find the Fourier transform of $x(n)$. Under what conditions does the Fourier transform exist?

2.2 The input to a linear, time-invariant system is

$$\begin{aligned} x(n) &= 1 \quad 0 \leq n \leq N-1 \\ &= 0 \quad \textit{otherwise} \end{aligned}$$

The impulse response of the system is

$$\begin{aligned} h(n) &= a^n \quad n \geq 0 \\ &= 0 \quad n < 0 . \end{aligned}$$

(a) Using discrete convolution, find the output, $y(n)$, of the system for all n.
(b) Find the output using z-transforms.

2.3 Find the z-transform and the Fourier transform of each of the following sequences. (Each of these are commonly used as "windows" in speech processing systems.)
(1) Exponential window

$$\begin{aligned} w_1(n) &= a^n \quad 0 \leq n \leq N-1 \\ &= 0 \quad \textit{otherwise} \end{aligned}$$

(2) Rectangular window

$$\begin{aligned} w_2(n) &= 1 \quad 0 \leq n \leq N-1 \\ &= 0 \quad \textit{otherwise} \end{aligned}$$

(3) Hamming window

$$\begin{aligned} w_3(n) &= 0.54 - 0.46 \cos[2\pi n/(N-1)] \quad 0 \leq n \leq N-1 \\ &= 0 \quad \textit{otherwise} \end{aligned}$$

Sketch the magnitude of the Fourier transforms in each case. Hint: obtain a relationship between $W_3(e^{j\omega})$ and $W_2(e^{j\omega})$.

2.4 The frequency response of an ideal lowpass filter is

$$\begin{aligned} H(e^{j\omega}) &= 1 \quad |\omega| < \omega_c \\ &= 0 \quad \omega_c < |\omega| \leq \pi \end{aligned}$$

($H(e^{j\omega})$ is, of course, periodic with period 2π.)
(a) Find the impulse response of the ideal lowpass filter.
(b) Sketch the impulse response for $\omega_c = \pi/4$.
The frequency response of an ideal bandpass filter is

$$\begin{aligned} H(e^{j\omega}) &= 1 \quad \omega_a < |\omega| < \omega_b \\ &= 0 \quad |\omega| < \omega_a \text{ and } \omega_b < |\omega| \leq \pi , \end{aligned}$$

(c) Find the impulse response of the ideal bandpass filter.
(d) Sketch the impulse response for $\omega_a = \pi/4$ and $\omega_b = 3\pi/4$.

2.5 The frequency response of an ideal differentiator is

$$H(e^{j\omega}) = j\omega e^{-j\omega\tau} \quad -\pi < \omega < \pi$$

(This response is repeated with period 2π.) The quantity τ is the delay c the system in samples.
(a) Sketch the magnitude and phase response of this system.
(b) Find the impulse response, $h(n)$, of this system.
(c) The impulse response of this ideal system can be truncated to

length N samples by a window such as those in Problem 2.3. In so doing the delay is set equal to $\tau = (N-1)/2$ so that the ideal impulse response can be truncated symmetrically [1]. If $\tau = (N-1)/2$ and N is an odd integer, show that the ideal impulse response decreases as $1/n$. Sketch the ideal impulse response for the case $N = 11$.

(d) In the case that N is even, show that $h(n)$ decreases as $1/n^2$. Sketch the ideal impulse response for the case $N = 10$.

2.6 The frequency response of an ideal Hilbert transformer(90° phase shifter) with delay τ is

$$\begin{aligned} H(e^{j\omega}) &= -je^{-j\omega\tau} \quad 0 < \omega < \pi \\ &= je^{-j\omega\tau} \quad -\pi < \omega < 0 . \end{aligned}$$

Find and sketch the impulse response of this system.

2.7 Consider a linear phase FIR digital filter. The impulse response of such a filter has the property

$$\begin{aligned} h(n) &= h(N-1-n) \quad 0 \leqslant n \leqslant N-1 \\ &= 0 \quad \textit{otherwise} \end{aligned}$$

(a) Show that if N is an even integer the convolution sum expression for the output of such a system can be expressed as

$$y(n) = \sum_{k=0}^{(N-2)/2} h(k)[x(n-k)+x(n-N+1+k)]$$

and if N is odd

$$y(n) = \sum_{k=0}^{(N-3)/2} h(k)[x(n-k)+x(n-N+1+k)]+h((N-1)/2)x(n-(N-1)/2) .$$

Thus, the number of multiplications required to compute each output sample is essentially halved.

(b) Draw the digital filter structures for each of the above equations.

2.8 Consider the first order system

$$y(n) = \alpha y(n-1) + x(n)$$

(a) Find the system function, $H(z)$, for this system.

(b) Find the impulse response of this system.

(c) For what values of α will the system be stable?

(d) Assume that the input is obtained by sampling with period T. Find the value of α such that

$$h(n) < e^{-1} \quad \text{for} \quad nT < 2 \text{ msec}$$

i.e., find the value of α that gives a time constant of 2 msec.

2.9 Consider a system function of the form of Eq. (2.24)

(a) Show that if $M < N$, $H(z)$ can be expressed as a partial fraction

expansion as in Eq. (2.30), where the coefficients A_m can be found from

$$A_m = H(z)(1-d_m z^{-1})|_{z=d_m} \quad m = 1,2,...,N$$

(b) Show that the z-transform of the sequence $A_k(d_k)^n u(n)$ is

$$\frac{A_k}{1-d_k z^{-1}} \quad |z| > |d_k| ,$$

and thus $h(n)$ is given by Eq. (2.31).

2.10 Consider two linear shift-invariant systems in cascade as shown in Fig. P2.10 – i.e., the output of the first system is the input to the second.

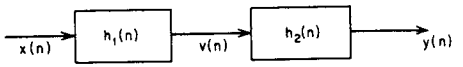


Fig. P2.10

(a) Show that the impulse response of the overall system is

$$h(n) = h_1(n) * h_2(n).$$

(b) Show that

$$h_1(n) * h_2(n) = h_2(n) * h_1(n)$$

and thus that the overall response does not depend on the order in which the systems are cascaded.

(c) Consider the system function of Eq. (2.23) written as

$$\begin{aligned} H(z) &= \left[\sum_{r=0}^{M} b_r z^{-r}\right]\left[\frac{1}{1-\sum_{k=1}^{N} a_k z^{-k}}\right] \\ &= H_1(z)\cdot H_2(z) \end{aligned}$$

i.e., as a cascade of two systems. Write the difference equations for the overall system from this point of view.

(d) Now consider the two systems of part (c) in the opposite order; i.e.,

$$H(z) = H_2(z)H_1(z)$$

Show that the difference equations of Eq. (2.32) result.

2.11 For the difference equation

$$y(n) = 2\cos(bT)y(n-1) - y(n-2)$$

find the two initial conditions $y(-1)$ and $y(-2)$ such that

(a) $y(n) = \cos(bTn) \quad n \geqslant 0$

(b) $y(n) = \sin(bTn) \quad n \geqslant 0$

2.12 Consider the set of difference equations

$$y_1(n) = Ay_1(n-1) + By_2(n-1) + x(n)$$
$$y_2(n) = Cy_1(n-1) + Dy_2(n-1)$$

(a) Draw the network diagram for this system.
(b) Find the transfer functions

$$H_1(z) = \frac{Y_1(z)}{X(z)} \quad \text{and} \quad H_2(z) = \frac{Y_2(z)}{X(z)}$$

(c) For the case $A = D = r\cos\theta$ and $C = -B = r\sin\theta$, find the impulse responses $h_1(n)$ and $h_2(n)$ that result when the system is excited by $x(n) = \delta(n)$.

2.13 A causal linear shift invariant system has the system function

$$H(z) = \frac{(1+2z^{-1}+z^{-2})(1+2z^{-1}+z^{-2})}{(1+\frac{7}{8}z^{-1}+\frac{5}{16}z^{-2})(1+\frac{3}{4}z^{-1}+\frac{7}{8}z^{-2})}$$

(a) Draw a digital network diagram of an implementation of this system in
(i) Cascade form
(ii) Direct form.
(b) In this system stable? Explain.

2.14 For the system of Fig. P2.14,
(a) Write the difference equations represented by the network.
(b) Find the system function for the network.

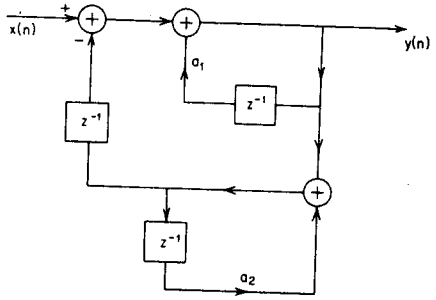


Fig. P2.14

2.15 Find a_1, a_2, and a_3 in terms of b_1, and b_2 so that the two networks of Fig P2.15 have the same transfer function.

2.16 The system function for a simple resonator is of the form

$$H(z) = \frac{1 - 2e^{-aT}\cos(bT) + e^{-2aT}}{1 - 2e^{-aT}\cos(bT)z^{-1} + e^{-2aT}z^{-2}}$$

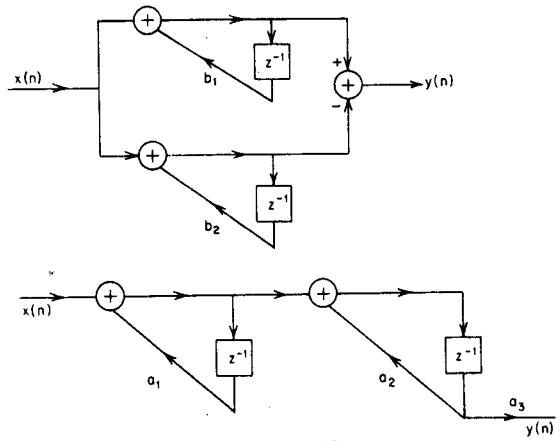


Fig. P2.15

(a) Find the poles and zeros of $H(z)$ and plot them in the z-plane.
(b) Find the impulse response of this system and sketch it for the con stants

$$T = 10^{-4}$$
$$b = 1000\pi$$
$$a = 200\pi$$

(c) Sketch the frequency response of this system as a function of analo frequency, Ω.

2.17 Consider the finite length sequence

$$x(n) = \delta(n) + 0.5\delta(n-5)$$

(a) Find the z-transform and Fourier transform of $x(n)$.
(b) Find the N-point DFT of $x(n)$ for $N = 50$, 10 and 5.
(c) How are the DFT values for $N = 5$ related to those of the DFT fo $N = 50$?
(d) What is the relationship between the N-point DFT of $x(n)$ and th Fourier transform of $x(n)$?

2.18 A speech signal is sampled at a rate of 20000 samples/sec (20 kHz). segment of length 1024 samples is selected and the 1024-point DFT i computed.
(a) What is the time duration of the segment of speech?
(b) What is the frequency resolution (spacing in Hz) between the DF values?
(c) How do your answers to parts (a) and (b) change if we compute th 1024-point DFT of 512 samples of the speech signal. (The 512 san ples would be augmented with 512 zero samples before th transform was computed.)

3

Digital Models for the Speech Signal

3.0 Introduction

In order to apply digital signal processing techniques to speech communication problems, it is essential to understand the fundamentals of the speech production process as well as the fundamentals of digital signal processing. This chapter provides a review of the acoustic theory of speech production and shows how this theory leads to a variety of ways of representing the speech signal. Specifically, we shall be concerned with obtaining discrete-time models for representing sampled speech signals. These models will serve as a basis for application of digital processing techniques.

This chapter plays a role similar to that of Chapter 2 in serving as a review of an established area of knowledge. Several excellent references provide much more detail on many of the topics of this chapter [1-5]. Particularly noteworthy are the books by Fant [1] and Flanagan [2]. Fant's book deals primarily with the acoustics of speech production and contains a great deal of useful data on vocal system measurements and models. Flanagan's book, which is much broader in scope, contains a wealth of valuable insights into the physical modelling of the speech production process and the way that such models are used in representing and processing speech signals. These books are indispensable to the serious student of speech communication.

Before discussing the acoustic theory and the resulting mathematical models for speech production, it is necessary to consider the various types of sounds that make up human speech. Thus, this chapter begins with a very

brief introduction to acoustic phonetics in the form of a summary of the phonemes of English and a discussion of the place and manner of articulation for each of the major phoneme classes. Then the fundamentals of the acoustic theory of speech production are presented. Topics considered include sound propagation in the vocal tract, transmission line analogies, and the steady state behaviour of the vocal system in the production of a single sustained sound. This theory provides the basis for the classical approach to modelling the speech signal as the output of a time-varying linear system (vocal tract) excited by either random noise or a quasi-periodic sequence of pulses. This approach is applied to obtain discrete time models for the speech signal. These models which are justified in terms of the acoustic theory and formulated in terms of digital filtering principles, serve as the basis for discussion of speech processing techniques throughout the remainder of this book.

3.1 The Process of Speech Production

Speech signals are composed of a sequence of sounds. These sounds and the transitions between them serve as a symbolic representation of information. The arrangement of these sounds (symbols) is governed by the rules of language. The study of these rules and their implications in human communication is the domain of *linguistics,* and the study and classification of the sounds of speech is called *phonetics.* A detailed discussion of phonetics and linguistics would take us too far afield. However, in processing speech signals to enhance or extract information, it is helpful to have as much knowledge as possible about the structure of the signal; i.e., about the way in which information is encoded in the signal. Thus, it is worthwhile to discuss the main classes of speech sounds before proceeding to a detailed discussion of mathematical models of the production of speech signals. Although this will be all that we shall have to say about linguistics and phonetics, this is not meant to minimize their importance – especially in the areas of speech recognition and speech synthesis.

3.1.1 The mechanism of speech production

Figure 3.1 is an X-ray photograph which places in evidence the important features of the human vocal system [6]. The *vocal tract,* outlined by the dotted lines in Fig. 3.1, begins at the opening between the vocal cords, or *glottis,* and ends at the lips. The vocal tract thus consists of the *pharynx* (the connection from the esophagus to the mouth) and the mouth or *oral cavity.* In the average male, the total length of the vocal tract is about 17 cm. The cross-section area of the vocal tract, determined by the positions of the tongue, lips, jaw, and velum varies from zero (complete closure) to about 20cm^2. The *nasal tract* begins at the velum and ends at the nostrils. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

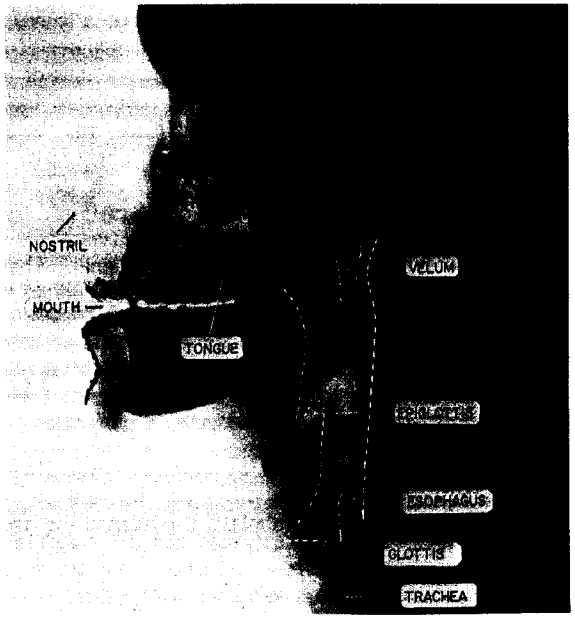


Fig. 3.1 Saggital plane X-ray of the human vocal apparatus. (After Flanagan et al. [6].)

In studying the speech production process, it is helpful to abstract the important features of the physical system in a manner which leads to a realistic yet tractable mathematical model. Figure 3.2 shows such a schematic diagram of the vocal system [6]. For completeness the diagram includes the sub-glottal system composed of the lungs, bronchi and trachea. This sub-glottal system serves as a source of energy for the production of speech. Speech is simply the acoustic wave that is radiated from this system when air is expelled from the lungs and the resulting flow of air is perturbed by a constriction somewhere in the vocal tract. As an example of a speech wave, Figure 3.3a shows the waveform of the utterance, "should we cha(se)," spoken by a male speaker. The general features of this waveform can be readily explained by a more detailed consideration of the mechanism of speech production.

Speech sounds can be classified into 3 distinct classes according to their mode of excitation. *Voiced* sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract. Voiced segments are labelled /U/, /d/, /w/, /i/ and /e/ in Fig. 3.3a. *Fricative or unvoiced sounds* are generated by forming a constriction at some point in the vocal tract (usually toward the mouth end), and forcing air through the constriction at a high enough velocity to produce turbulence. This

creates a broad-spectrum noise source to excite the vocal tract. The segment labelled /$\int$/ in Fig. 3.3a is the fricative "sh." *Plosive sounds* result from making a complete closure (again, usually toward the front of the vocal tract), building up pressure behind the closure, and abruptly releasing it. Plosive excitation is involved in creating the sound labelled /$t\int$/ at the beginning of the fourth line of Fig. 3.3a. Note the gap (region of very small amplitude) at the end of the third line which precedes the burst of noise-like waveform. This gap corresponds to the time of complete closure of the vocal tract.

The vocal tract and nasal tract are shown in Figure 3.2 as tubes of nonuniform cross-sectional area. As sound, generated as discussed above, propagates down these tubes, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed with organ pipes or wind instruments. In the context of speech production, the resonance frequencies of the vocal tract tube are called *formant frequencies* or simply *formants*. The formant frequencies depend upon the shape and dimensions of the vocal tract; each shape is characterized by a set of formant frequencies. Different sounds are formed by varying the shape of the vocal tract. Thus, the spectral properties of the speech signal vary with time as the vocal tract shape varies.

The time-varying spectral characteristics of the speech signal can be graphically displayed through the use of the sound spectrograph [2,7]. This device produces a two-dimensional pattern called a *spectrogram* in which the vertical dimension corresponds to frequency and the horizontal dimension to time. The darkness of the pattern is proportional to signal energy. Thus, the resonance frequencies of the vocal tract show up as dark bands in the spectrogram. Voiced regions are characterized by a striated appearance due to the periodicity of the time waveform, while unvoiced intervals are more solidly filled in. A spectrogram of the utterance of Fig. 3.3a is shown in Figure 3.3b. The spectrogram is labelled to correspond to the labelling of Fig. 3.3a so that the time domain and frequency domain features can be correlated.

The sound spectrograph has long been a principal tool in speech research, and although more flexible displays can be generated using digital processing techniques (see Chapter 6), its basic principles are still widely used. An early,

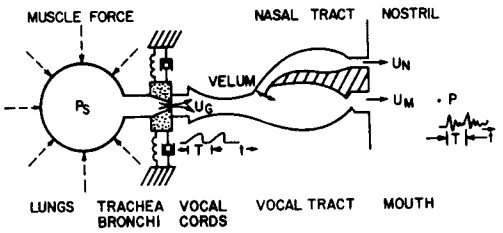


Fig. 3.2 Schematized diagram of the vocal apparatus (After Flanagan et al. [6].)

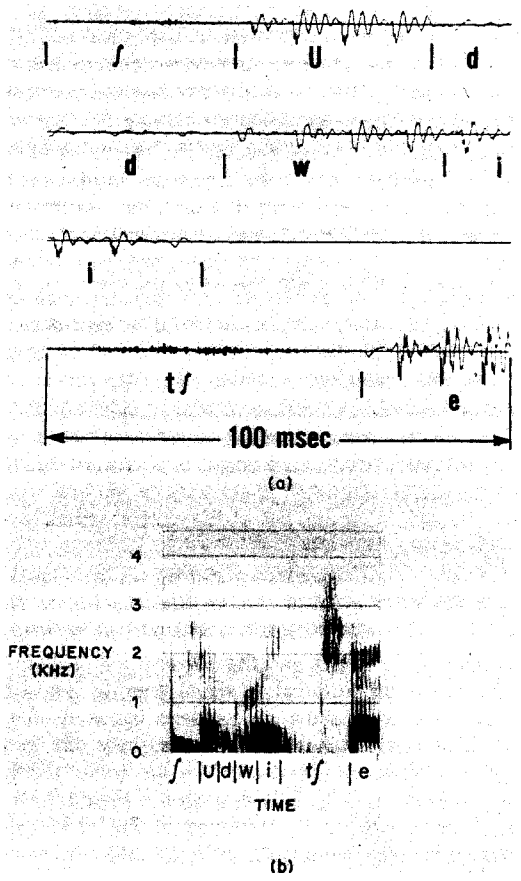


Fig. 3.3 (a) Waveform of the utterance "Should we cha(se)"; (b) corresponding spectrogram.

but still very useful, reference on spectrographic representations of speech is the book *Visible Speech* [8]. This book, although written for the purpose of teaching people literally to "read" spectrograms, provides an excellent introduction to acoustic phonetics.

3.1.2 Acoustic phonetics

Most languages, including English, can be described in terms of a set of distinctive sounds, or *phonemes*. In particular, for American English, there are about 42 phonemes including vowels, diphthongs, semivowels and consonants. There are a variety of ways of studying phonetics; e.g., linguists study the distinctive features or characteristics of the phonemes [9,10]. For our purposes it is sufficient to consider an acoustic characterization of the various sounds including the place and manner of articulation, waveforms, and spectrographic characterizations of these sounds.

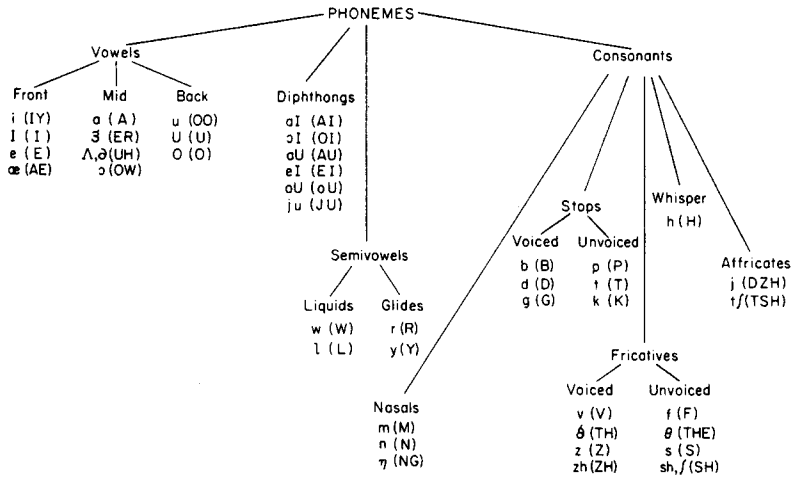
Table 3.1 shows how the sounds of American English are broken into phoneme classes.[1] The four broad classes of sounds are vowels, diphthongs, semivowels, and consonants. Each of these classes may be further broken down into sub-classes which are related to the manner, and place of articulation of the sound within the vocal tract.

Each of the phonemes in Table 3.1 can be classified as either a continuant, or a noncontinuant sound. Continuant sounds are produced by a fixed (non-time-varying) vocal tract configuration excited by the appropriate source. The class of continuant sounds includes the vowels, the fricatives (both unvoiced and voiced), and the nasals. The remaining sounds (diphthongs, semivowels, stops and affricates) are produced by a changing vocal tract configuration. These are therefore classed as noncontinuants.

3.1.2a Vowels

Vowels are produced by exciting a fixed vocal tract with quasi-periodic pulses of air caused by vibration of the vocal cords. As we shall see later in this chapter, the way in which the cross-sectional area varies along the vocal

Table 3.1 Phonemes in American English.



[1]Table 3.1 gives both a phonetic representation and an orthographic representation for each phoneme. The phonetic and orthographic representations are used interchangably throughout this text.

tract determines the resonant frequencies of the tract (formants) and thus the sound that is produced. The dependence of cross-sectional area upon distance along the tract is called the *area function* of the vocal tract. The area function for a particular vowel is determined primarily by the position of the tongue, but the positions of the jaw, lips, and, to a small extent, the velum also influence the resulting sound. For example, in forming the vowel /a/ as in "father," the vocal tract is open at the front and somewhat constricted at the back by the main body of the tongue. In contrast, the vowel /i/ as in "eve" is formed by raising the tongue toward the palate, thus causing a constriction at the front and increasing the opening at the back of the vocal tract. Thus, each vowel sound can be characterized by the vocal tract configuration (area function) that is used in its production. It is obvious that this is a rather imprecise characterization because of the inherent differences between the vocal tracts of speakers. An alternative representation is in terms of the resonance frequencies of the vocal tract. Again a great deal of variability is to be expected among speakers producing the same vowel. Peterson and Barney [11] measured the formant (resonance) frequencies (using a sound spectrograph) of vowels that were perceived to be equivalent. Their results are shown in Fig. 3.4 which is a plot of second

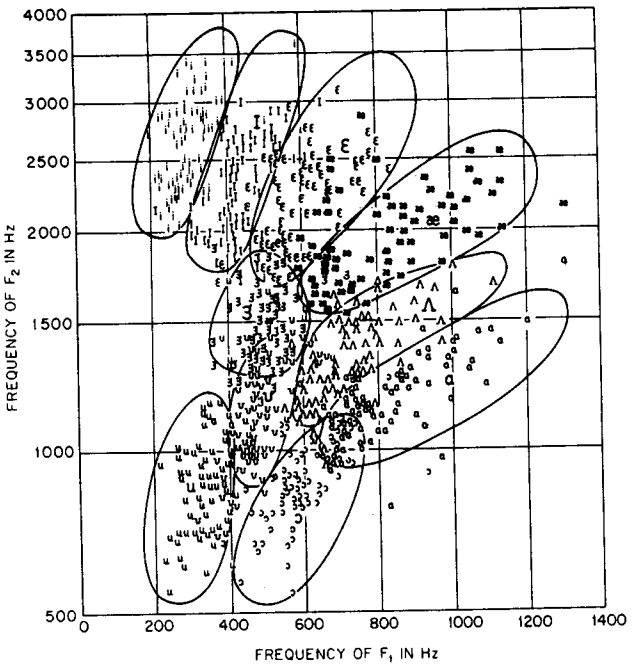


Fig. 3.4 Plot of second formant frequency versus first formant frequency for vowels by a wide range of speakers. (After Peterson and Barney [11].)

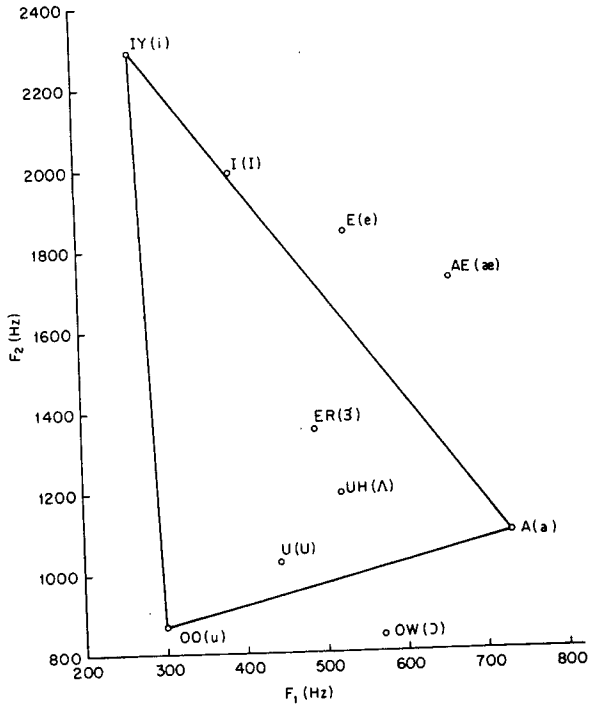


Fig. 3.5 The vowel triangle.

formant frequency as a function of first formant frequency for several vowels spoken by men and children. The broad ellipses in Figure 3.4 show the approximate range of variation in formant frequencies for each of these vowels. Table 3.2 gives average values of the first three formant frequencies of the vowels for male speakers. Although a great deal of variation clearly exists in the vowel formants, the data of Table 3.2 serve as a useful characterization of the vowels.

Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	ɪ	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	ɑ	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

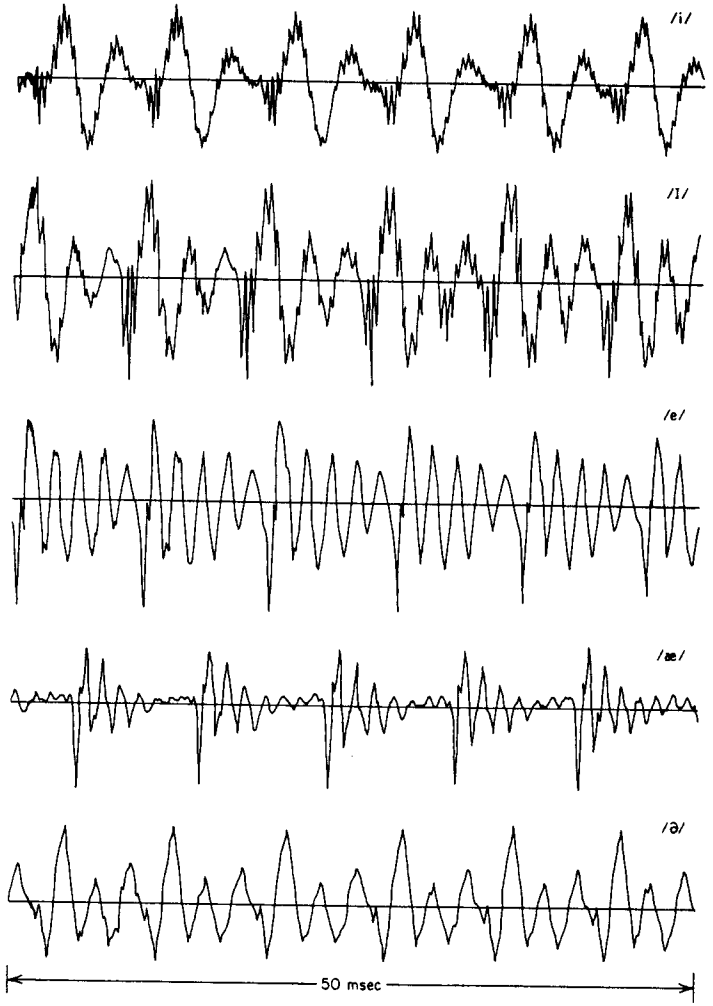


Fig. 3.6 The acoustic waveforms for several American English vowels and corresponding spectrograms.

Figure 3.5 shows a plot of the second formant frequency versus the first formant frequency for the vowels of Table 3.2. The so-called "vowel triangle" is readily seen in this figure. At the upper left hand corner of the triangle is the vowel /i/ with a low first formant, and a high second formant. At the lower left hand corner is the vowel /u/ with low first and second formants. The third vertex of the triangle is the vowel /a/ with a high first formant, and a low second formant. Later in this chapter we shall see how vocal tract shape affects the formant frequencies of vowels.

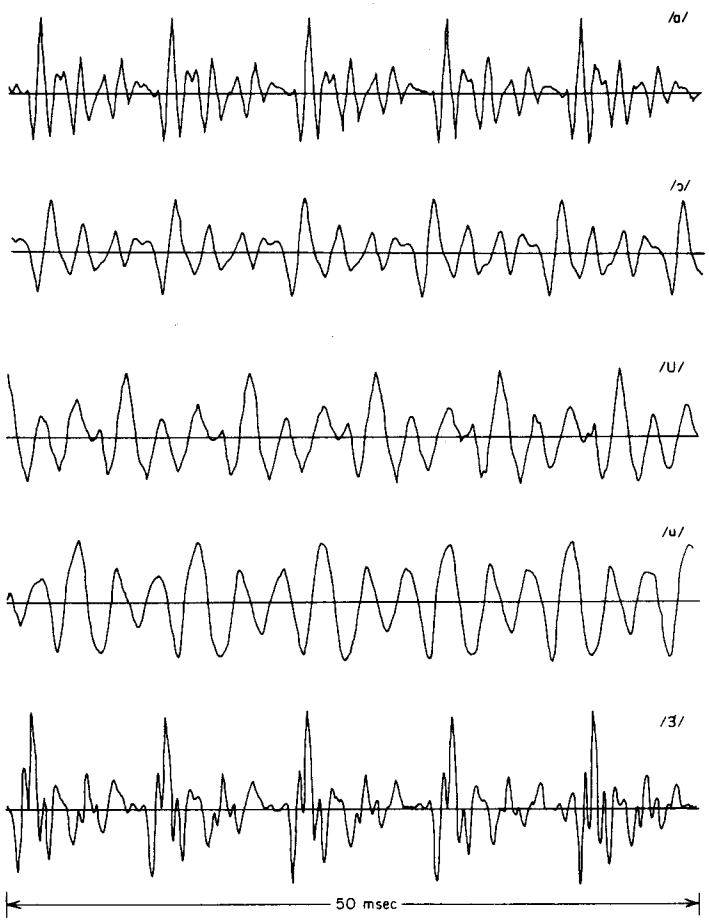


Fig. 3.6 (Continued)

The acoustic waveforms and spectrograms for each of the vowels of English are shown in Fig. 3.6. The spectrograms clearly show a different pattern of resonances for each vowel. The acoustic waveforms, in addition to showing the periodicity characteristic of voiced sounds, also display the gross spectral properties if a single "period" is considered. For example, the vowel /i/ shows a low frequency damped oscillation upon which is superimposed a relatively strong high frequency oscillation. This is consistent with a low first formant and high second and third formant (see Table 3.2). (Two resonances in proximity tend to boost the spectrum.) In contrast the vowel /u/ shows relatively little high frequency energy as a consequence of the low first and second formant frequencies. Similar correspondences can be observed for all the vowels in Fig. 3.6.

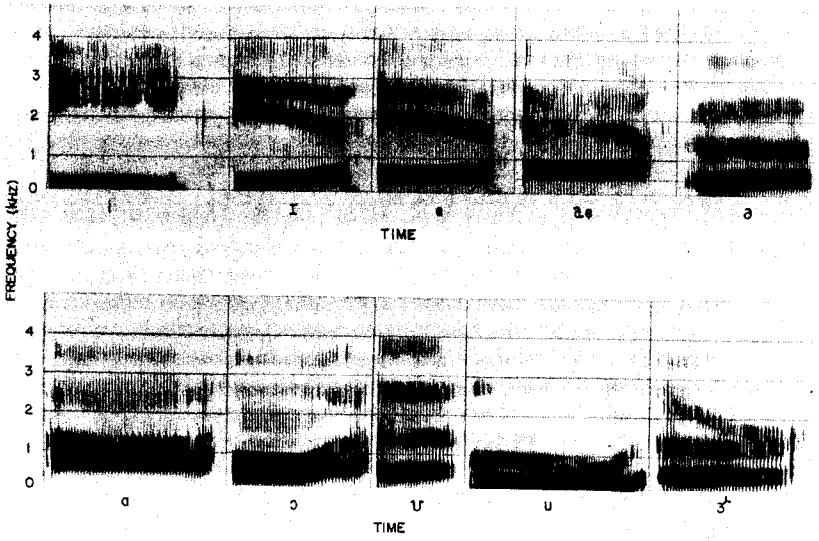


Fig. 3.6 (Continued)

3.1.2b Diphthongs

Although there is some ambiguity and disagreement as to what is and what is not a diphthong, a reasonable definition is that a diphthong is a gliding monosyllabic speech item that starts at or near the articulatory position for one vowel and moves to or toward the position for another. According to this definition, there are six diphthongs in American English including /eI/ (as in bay), /oU/ as in (boat), /aI/ (as in buy), /aU/ (as in how), /oI/ (as in boy), and /ju/ (as in you).

The diphthongs are produced by varying the vocal tract smoothly between vowel configurations appropriate to the diphthong. To illustrate this point, Figure 3.7 shows a plot of measurements of the second formant versus the first formant (as a function of time) for the diphthongs [12]. The arrows in this figure indicate the direction of motion of the formants (in the (F_1-F_2) plane) as time increases. The dashed circles in this figure indicate average positions of the vowels. Based on these data, and other measurements, the diphthongs can be characterized by a time varying vocal tract area function which varies between two vowel configurations.

3.1.2c Semivowels

The group of sounds consisting of /w/, /l/, /r/, and /y/ is quite difficult to characterize. These sounds are called semivowels because of their vowel-like nature. They are generally characterized by a gliding transition in vocal tract

area function between adjacent phonemes. Thus the acoustic characteristics of these sounds are strongly influenced by the context in which they occur. For our purposes they are best described as transitional, vowel-like sounds, and hence are similar in nature to the vowels and diphthongs. An example of the semivowel /w/ is shown in Figure 3.3.

3.1.2d Nasals

The nasal consonants /m/, /n/, and /ŋ/ are produced with glottal excitation and the vocal tract totally constricted at some point along the oral passageway. The velum is lowered so that air flows through the nasal tract, with sound being radiated at the nostrils. The oral cavity, although constricted toward the front, is still acoustically coupled to the pharynx. Thus, the mouth serves as a resonant cavity that traps acoustic energy at certain natural frequencies. As far as the radiated sound is concerned these resonant frequencies of the oral cavity appear as anti-resonances, or zeros of sound transmission [2]. Furthermore, nasal consonants and nasalized vowels (i.e., some vowels preceding or following nasal consonants) are characterized by resonances which are spectrally broader, or more highly damped, than those for vowels. The broadening of the nasal

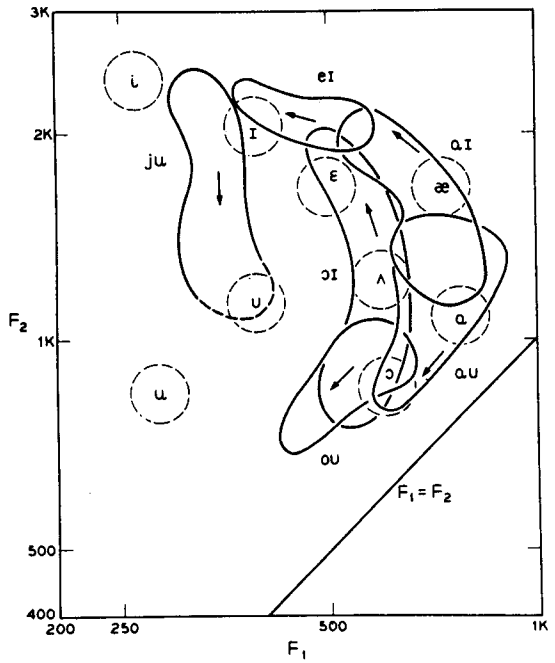


Fig. 3.7 Time variations of the first two formants for diphthongs. (After Holbrook and Fairbanks [27].)

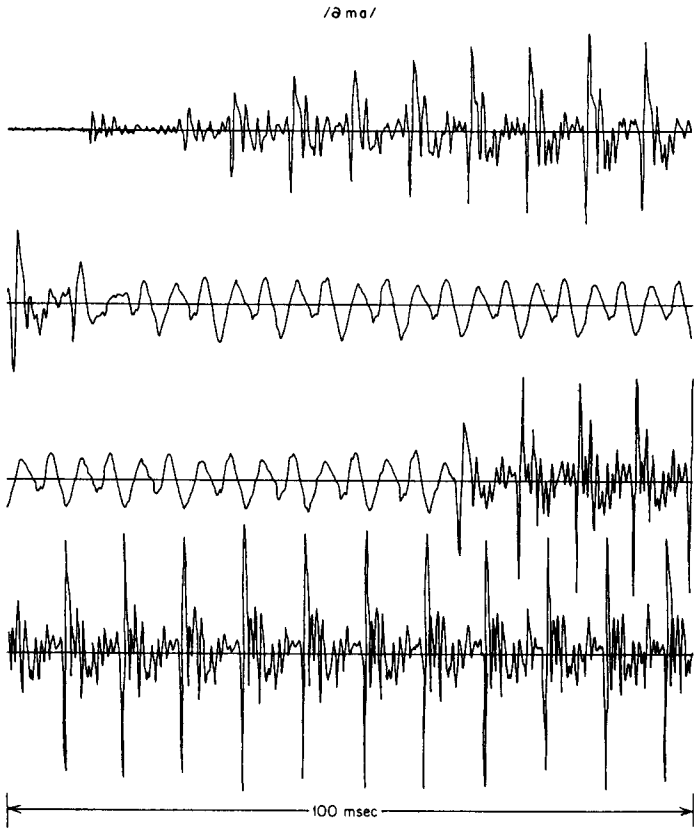


Fig. 3.8 Acoustic waveforms and spectrograms for utterances */UH–M–A/* and */UH–N–A/*.

resonances is due to the fact that the inner surface of the nasal tract is convoluted, so that the nasal cavity has a relatively large ratio of surface area to cross-sectional area. Therefore, heat conduction and viscous losses are larger than normal.

The three nasal consonants are distinguished by the place along the oral tract at which a total constriction is made. For /m/, the constriction, is at the lips; for /n/ the constriction is just back of the teeth; and for /η/ the constriction is just forward of the velum itself. Figure 3.8 shows typical speech waveforms and spectrograms for two nasal consonants in the context vowel-nasal-vowel. It is clear that the waveforms of /m/ and /n/ look very similar. The spectrograms show a concentration of low frequency energy with a mid-range of frequencies that contains no prominent peaks. This is because of the particular combination of resonances and anti-resonances that result from the coupling of the nasal and oral tracts [13].

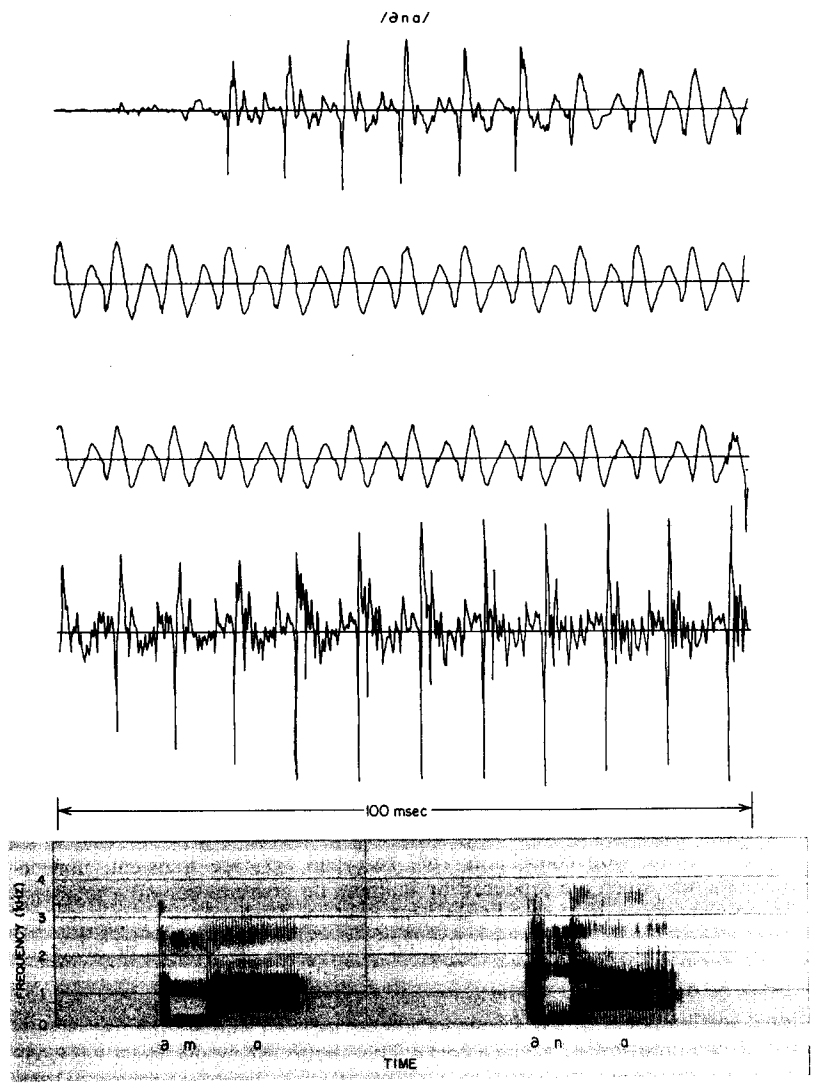


Fig. 3.8 (Continued)

3.1.2e Unvoiced Fricatives

The unvoiced fricatives /f/, /θ/, /s/, and /sh/ are produced by exciting the vocal tract by a steady air flow which becomes turbulent in the region of a constriction in the vocal tract. The location of the constriction serves to determine which fricative sound is produced. For the fricative /f/ the constriction is

near the lips; for /θ/ it is near the teeth; for /s/ it is near the middle of the oral tract; and for /sh/ it is near the back of the oral tract. Thus the system for producing unvoiced fricatives consists of a source of noise at a constriction, which separates the vocal tract into two cavities. Sound is radiated from the lips; i.e. from the front cavity. The back cavity serves, as in the case of nasals, to trap energy and thereby introduce anti-resonances into the vocal output [2,14]. Figure 3.9 shows the waveforms and spectrograms of the fricatives /f/, /s/ and /sh/. The nonperiodic nature of fricative excitation is obvious in the waveform plots. The spectral differences among the fricatives are readily seen by comparing the three spectrograms.

3.1.2f Voiced Fricatives

The voiced fricatives /v/, /th/, /z/ and /zh/ are the counterparts of the unvoiced fricatives /f/, /θ/, /s/, and /sh/, respectively, in that the place of constriction for each of the corresponding phonemes is essentially identical. However, the voiced fricatives differ markedly from their unvoiced counterparts in that two excitation sources are involved in their production. For voiced fricatives the vocal cords are vibrating, and thus one excitation source is at the glottis. However, since the vocal tract is constricted at some point forward of the glottis, the air flow becomes turbulent in the neighborhood of the constriction. Thus the spectra of voiced fricatives can be expected to display two distinct components. These excitation features are readily observed in Figure 3.10 which shows typical waveforms and spectra for several voiced fricatives. The similarity of the unvoiced fricative /f/ to the voiced fricative /v/ is easily seen by comparing their corresponding spectrograms in Figures 3.9 and 3.10. Likewise it is instructive to compare the spectrograms of /sh/ and /zh/.

3.1.2g Voiced Stops

The voiced stop consonants /b/, /d/, and /g/, are transient, noncontinuant sounds which are produced by building up pressure behind a total constriction somewhere in the oral tract, and suddenly releasing the pressure. For /b/ the constriction is at the lips; for /d/ the constriction is back of the teeth; and for /g/ it is near the velum. During the period when there is a total constriction in the tract there is no sound radiated from the lips. However, there is often a small amount of low frequency energy radiated through the walls of the throat (sometimes called a voice bar). This occurs when the vocal cords are able to vibrate even though the vocal tract is closed at some point.

Since the stop sounds are dynamical in nature, their properties are highly influenced by the vowel which follows the stop consonant [15]. As such, the waveforms for stop consonants give little information about the particular stop consonant. Figure 3.11 shows the waveform and spectrogram of the syllable /UH-B-A/. The waveform of /b/ shows few distinguishing features except for the voiced excitation and lack of high frequency energy.

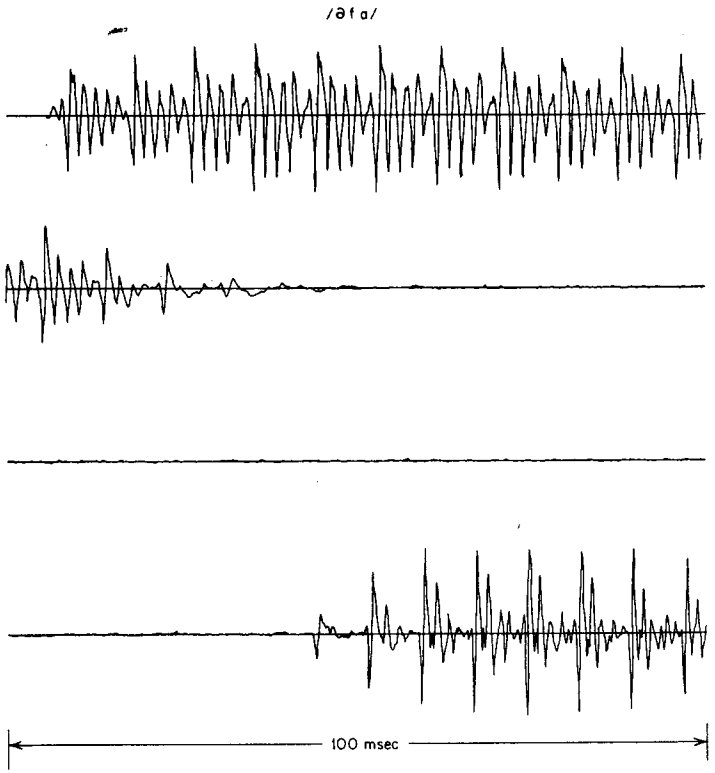


Fig. 3.9 Acoustic waveforms and spectrograms for */UH−F−A/*, */UH−S−A/*, and */UH−SH−A/*.

3.1.2h Unvoiced Stops

The unvoiced stop consonants /p/, /t/, and /k/ are similar to their voiced counterparts /b/, /d/, and /g/ with one major exception. During the period of total closure of the tract, as the pressure builds up, the vocal cords do not vibrate. Thus, following the period of closure, as the air pressure is released, there is a brief interval of friction (due to sudden turbulence of the escaping air) followed by a period of aspiration (steady air flow from the glottis exciting the resonances of the vocal tract) before voiced excitation begins.

Figure 3.12 shows waveforms and spectrograms of the voiceless stop consonants /p/ and /t/. The "stop gap," or time interval during which the pressure is built up is clearly in evidence. Also, it is readily seen that the duration and frequency content of the frication noise and aspiration varies greatly with the stop consonant.

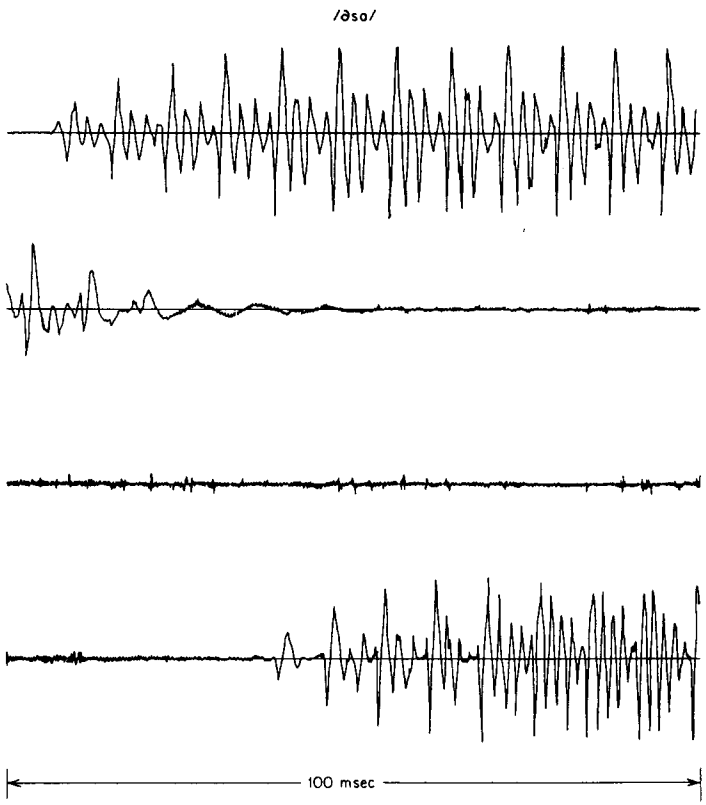


Fig. 3.9 (Continued)

3.1.2i Affricates and /h/

The remaining consonants of American English are the affricates /t∫/ and /j/, and the phoneme /h/. The unvoiced affricate /t∫/ is a dynamical sound which can be modelled as the concatenation of the stop /t/ and the fricative /∫/. (See Fig. 3.3a for an example.) The voiced affricate /j/ can be modelled as the concatenation of the stop /d/ and the fricative /zh/. Finally, the phoneme /h/ is produced by exciting the vocal tract by a steady air flow – i.e., without the vocal cords vibrating, but with turbulent flow being produced at the glottis.[2] The characteristics of /h/ are invariably those of the vowel which follows /h/ since the vocal tract assumes the position for the following vowel during the production of /h/.

[2]Note that this is also the mode of excitation for whispered speech.

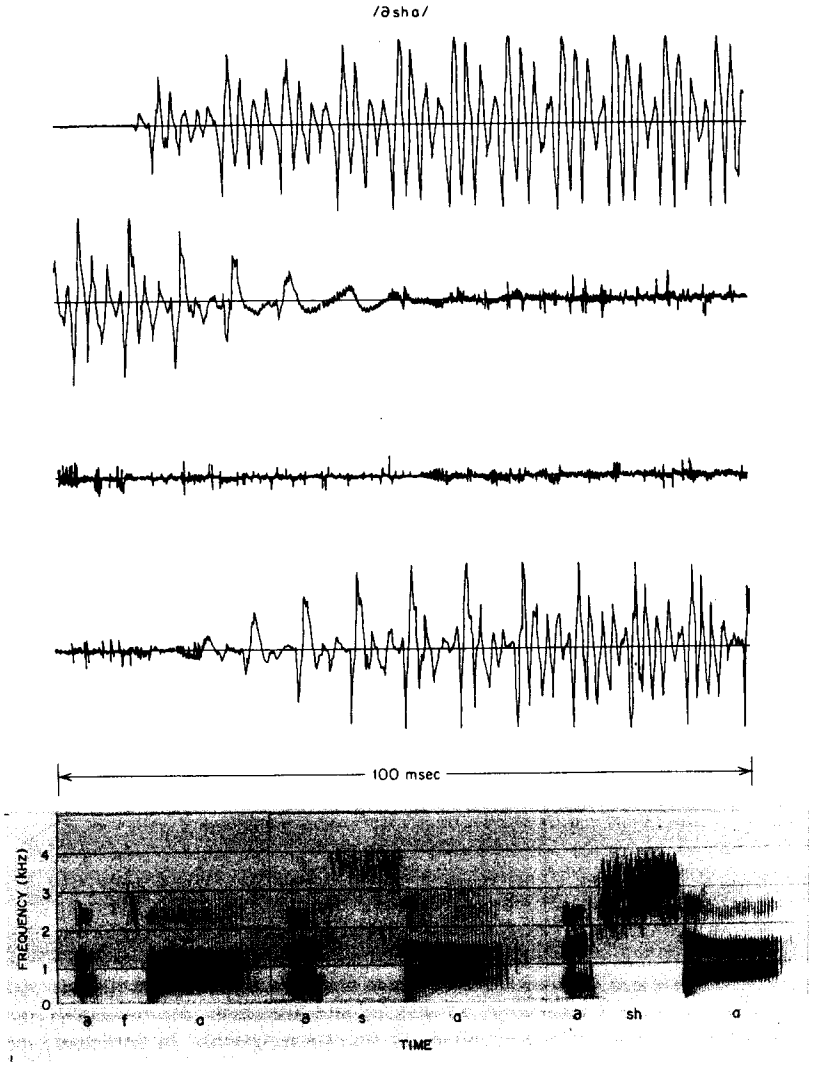


Fig. 3.9 (Continued)

3.2 The Acoustic Theory of Speech Production

The previous section was a review of the qualitative description of the sounds of speech and the way that they are produced. In this section we shall consider mathematical representations of the process of speech production. Such mathematical representations serve as the basis for the analysis and synthesis of speech.

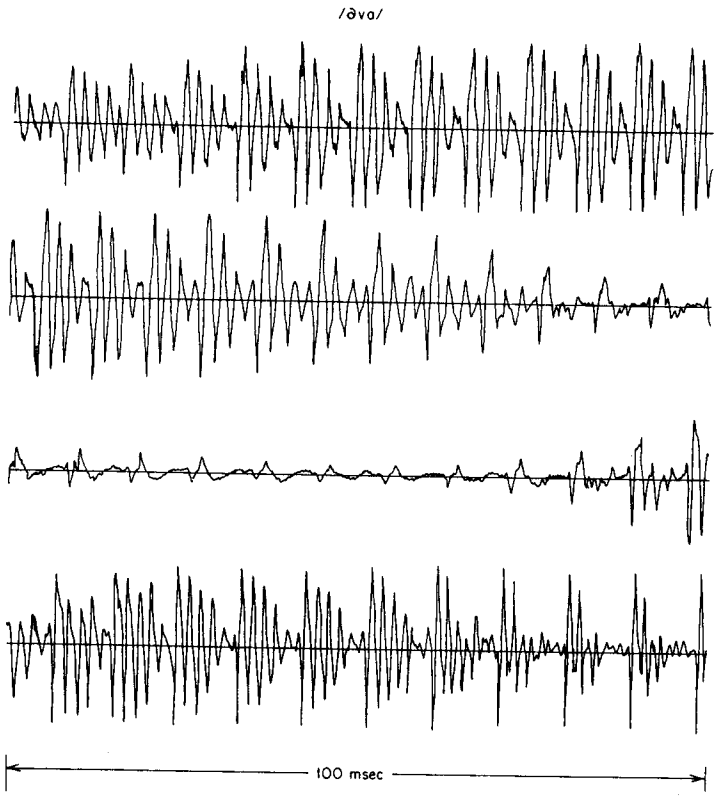


Fig. 3.10 Acoustic waveforms and spectrograms for utterances */UH–V–A/* and */UH–ZH–A/*.

3.2.1 Sound propagation

Sound is almost synonymous with vibration. Sound waves are created by vibration and are propagated in air or other media by vibrations of the particles of the media. Thus, the laws of physics are the basis for describing the generation and propagation of sound in the vocal system. In particular, the fundamental laws of conservation of mass, conservation of momentum, and conservation of energy along with the laws of thermodynamics and fluid mechanics, all apply to the compressible, low viscosity fluid (air) that is the medium for sound propagation in speech. Using these physical principles, a set of partial differential equations can be obtained that describe the motion of air in the vocal system [16-20]. The formulation and solution of these equations is extremely difficult except under very simple assumptions about vocal tract shape and energy losses in the vocal system. A detailed acoustic theory must consider the effects of the following:

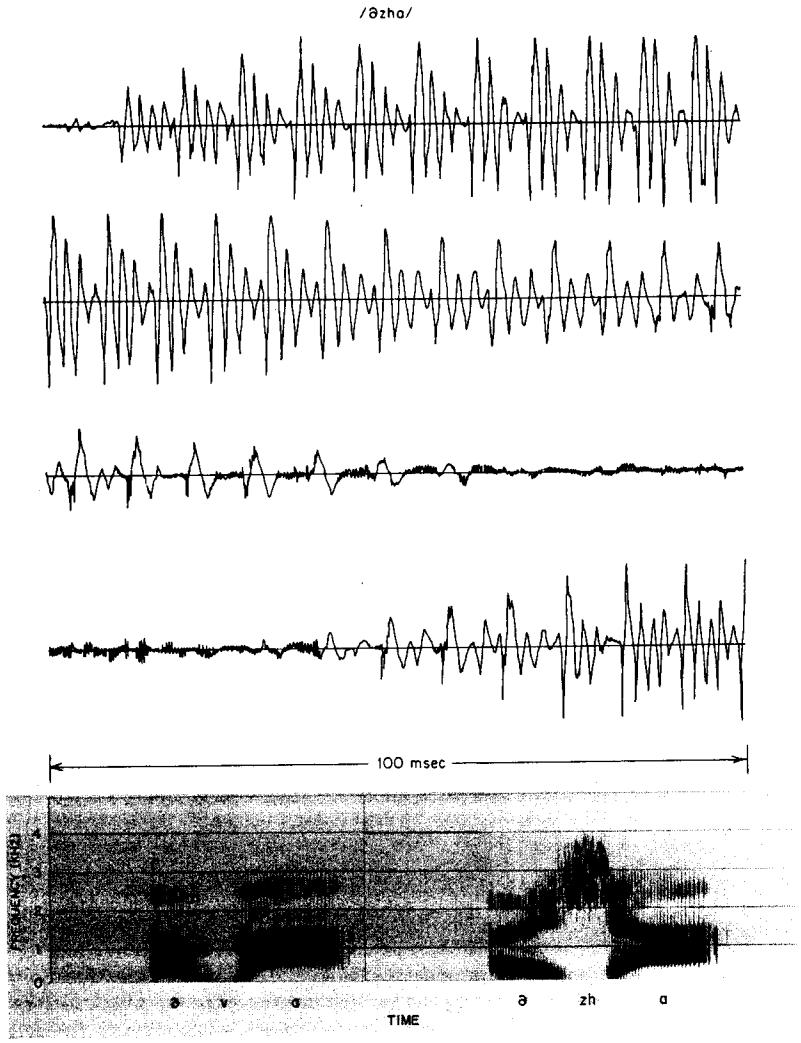


Fig. 3.10 (Continued)

1. Time variation of the vocal tract shape.
2. Losses due to heat conduction and viscous friction at the vocal tract walls.
3. Softness of the vocal tract walls.
4. Radiation of sound at the lips.
5. Nasal coupling.
6. Excitation of sound in the vocal tract.

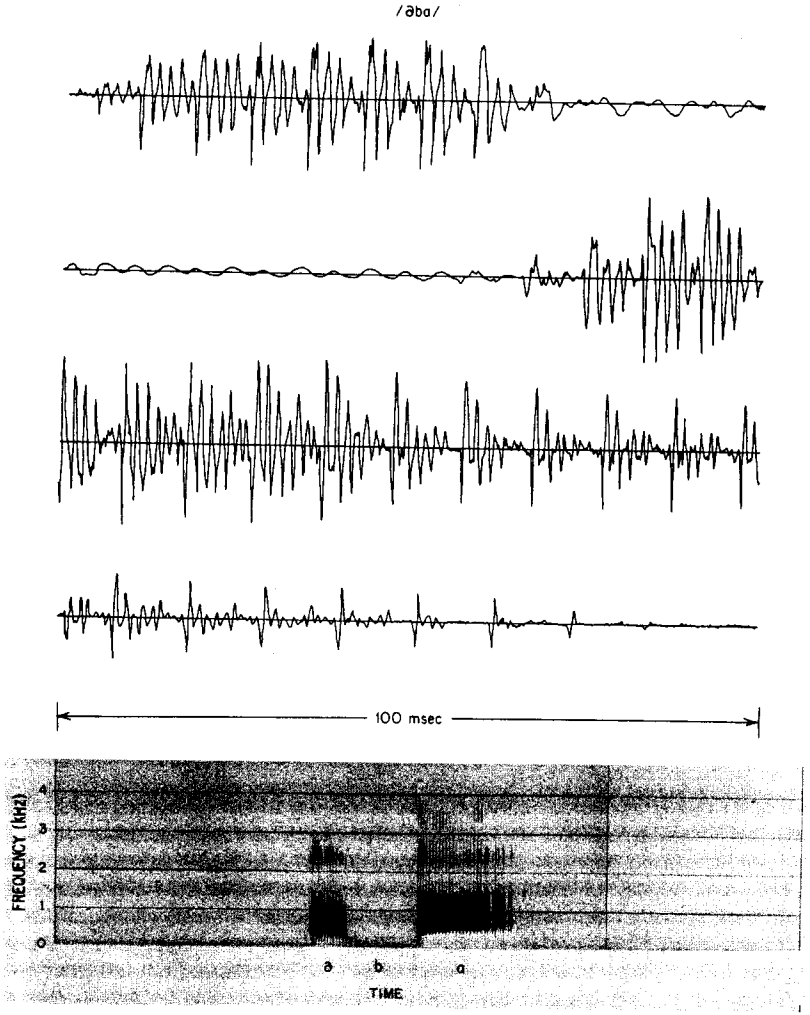


Fig. 3.11 Acoustic waveform and spectrogram for utterance */UH−B−A/*.

A completely detailed acoustic theory incorporating all the above effects is beyond the scope of this chapter, and indeed, such a theory is not yet available. We must be content to survey these factors, providing references to details when available, and qualitative discussions when suitable references are unavailable.

The simplest physical configuration that has a useful interpretation in terms of the speech production process is depicted in Figure 3.13a. The vocal

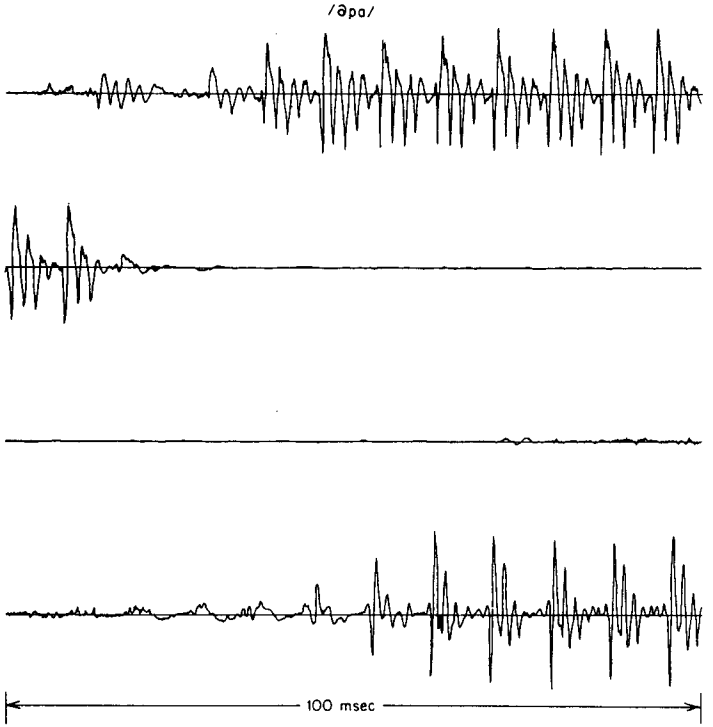


Fig. 3.12 Acoustic waveforms and spectrograms for utterances */UH−P−A/* and */UH−T−A/*.

tract is modeled as a tube of nonuniform, time-varying, cross-section. For frequencies corresponding to wavelengths that are long compared to the dimensions of the vocal tract (less than about 4000 Hz), it is reasonable to assume plane wave propagation along the axis of the tube. A further simplifying assumption is that there are no losses due to viscosity or thermal conduction, either in the bulk of the fluid or at the walls of the tube. With these assumptions, and the laws of conservation of mass, momentum and energy, Portnoff [18] has shown that sound waves in the tube satisfy the following pair of equations:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t} \tag{3.1a}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2} \frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \tag{3.1b}$$

where

$p = p(x,t)$ is the variation in sound pressure in the tube at position x and time t.

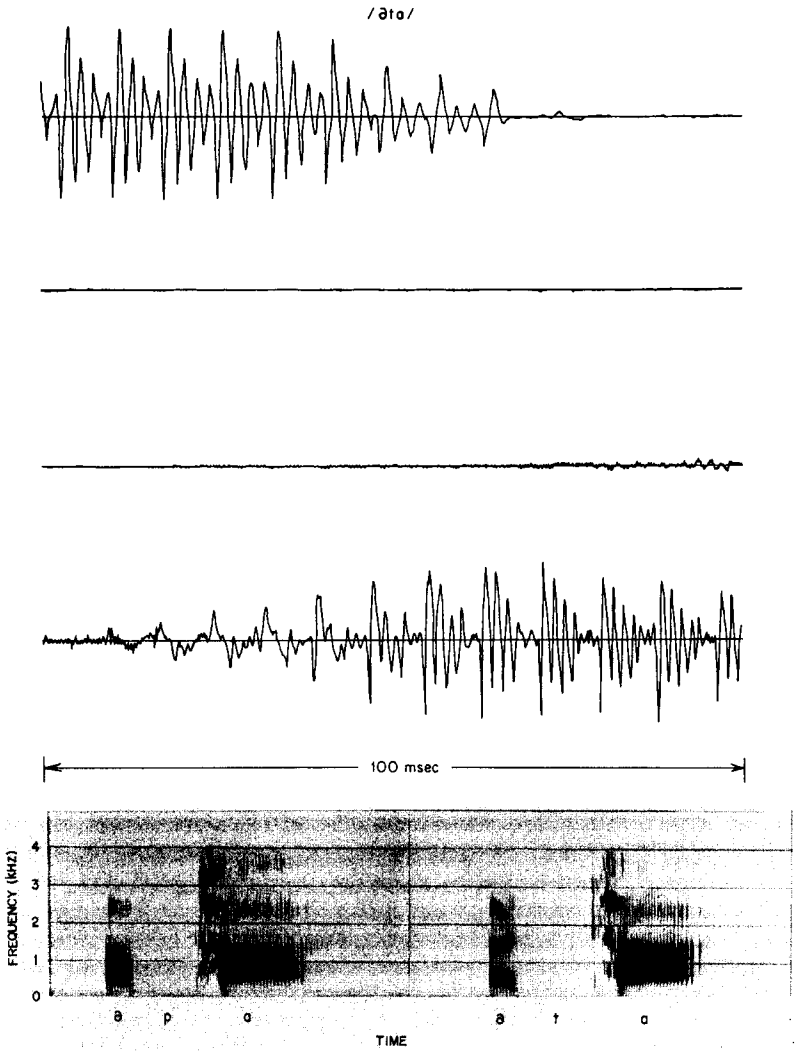


Fig. 3.12 (Continued)

$u = u(x,t)$ is the variation in volume velocity flow at position x and time t.
ρ is the density of air in the tube.
c is the velocity of sound
$A = A(x,t)$ is the "area function" of the tube; i.e., the value of cross-sectional area

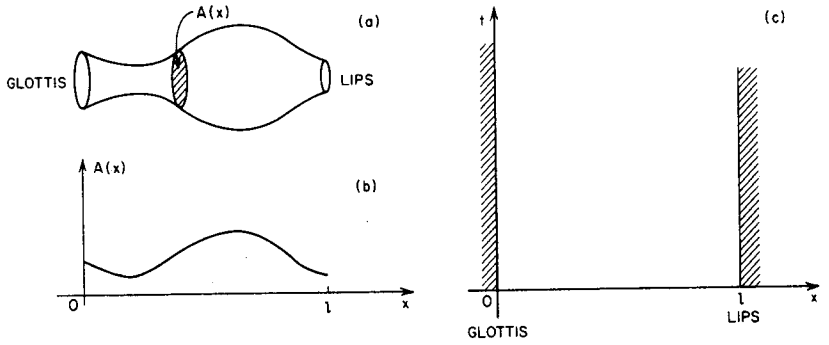


Fig. 3.13 (a) Schematic vocal tract; (b) corresponding area function; (c) $x-t$ plane for solution of wave equation.

normal to the axis of the tube as a function of a distance along the tube and as a function of time.

A similar set of equations has been derived by Sondhi [20].

Closed form solutions to Eqs. (3.1) are not possible except for the simplest configurations. Numerical solutions can be obtained, however. Complete solution of the differential equations requires that pressure and volume velocity be found for values of x and t in the region bounded by the glottis and the lips. To obtain the solution, boundary conditions must be given at each end of the tube. At the lip end, the boundary condition must account for the effects of sound radiation. At the glottis (or possibly some internal point), the boundary condition is imposed by the nature of the excitation.

In addition to the boundary conditions, the vocal tract area function, $A(x,t)$, must be known. Figure 3.13b shows the area function for the tube in Fig. 3.13a, at a particular time. For continuant sounds, it is reasonable to assume that $A(x,t)$ does not change with time; however this is not the case for noncontinuants. Detailed measurements of $A(x,t)$ are extremely difficult to obtain even for continuant sounds. One approach to such measurements is through the use of X-ray motion pictures. Fant [1] and Perkell [21] provide some data of this form; however, such measurements can only be obtained on a limited scale. Another approach is to infer the vocal tract shape from acoustic measurements. Sondhi and Gopinath [22] have described an approach which involves the excitation of the vocal tract by an external source. Both of these approaches are useful for obtaining knowledge of the dynamics of speech production, but they are not directly applicable to the representation of speech signals (e.g. for purposes of transmission). Atal [23] has described investigations directed toward obtaining $A(x,t)$ directly from the speech signal produced under normal speaking conditions.

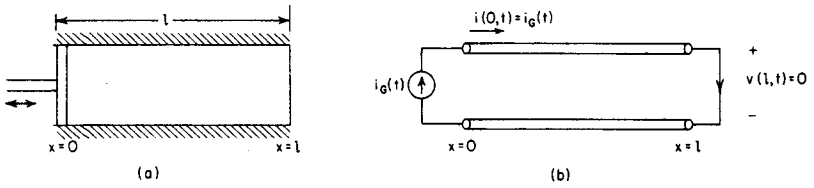


Fig. 3.14 (a) Uniform lossless tube with ideal terminations; (b) corresponding electrical transmission line analogy.

The complete solution of Eqs. (3.1) is very complicated [18] even if $A(x,t)$ is accurately determined. Fortunately, it is not necessary to solve the equations under the most general conditions to obtain insight into the nature of the speech signal. A variety of reasonable approximations and simplifications can be invoked to make the solution possible.

3.2.2 Example: uniform lossless tube

Useful insight into the nature of the speech signal can be obtained by considering a very simple model in which the vocal tract area function is assumed constant in both x and t (time invariant with uniform cross-section). This configuration is approximately correct for the neutral vowel /UH/. We shall examine this model first, returning later to examine more realistic models. Figure 3.14a depicts a tube of uniform cross-section being excited by an ideal source of volume velocity flow. This ideal source is represented by a piston that can be caused to move in any desired fashion, independent of pressure variations in the tube. A further assumption is that at the open end of the tube, there are no variations in air pressure – only in volume velocity. These are obviously gross simplifications which in fact are impossible to achieve in reality; however, we are justified in considering this example since the basic approach of the analysis and the essential features of the resulting solution have much in common with more realistic models. Furthermore we shall show that more general models can be constructed by concatenation of uniform tubes.

If $A(x,t) = A$ is a constant, then the partial differential equations Eqs. (3.1) reduce to the form

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \tag{3.2a}$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{3.2b}$$

It can be shown (see Problem 3.3) that the solution to Eqs. (3.2) has the form

$$u(x,t) = [u^+(t-x/c) - u^-(t+x/c)] \tag{3.3a}$$

$$p(x,t) = \frac{\rho c}{A}[u^+(t-x/c) + u^-(t+x/c)] \tag{3.3b}$$

In Eqs. (3.3) the functions $u^+(t-x/c)$ and $u^-(t+x/c)$ can be interpreted as traveling waves in the positive and negative directions respectively. The relationship between these traveling waves is determined by the boundary conditions.

Anyone familiar with the theory of electrical transmission lines will recall that for a lossless uniform line the voltage $v(x,t)$ and current $i(x,t)$ on the line satisfy the equations

$$-\frac{\partial v}{\partial x} = L\frac{\partial i}{\partial t} \tag{3.4a}$$

$$-\frac{\partial i}{\partial x} = C\frac{\partial v}{\partial t} \tag{3.4b}$$

where L and C are the inductance and capacitance per unit length respectively. Thus the theory of lossless uniform electric transmission lines [24,25] applies directly to the uniform acoustic tube if we make the analogies shown in Table 3.3.

Table 3.3 Analogies Between Acoustic and Electric Quantities

Acoustic Quantity	*Analogous Electric Quantity*
p - pressure	v - voltage
u - volume velocity	i - current
ρ/A - acoustic inductance	L - inductance
$A/(\rho c^2)$ - acoustic capacitance	C - capacitance

Using these analogies, the uniform acoustic tube behaves identically to a lossless uniform transmission line terminated in a short circuit ($v(l,t)=0$) at one end and excited by a current source ($i(0,t)=i_G(t)$) at the other end. This is depicted in Fig. 3.14b.

Frequency domain representations of linear systems such as transmission lines and circuits are exceedingly useful. By analogy we can obtain similar representations of the lossless uniform tube. The frequency-domain representation of this model is obtained by assuming a boundary condition at $x = 0$ of

$$u(0,t) = u_G(t) = U_G(\Omega)e^{j\Omega t} \tag{3.5}$$

That is, the tube is excited by a complex exponential variation of volume velocity of radian frequency Ω and complex amplitude, $U_G(\Omega)$. Since Equations (3.2) are linear, the solution $u^+(t-x/c)$ and $u^-(t+x/c)$ must be of the form

$$u^+(t-x/c) = K^+e^{j\Omega(t-x/c)} \tag{3.6a}$$

$$u^-(t+x/c) = K^-e^{j\Omega(t+x/c)} \tag{3.6b}$$

Substituting these equations into Eqs. (3.3) and applying the boundary condition

$$p(l,t) = 0 \tag{3.7}$$

at the lip end of the tube and Eq. (3.5) at the glottis end we can solve for the unknown constants K^+ and K^-. The resulting sinusoidal steady state solutions for $p(x,t)$ and $u(x,t)$ are

$$p(x,t) = jZ_0 \frac{\sin[\Omega(l-x)/c]}{\cos[\Omega l/c]} U_G(\Omega) e^{j\Omega t} \tag{3.8a}$$

$$u(x,t) = \frac{\cos[\Omega(l-x)/c]}{\cos[\Omega l/c]} U_G(\Omega) e^{j\Omega t} \tag{3.8b}$$

where

$$Z_0 = \frac{\rho c}{A} \tag{3.9}$$

is by analogy called the *characteristic acoustic impedance* of the tube.

An alternative approach which we will use subsequently avoids solution for the forward and backward traveling waves by expressing $p(x,t)$ and $u(x,t)$ for a complex exponential excitation directly as[3]

$$p(x,t) = P(x,\Omega) e^{j\Omega t} \tag{3.10a}$$

$$u(x,t) = U(x,\Omega) e^{j\Omega t} \tag{3.10b}$$

Substituting these solutions into Eqs. (3.1) gives the ordinary differential equations relating the complex amplitudes

$$-\frac{dP}{dx} = ZU \tag{3.11a}$$

$$-\frac{dU}{dx} = YP \tag{3.11b}$$

where

$$Z = j\Omega \frac{\rho}{A} \tag{3.12}$$

can be called the *acoustic impedance* per unit length and

$$Y = j\Omega \frac{A}{\rho c^2} \tag{3.13}$$

is the *acoustic admittance* per unit length. The differential equations of Eqs. (3.11) have solutions of the form

$$P(x,\Omega) = Ae^{\gamma x} + Be^{-\gamma x} \tag{3.14a}$$

$$U(x,\Omega) = Ce^{\gamma x} + De^{-\gamma x} \tag{3.14b}$$

where

$$\gamma = \sqrt{ZY} = j\Omega/c \tag{3.14c}$$

[3]Henceforth our convention will be to denote time domain variables with lower case letters (e.g. $u(x,t)$) and their corresponding frequency domain representations with capital letters (i.e. $U(x,\Omega)$).

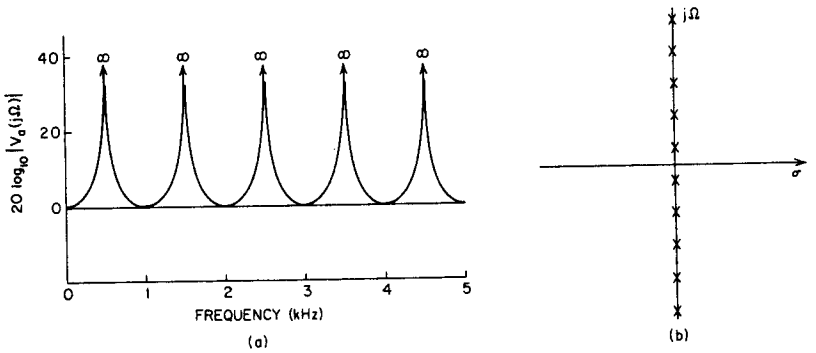


Fig. 3.15 (a) Frequency response; and (b) pole locations for a uniform lossless tube.

The unknown coefficients can be found by applying the boundary conditions

$$P(l,\Omega) = 0 \tag{3.15a}$$

$$U(0,\Omega) = U_G(\Omega) \tag{3.15b}$$

The result is, of course, the same as Eqs. (3.8). Equations (3.8) express the relationship between the sinusoidal volume velocity source and the pressure and volume velocity at any point in the tube. In particular, if we consider the relationship between the volume velocity at the lips and the volume velocity source, we obtain from Eq. (3.8b),

$$u(l,t) = U(l,\Omega) e^{j\Omega t} = \frac{1}{\cos(\Omega l/c)} U_G(\Omega) e^{j\Omega t} \tag{3.16}$$

The ratio

$$\frac{U(l,\Omega)}{U_G(\Omega)} = V_a(j\Omega) = \frac{1}{\cos(\Omega l/c)} \tag{3.17}$$

is the frequency response relating the input and output volume velocities. This function is plotted in Figure 3.15a for values $l = 17.5$ cm and $c = 35000$ cm/sec. Replacing Ω by s/j, we obtain the Laplace transform or system function

$$V_a(s) = \frac{2e^{-sl/c}}{1 + e^{-s2l/c}} \tag{3.18}$$

Note that $V_a(s)$ has an infinite number of poles equally spaced on the $j\Omega$ axis at

$$s_n = \pm j\left[\frac{(2n+1)\pi c}{2l}\right] \qquad n = 0, \pm 1, \pm 2, \ldots \tag{3.19}$$

These pole locations are shown in Fig. 3.15b. The poles of the system function of a linear time-invariant system are the natural frequencies (or eigenfrequencies) of the system. The poles also correspond to resonance frequencies of the system. These resonant frequencies are, of course, called the formant frequencies when considering speech production. As we shall see, similar resonance effects will be observed regardless of the vocal tract shape.

It should be recalled at this point that the frequency response function allows us to determine the response of the system not only to sinusoids but to arbitrary inputs through the use of Fourier analysis. Indeed, Eq. (3.17) has the more general intepretation that $V_a(j\Omega)$ is the ratio of the Fourier transform of the volume velocity at the lips (output) to the Fourier transform of the volume velocity at the glottis (input or source). Thus the frequency response is a convenient characterization of the model for the vocal system. Now that we have demonstrated a method for determining the frequency response of acoustic models for speech production by considering the simplest possible model, we can begin to consider more realistic models.

3.2.3 Effects of losses in the vocal tract

The equations of motion for sound propagation in the vocal tract that we have given were derived under the assumption of no energy loss in the tube. In reality, energy will be lost as a result of viscous friction between the air and the walls of the tube, heat conduction through the walls of the tube, and vibration of the tube walls. To include these effects, we might attempt to return to the basic laws of physics and derive a new set of equations of motion. This is made extremely difficult by the frequency dependence of these losses. As a result, a common approach is to modify the frequency domain representation of the equations of motion [2,18]. We shall survey the results of this approach in this section.

Let us first consider the effects of the vibration of the vocal tract wall. The variations of air pressure inside the tract will cause the walls to experience a varying force. Thus, if the walls are elastic, the cross-sectional area of the tube will change depending upon the pressure in the tube. Assuming that the walls are "locally reacting" [17,18], then the area $A(x,t)$ will be a function of $p(x,t)$. Since the pressure variations are very small, the resulting variation in cross-sectional area can be treated as a small perturbation of the "nominal" area; i.e., we can assume that

$$A(x,t) = A_0(x,t) + \delta A(x,t) \tag{3.20}$$

where $A_0(x,t)$ is the nominal area and $\delta A(x,t)$ is a small perturbation. This is depicted in Fig. 3.16. Because of the mass and elasticity of the vocal tract wall, the relationship between the area perturbation $\delta A(x,t)$, and the pressure variations, $p(x,t)$, can be modeled by a differential equation of the form

$$m_w \frac{d^2(\delta A)}{dt^2} + b_w \frac{d(\delta A)}{dt} + k_w(\delta A) = p(x,t) \tag{3.21}$$

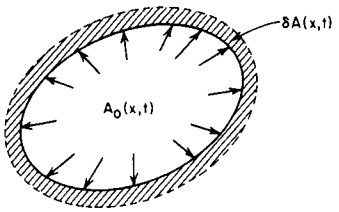


Fig. 3.16 Illustration of the effects of wall vibration.

where

$m_w(x)$ is the mass/unit length of the vocal tract wall
$b_w(x)$ is the damping/unit length of the vocal tract wall
$k_w(x)$ is the stiffness/unit length of the vocal tract wall.

Neglecting second order terms in the quantities u/A and pA, we can write Eqs. (3.1) as

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A_0)}{\partial t} \tag{3.22a}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (pA_0)}{\partial t} + \frac{\partial A_0}{\partial t} + \frac{\partial (\delta A)}{\partial t} \tag{3.22b}$$

Thus, sound propagation in a soft walled tube such as the vocal tract is described by the set of equations, Eqs. (3.20), (3.21) and (3.22).

To examine this effect in more detail let us obtain a frequency domain representation, as before, by considering a time invariant tube, excited by a complex volume velocity source; i.e., the boundary condition at the glottis is

$$u(0,t) = U_G(\Omega)e^{j\Omega t} \tag{3.23}$$

Then because the differential equations Eqs. (3.21) and (3.22) are linear and time invariant for this case, the volume velocity and pressure are also of the form

$$p(x,t) = P(x,\Omega)e^{j\Omega t} \tag{3.24a}$$

$$u(x,t) = U(x,\Omega)e^{j\Omega t} \tag{3.24b}$$

Substituting Eqs. (3.24) into Eqs. (3.21) and (3.22) yields the equations

$$-\frac{\partial P}{\partial x} = ZU \tag{3.25a}$$

$$-\frac{\partial U}{\partial x} = YP + Y_w P \tag{3.25b}$$

where

$$Z(x,\Omega) = j\Omega \frac{\rho}{A_0(x)} \tag{3.26a}$$

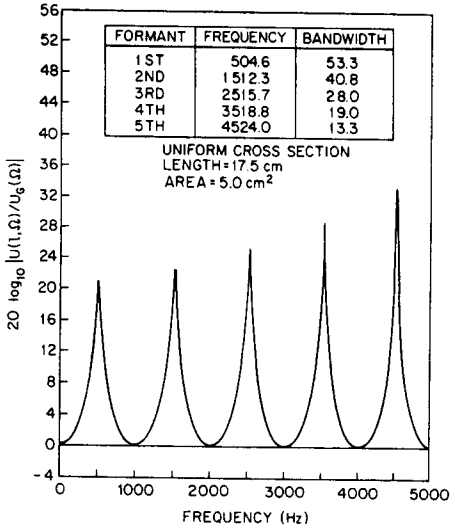


Fig. 3.17 Frequency response of uniform tube with yielding walls and no other losses. Terminated in a short circuit ($p(l,t)$=0). (After Portnoff [18].)

$$Y(x, \Omega) = j\Omega \frac{A_0(x)}{\rho c^2} \tag{3.26b}$$

and

$$Y_w(x, \Omega) = \frac{1}{j\Omega m_w(x) + b_w(x) + \dfrac{k_w(x)}{j\Omega}} \tag{3.26c}$$

Note that Eqs. (3.25) are identical to Eqs. (3.11) except for the addition of the wall admittance term Y_w and for the fact that the acoustic impedance and admittances are in this case functions of x. If we consider a uniform tube, then $A_0(x)$ is constant, and Eqs. (3.12) and (3.13) are identical to Eqs. (3.26a) and (3.26b).

Using estimates obtained from measurements on body tissues [2], the parameters in Eq. (3.26c) were estimated and the differential equations, Eqs. (3.25), were solved with boundary condition $p(l,t) = 0$ at the lip end [18,19]. The ratio

$$V_a(j\Omega) = \frac{U(l, \Omega)}{U_G(\Omega)} \tag{3.27}$$

is plotted as a function of Ω in Fig. 3.17 for the case of a uniform tube of length 17.5 cm [18]. The results are similar to Fig. 3.15 but different in an important way. It is clear that the resonances are no longer exactly on the $j\Omega$ axis of the s-plane. This is evident since the frequency response no longer is

infinite at frequencies 500 Hz, 1500 Hz, 2500 Hz, etc., although the response is peaked in the vicinity of these frequencies. The center frequencies and bandwidths[4] of the resonances in Figure 3.17 are given in the associated table. Several important effects are evident in this example. First we note that the center frequencies are slightly higher than for the lossless case. Second, the bandwidths of the resonances are no longer zero as in the lossless case, since the peak value is no longer infinite. It can be seen that the effect of yielding walls is most pronounced at low frequencies. This is to be expected since we would expect very little motion of the massive walls at high frequencies. The results of this example are typical of the general effects of vocal tract wall vibration; i.e., the center frequencies are slightly increased and the low frequency resonances are broadened as compared to the rigid wall case.

The effects of viscous friction and thermal conduction at the walls are much less pronounced than the effects of wall vibration. Flanagan [2] has considered these losses in detail and has shown that the effect of viscous friction can be accounted for in the frequency domain representation (Eq. (3.25)) by including a real, frequency dependent term in the expression for the acoustic impedance, Z; i.e.,

$$Z(x, \Omega) = \frac{S(x)}{[A_0(x)]^2} \sqrt{\Omega \rho \mu / 2} + j\Omega \frac{\rho}{A_0(x)} \tag{3.28a}$$

where $S(x)$ is the circumference of the tube, μ is the coefficient of friction, and ρ is the density of air in the tube. The effects of heat conduction through the vocal tract wall can likewise be accounted for by adding a real frequency dependent term to the acoustic admittance, $Y(x, \Omega)$; i.e.,

$$Y(x, \Omega) = \frac{S(x)(\eta - 1)}{\rho c^2} \sqrt{\frac{\lambda \Omega}{2 c_p \rho}} + j\Omega \frac{A_0(x)}{\rho c^2} \tag{3.28b}$$

where c_p is the specific heat at constant pressure, η is the ratio of specific heat at constant pressure to that at constant volume, and λ is the coefficient of heat conduction [2]. Typical values for the constants in Eqs. (3.28) are given by Flanagan [2]. For our purposes, it is sufficient to note that the loss due to friction is proportional to the real part of $Z(x, \Omega)$, and thus to $\Omega^{1/2}$. Likewise the thermal loss is proportional to the real part of $Y(x, \Omega)$, which in turn is proportional to $\Omega^{1/2}$. Using the values given by Eqs. (3.28) for $Z(x, \Omega)$ and $Y(x, \Omega)$ and the values of $Y_w(x, \Omega)$ given by Eq. (3.26c), Eqs. (3.25) were again solved numerically [18]. The resulting frequency response for the boundary condition of $p(l,t) = 0$ is shown in Fig. 3.18. Again the center frequencies and bandwidths were determined and are shown in the associated table. Comparing Fig. 3.18 with Fig. 3.17, we observe that the center frequencies are decreased by the addition of friction and thermal loss, while the bandwidths are increased. Since friction and thermal losses increase with $\Omega^{1/2}$, the higher frequency resonances experience a greater broadening than do the lower resonances.

[4]The bandwidth of a resonance is defined as the frequency interval around a resonance in which the frequency response is greater than 0.707 times the peak value at the center frequency [26].

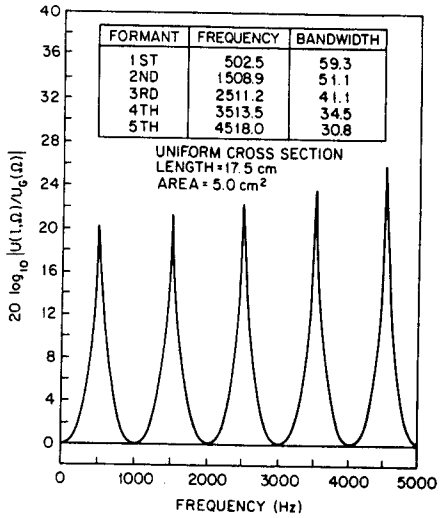


Fig. 3.18 Frequency response of uniform tube with yielding walls, friction and thermal losses, and terminated in a short circuit ($p(l,t)=0$). (After Portnoff [18].)

The examples depicted in Figs. 3.17 and 3.18 are typical of the general effects of losses in the vocal tract. To summarize, viscous and thermal losses increase with frequency and have their greatest effect in the high frequency resonances, while wall loss is most pronounced at low frequencies. The yielding walls tend to raise the resonant frequencies while the viscous and thermal losses tend to lower them. The net effect for the lower resonances is a slight upward shift as compared to the lossless, rigid walled model. The effect of friction and thermal loss is small compared to the effects of wall vibration for frequencies below 3-4 kHz. Thus, Eqs. (3.21) and (3.22), which neglect these losses, are nevertheless a good representation of sound transmission in the vocal tract. As we shall see in the next section, the radiation termination at the lips is a much greater source of high frequency loss. This provides further justification for neglecting friction and thermal loss in models or simulations of speech production.

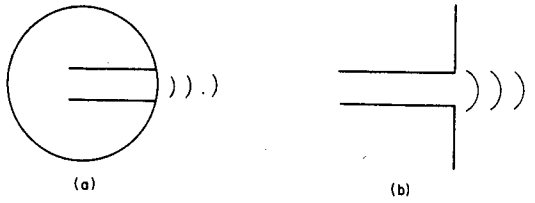


Fig. 3.19 (a) Radiation from a spherical baffle; (b) radiation from an infinite plane baffle.

3.2.4 Effects of radiation at the lips

So far we have discussed the way that internal losses affect the sound transmission properties of the vocal tract. In our examples we have assumed the boundary condition $p(l,t) = 0$ at the lips. In the electric transmission line analogy this corresponds to a short circuit. The acoustic counterpart of a short circuit is as difficult to achieve as an electrical short circuit since it requires a configuration in which volume velocity changes can occur at the end of the vocal tract tube without corresponding pressure changes. In reality, the vocal tract tube terminates with the opening between the lips (or the nostrils in the case of nasals). Thus a reasonable model is as depicted in Fig. 3.19a, which shows the lip opening as an orifice in a sphere. In this model, at low frequencies, the opening can be considered a radiating surface, with the radiated sound waves being diffracted by the spherical baffle that represents the head.

The resulting diffraction effects are complicated and difficult to represent; however, for determining the boundary condition at the lips, all that is needed is a relationship between pressure and volume velocity at the radiating surface. Even this is very complicated for the configuration of Fig. 3.19a. However, if the radiating surface (lip opening) is small compared to the size of the sphere, a reasonable approximation assumes that the radiating surface is set in a plane baffle of infinite extent as depicted in Fig. 3.19b. In this case, it can be shown [2,17,18] that the sinusoidal steady state relation between the complex amplitudes of pressure and volume velocity at the lips is

$$P(l,\Omega) = Z_L(\Omega)\cdot U(l,\Omega) \tag{3.29a}$$

where the "radiation impedance" or "radiation load" at the lips is approximately of the form

$$Z_L(\Omega) = \frac{j\Omega L_r R_r}{R_r + j\Omega L_r} \tag{3.29b}$$

The electrical analog to this radiation load is a parallel connection of a radiation resistance, R_r, and radiation inductance, L_r. Values of R_r and L_r that provide a good approximation to the infinite plane baffle are [2]

$$R_r = \frac{128}{9\pi^2} \tag{3.30a}$$

$$L_r = \frac{8a}{3\pi c} \tag{3.30b}$$

where a is the radius of the opening and c is the velocity of sound.

The behavior of the radiation load influences the nature of wave propagation in the vocal tract through the boundary condition of Eqs. (3.29). Note that it is easily seen from Eq. (3.29b) that at very low frequencies $Z_L(\Omega) \approx 0$; i.e., at very low frequencies the radiation impedance approximates the ideal short circuit termination that has been assumed up to this point. Likewise, it is clear from Eq. (3.29b) that for a mid range of frequencies, (when

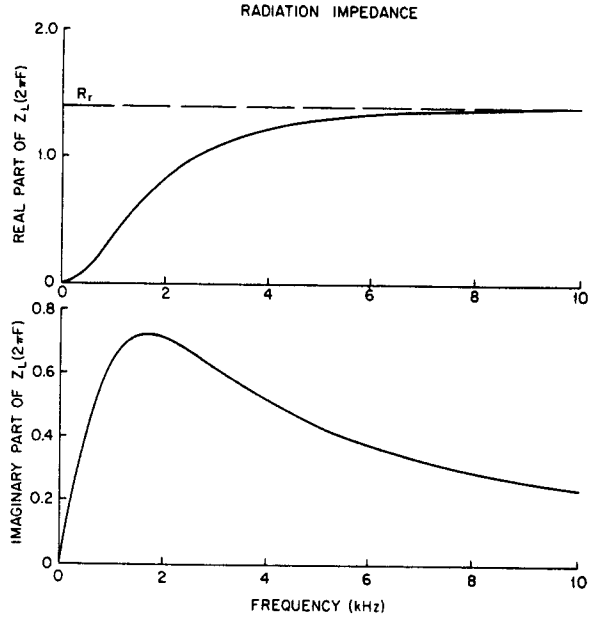


Fig. 3.20 Real and imaginary parts of the radiation impedance.

$\Omega L_r << R_r$), $Z_L(\Omega) \approx j\Omega L_r$. At higher frequencies ($\Omega L_r >> R_r$), $Z_L(\Omega) \approx R_r$. This is readily seen in Fig. 3.20 which shows the real and imaginary parts of $Z_L(\Omega)$ as a function of Ω for typical values of the parameters. The energy dissipated due to radiation is proportional to the real part of the radiation impedance. Thus we can see that for the complete speech production system (vocal tract and radiation), the radiation losses will be most significant at higher frequencies. To assess the magnitude of this effect, Eqs. (3.25), (3.26c) and (3.29) were solved simultaneously for the case of a uniform time invariant tube with yielding walls, friction and thermal losses, and radiation loss corresponding to an infinite plane baffle. Figure 3.21 shows the resulting frequency response.

$$V_a(j\Omega) = \frac{U(l,\Omega)}{U_G(\Omega)} \tag{3.31}$$

for an input $U(0,t) = U_G(\Omega)e^{j\Omega t}$. Comparing Figure 3.21 to Figures 3.17 and 3.18 shows that the major effect is to broaden the resonances (increase loss) and to lower the resonance frequencies (formant frequencies). As expected the major effect on the resonance bandwidths occurs at higher frequencies. The first resonance (formant) bandwidth is primarily determined by the wall loss, while the higher formant bandwidths are primarily determined by radiation loss. The second and third formant bandwidths can be said to be determined by a combination of these two loss mechanisms.

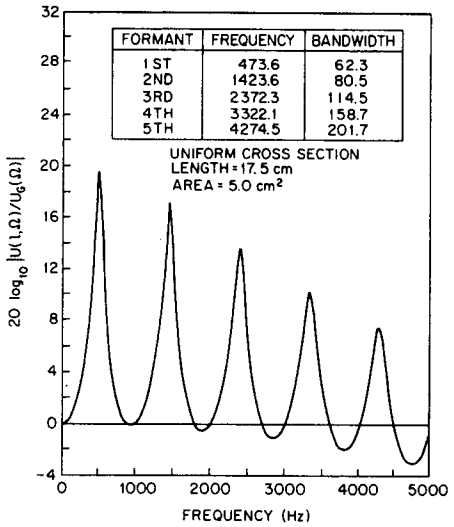


Fig. 3.21 Frequency response of uniform tube with yielding walls, friction and thermal loss. (After Portnoff [18].)

The frequency response shown in Figure 3.21 relates the volume velocity at the lips to the input volume velocity at the lips. The relationship between pressure at lips and volume velocity at the glottis may be of interest, especially if a pressure sensitive microphone is used in converting the acoustic wave to an electrical wave. Since $P(l,\Omega)$ and $U(l,\Omega)$ are related by Eq. (3.29a), the pressure transfer function is simply

$$H_a(\Omega) = \frac{P(l,\Omega)}{U_G(\Omega)} = \frac{P(l,\Omega)}{U(l,\Omega)} \cdot \frac{U(l,\Omega)}{U_G(\Omega)}$$

$$= Z_L(\Omega) \cdot V_a(\Omega) \tag{3.32}$$

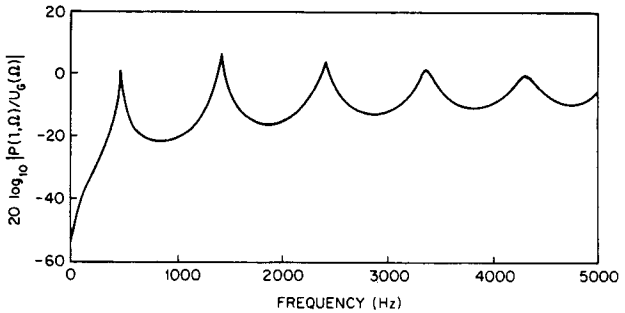


Fig. 3.22 Frequency response relating pressure at lips to volume velocity at glottis for uniform tube.

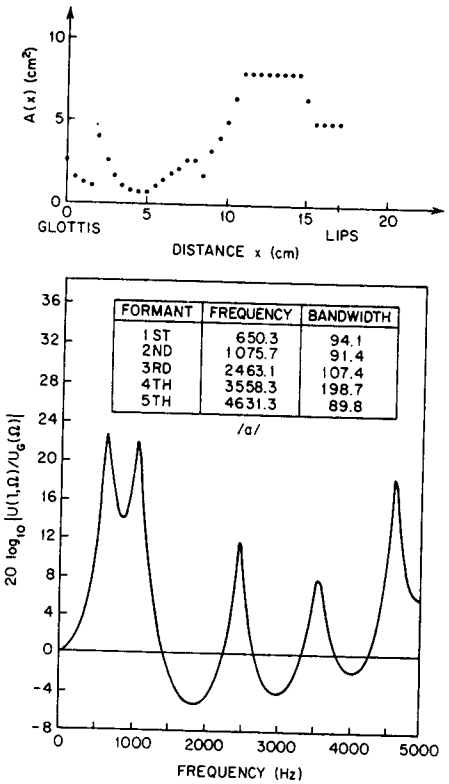


Fig. 3.23 Area function (after Fant [1]) and frequency response (after Portnoff [18]) for the Russian vowel /*a*/.

It can be seen from Fig. 3.21 that the major effects will be an emphasis of high frequencies and the introduction of a zero at $\Omega = 0$. Figure 3.22 shows the frequency response $20 \log_{10} |H_a(\Omega)|$ including wall losses and the radiation loss of an infinite plane baffle. A comparison of Figures 3.21 and 3.22 places in evidence the zero at $\Omega = 0$ and the high frequency emphasis.

3.2.5 Vocal tract transfer functions for vowels

The equations discussed in Sections 3.2.3 and 3.2.4 constitute a detailed model for sound propagation and radiation in speech production. Using numerical integration techniques, either the time domain or frequency domain forms can be solved for a variety of vocal tract response functions. Such solutions provide considerable insight into the nature of the speech production process and the speech signal.

As an example [18], the frequency domain equations, Eqs. (3.25), (3.26c), (3.28), and (3.29), were used to compute frequency response functions for a set of area functions measured by Fant [1]. Figures 3.23-3.26 show the vocal tract area functions and corresponding frequency responses $(U(l, \Omega)/U_G(\Omega))$ for the Russian vowels /a/, /e/, /i/, and /u/. These figures illustrate the effects of all the loss mechanisms discussed in Sections 3.2.3 and 3.2.4. The formant frequencies and bandwidths compare favorably with measurements on natural vowels of formant frequencies obtained by Peterson and Barney [11] and formant bandwidths by Dunn [27].

In summary, we may conclude from these examples and those of the previous sections that:

1. The vocal system is characterized by a set of resonances (formants) that depend primarily upon the vocal tract area function, although there is some shift due to losses, as compared to the lossless case.

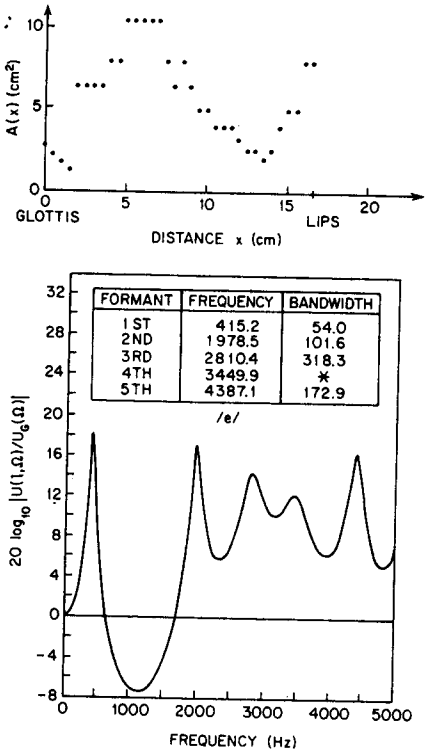


Fig. 3.24 Area function (after Fant [1]) and frequency response (after Portnoff [18]) for the Russian vowel /*e*/.

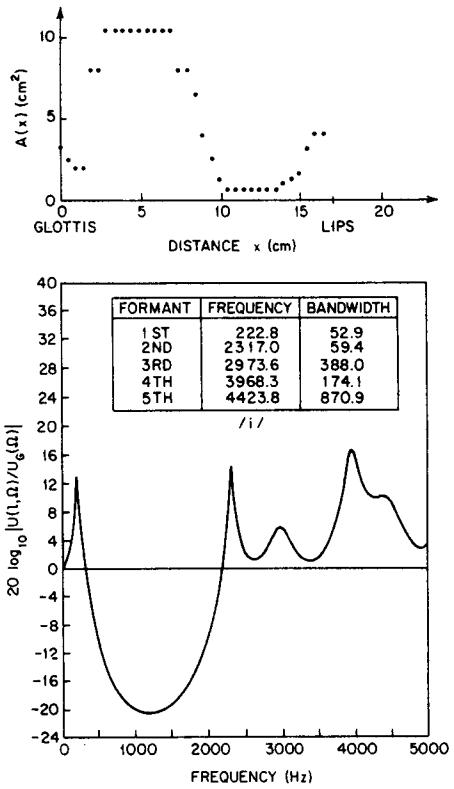


Fig. 3.25 Area function (after Fant [1]) and frequency response (after Portnoff [18]) for the Russian vowel /*i*/.

2. The bandwidths of the lowest formant frequencies (first and second) depend primarily upon the vocal tract wall loss.[5]
3. The bandwidths of the higher formant frequencies depend primarily upon the viscous friction and thermal losses in the vocal tract and the radiation loss.

3.2.6 The effect of nasal coupling

In the production of the nasal consonants /m/, /n/, and /η/ the velum is lowered like a trap-door to couple the nasal tract to the pharynx. Simultaneously a complete closure is formed in the oral tract (e.g., at the lips for /m/).

[5]We shall see in Section 3.2.7 that loss associated with the excitation source also effects the lower formants.

This configuration can be represented as in Fig. 3.27a, which shows two branches, one of which is completely closed. At the point of branching the sound pressure is the same at the input to each tube, while the volume velocity must be continuous at the branching point; i.e., the volume velocity at the output of the pharynx tube must be the sum of the volume velocities at the inputs to the nasal and oral cavities. The corresponding electrical transmission line analog is shown in Fig. 3.27b. Note that continuity of volume velocity at the junction of the 3 tubes corresponds to Kirchoff's current law at the junction of the transmission lines.

For nasal consonants the radiation of sound occurs primarily at the nostrils. Thus the nasal tube is terminated with a radiation impedance appropriate for the size of the nostril openings. The oral tract, which is completely closed, is terminated by the equivalent of an open electrical circuit; i.e., no flow occurs. Nasalized vowels are produced by the same system with the oral tract ter-

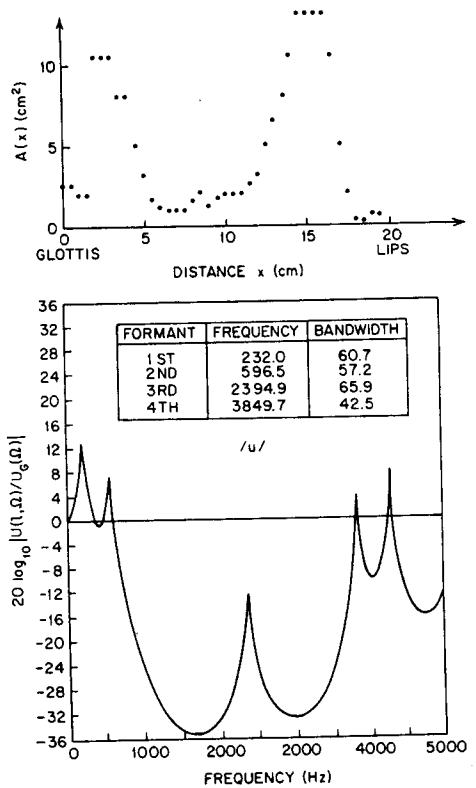


Fig. 3.26 Area function (after Fant [1]) and frequency response (after Portnoff [18]) for the Russian vowel /*u*/.

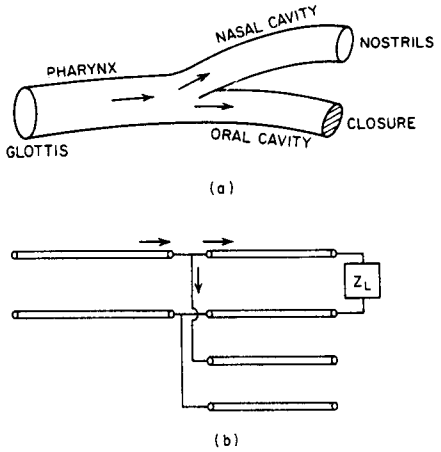


Fig. 3.27 (a) Tube model for production of nasals; (b) corresponding electrical analog.

minated as for vowels. The speech signal would then be the superposition of the nasal and oral outputs.

The mathematical model for this configuration consists of three sets of partial differential equations with boundary conditions being imposed by the form of glottal excitation, terminations of the nasal and oral tracts, and continuity relations at the junction. This leads to a rather complicated set of equations which could in principle be solved, given adequate measurements of area functions for all three tubes. However, the transfer function of the complete system would have many features in common with the previous examples. That is, the system would be characterized by a set of resonances or formants that would be dependent upon the shape and length of the 3 tubes. An important difference results from the fact that the closed oral cavity can trap energy at certain frequencies, preventing those frequencies from appearing in the nasal output. In the electrical transmission line analogy, these are frequencies at which the input impedance of the open circuited line is zero. At these frequencies the junction is short circuited by the transmission line corresponding to the oral cavity. The result is that for nasal sounds, the vocal system transfer function will be characterized by anti-resonances (zeros) as well as resonances. It has also been observed [13] that nasal formants have broader bandwidths than non-nasal voiced sounds. This is attributed to the greater viscous friction and thermal loss due to the large surface area of the nasal cavity.

3.2.7 Excitation of sound in the vocal tract

The previous sub-sections have described how the laws of physics can be applied to describe the propagation and radiation of sound in speech production. To complete our discussion of acoustic principles we must now consider the

mechanisms whereby sound waves are generated in the vocal system. Recall that in our general overview of speech production in Section 3.1.1, we identified 3 major mechanisms of excitation. These are:

1. Air flow from the lungs is modulated by the vocal cord vibration, resulting in a quasi-periodic pulse-like excitation.
2. Air flow from the lungs becomes turbulent as the air passes through a constriction in the vocal tract, resulting in a noise-like excitation.
3. Air flow builds up pressure behind a point of total closure in the vocal tract. The rapid release of this pressure, by removing the constriction, causes a transient excitation.

A detailed model of excitation of sound in the vocal system involves the subglottal system (lungs, bronchi, and trachea), the glottis, and the vocal tract. Indeed, a model which is complete in all necessary details is also fully capable of simulating breathing as well as speech production! [2]. The first comprehensive effort toward a detailed physical model of sound generation in the vocal system was by Flanagan [2,28]. Subsequent research has produced a much refined model that provides a very detailed representation of the process of generation of both voiced and unvoiced speech [28-31]. This model, which is based upon classical mechanics and fluid mechanics, is beyond the scope of our discussion here. However, a brief qualitative discussion of the basic principles of sound generation will be helpful in pointing the way toward the simple models that are widely used as the basis for speech processing.

The vibration of the vocal cords in voiced speech production can be explained by considering the schematic representation of the vocal system shown in Fig. 3.28. The vocal cords constrict the path from the lungs to the vocal tract. As lung pressure is increased, air flows out of the lungs and through the opening between the vocal cords (glottis). Bernoulli's law states that when a fluid flows through an orifice, the pressure is lower in the constriction than on either side. If the tension in the vocal cords is properly adjusted, the reduced pressure allows the cords to come together, thereby completely constricting air flow. (This is indicated by the dotted lines in Figure 3.28.) As a result, pressure increases behind the vocal cords. Eventually it builds up to a level sufficient to force the vocal cords to open and thus allow air to flow through the glottis again. Again the air pressure in the glottis falls, and the

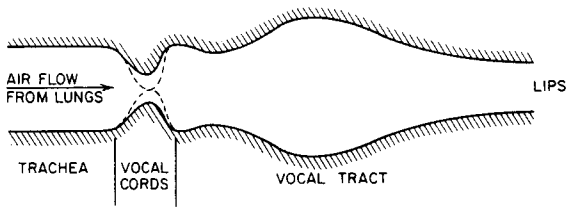


Fig. 3.28 Schematic representation of the vocal system.

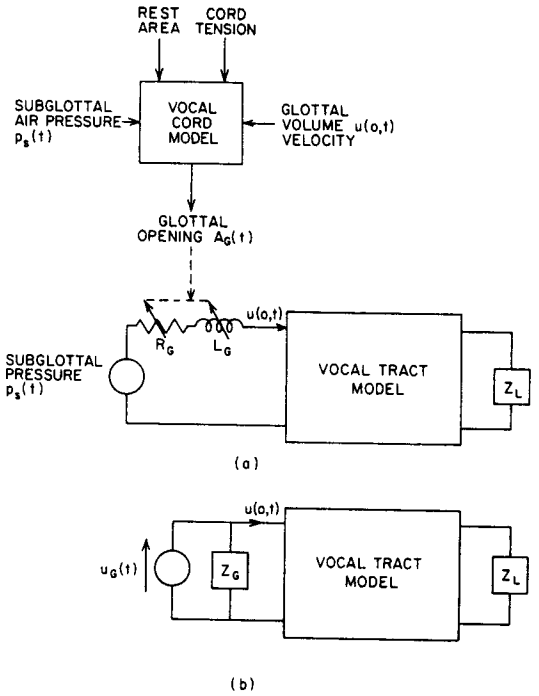


Fig. 3.29 (a) Diagram of vocal cord model; (b) approximate model for vocal cords.

cycle is repeated. Thus, the vocal cords enter a condition of sustained oscillation. The rate at which the glottis opens and closes is controlled by the air pressure in the lungs, the tension and stiffness of the vocal cords, and the area of the glottal opening under rest conditions. These are the control parameters of a detailed model of vocal cord behavior. Such models must also include the effects of the vocal tract, since pressure variations in the vocal tract influence the pressure variations in the glottis. In terms of the electrical analog, the vocal tract acts as a load on the vocal cord oscillator. A schematic diagram of the vocal cord model (adapted from [30]) is shown in Figure 3.29a. The vocal cord model consists of a set of complicated nonlinear differential equations. The coupling of these differential equations to the partial differential equations describing vocal tract transmission can be represented by a time varying acoustic resistance and inductance as shown [30]. These impedance elements are functions of $1/A_G(t)$. For example, when $A_G(t) = 0$ (glottis closed) the impedance is infinite and the volume velocity is zero. Thus, the glottal flow is automatically chopped up into pulses. An example of the signals generated by such a model is shown in Fig. 3.30 [30]. The upper waveform is the volume velocity and the lower waveform is the pressure at the lips for a vocal tract

configuration appropriate for the vowel /a/. The pulse-like nature of the glottal flow is certainly consistent with our previous discussion and with direct observation through the use of high-speed motion pictures [2]. The damped oscillations of the output are, of course, consistent with our previous discussion of the nature of sound propagation in the vocal tract.

Since glottal area is a function of the flow into the vocal tract, the overall system of Fig. 3.29a is nonlinear, even though the vocal tract transmission and radiation systems are linear. The coupling between the vocal tract and the glottis is weak, however, and it is common to neglect this interaction. This leads to a separation and linearization of the excitation and transmission system as depicted in Figure 3.29b. In this case $u_G(t)$ is a volume velocity source whose wave shape is of the form of the upper waveform in Fig. 3.30. The glottal acoustic impedance, Z_G, is obtained by linearization of the relations between pressure and volume velocity in the glottis [2]. This impedance is of the form

$$Z_G(\Omega) = R_G + j\Omega L_G \tag{3.33}$$

where R_G and L_G are constants. With this configuration the ideal frequency domain boundary condition of $U(0, \Omega) = U_G(\Omega)$ is replaced by

$$U(0, \Omega) = U_G(\Omega) - P(0, \Omega)/Z_G(\Omega) \tag{3.34}$$

The glottal source impedance has significant effects upon resonance bandwidths for the speech production system. The major effect is a broadening of the lowest resonance. This is because $Z_G(\Omega)$ increases with frequency so that at high frequencies Z_G appears as an open circuit and all of the glottal source flows into the vocal tract system. Thus, yielding walls and glottal loss control the bandwidths of the lower formants while radiation, friction, and thermal losses control the bandwidths of the higher formants.

The mechanism of production of voiceless sounds involves the turbulent flow of air. This can occur at a constriction whenever the volume velocity exceeds a certain critical value [2,29]. Such excitation can be modeled by

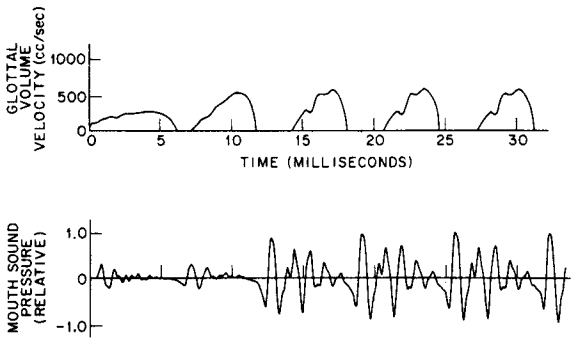


Fig. 3.30 Glottal volume velocity and sound pressure at the mouth for vowel /a/. (After Ishizaka and Flanagan [30].)

inserting a randomly time varying source at the point of constriction. The strength of the source is made dependent (nonlinearly) upon the volume velocity in the tube. In this way, frication is automatically inserted when needed [2,29,31]. For fricative sounds, the vocal cord parameters are adjusted so that the cords do not vibrate. For voiced fricatives, the vocal cords vibrate and turbulent flow occurs at a constriction whenever the volume velocity exceeds the critical value. This usually occurs at the peaks of the volume velocity pulses. For plosives, the vocal tract is closed for a period of time while pressure is built up behind the closure with the vocal cords not vibrating. When the constriction is released, the air rushes out at a high velocity thus causing turbulent flow.

3.2.8 Models based upon the acoustic theory

Section 3.2 has discussed in some detail the important features of the acoustic theory of speech production. The detailed models for sound generation, propagation, and radiation can in principle be solved with suitable values of the excitation and vocal tract parameters to compute an output speech waveform. Indeed, it can be argued effectively that this may be the best

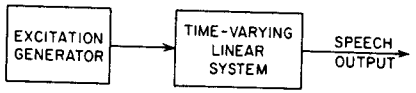


Fig. 3.31 Source-system model of speech production.

approach to the synthesis of natural sounding synthetic speech [31]. However, for many purposes such detail is impractical or unnecessary. In such cases the acoustic theory points the way to a simplified approach to modeling speech signals. Figure 3.31 shows a general block diagram that is representative of numerous models that have been used as the basis for speech processing. These models all have in common that the excitation features are separated from the vocal tract and radiation features. The vocal tract and radiation effects are accounted for by the time-varying linear system. Its purpose is to model the resonance effects that we have discussed. The excitation generator creates a signal that is either a train of (glottal) pulses, or randomly varying (noise). The parameters of the source and system are chosen so that the resulting output has the desired speech-like properties. If this can be done, the model may serve as a useful basis for speech processing. In the remainder of this chapter we shall discuss some models of this type.

3.3 Lossless Tube Models

A widely used model for speech production is based upon the assumption that the vocal tract can be represented as a concatenation of lossless acoustic tubes, as depicted in Fig. 3.32. The constant cross-sectional areas $\{A_k\}$, of the tubes

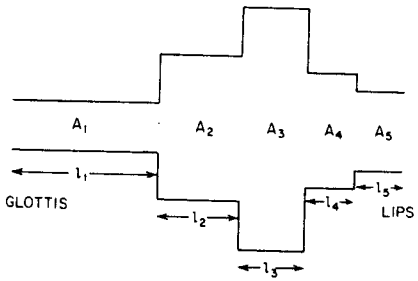


Fig. 3.32 Concatenation of 5 lossless acoustic tubes.

are chosen so as to approximate the area function, $A(x)$, of the vocal tract. If a large number of tubes of short length is used, we can reasonably expect the resonant frequencies of the concatenated tubes to be close to those of a tube with continuously varying area function. However, since this approximation neglects the losses due to friction, heat conduction, and wall vibration, we may also reasonably expect the bandwidths of the resonances to differ from those of a detailed model which includes these losses. However, losses can be accounted for at the glottis and lips, and as we shall see here and in Chapter 8, this can be done so as to accurately represent the resonance properties of the speech signal.

More important for our present discussion is the fact that lossless tube models provide a convenient transition between continuous-time models and discrete-time models. Thus we shall consider models of the form of Figure 3.32 in considerable detail.

3.3.1 Wave propagation in concatenated lossless tubes

Since each tube in Figure 3.32 is assumed lossless, sound propagation in each tube is described by Equations (3.2) with appropriate values of the cross-sectional area. Thus if we consider the k^{th} tube with cross-sectional area, A_k, the pressure and volume velocity in that tube have the form

$$p_k(x,t) = \frac{\rho c}{A_k}\left[u_k^+(t-x/c) + u_k^-(t+x/c)\right] \tag{3.35a}$$

$$u_k(x,t) = u_k^+(t-x/c) - u_k^-(t+x/c) \tag{3.35b}$$

where x is distance measured from the left-hand end of the k^{th} tube ($0 \leqslant x \leqslant l_k$) and $u_k^+(\,)$ and $u_k^-(\,)$ are positive-going and negative-going traveling waves in the k^{th} tube. The relationship between the traveling waves in adjacent tubes can be obtained by applying the physical principle that pressure and volume velocity must be continuous in both time and space everywhere in the system. This provides boundary conditions that can be applied at both ends of each tube.

Consider in particular the junction between the k^{th} and $(k+1)^{st}$ tubes as

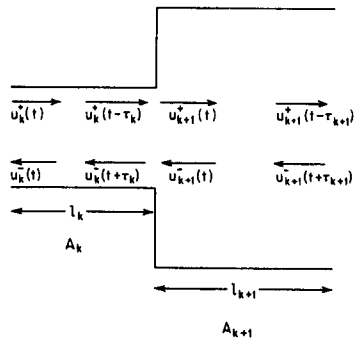


Fig. 3.33 Illustration of the junction between two lossless tubes.

depicted in Figure 3.33. Applying the continuity conditions at the junction gives

$$p_k(l_k,t) = p_{k+1}(0,t) \tag{3.36a}$$

$$u_k(l_k,t) = u_{k+1}(0,t) \tag{3.36b}$$

Substituting Eqs. (3.35) into Eqs. (3.36) gives

$$\frac{A_{k+1}}{A_k}\left[u_k^+(t-\tau_k) + u_k^-(t+\tau_k)\right] = u_{k+1}^+(t) + u_{k+1}^-(t) \tag{3.37a}$$

$$u_k^+(t-\tau_k) - u_k^-(t+\tau_k) = u_{k+1}^+ - u_{k+1}^-(t) \tag{3.37b}$$

where $\tau_k = l_k/c$ is the time for a wave to travel the length of the k^{th} tube. From Figure 3.33 we observe that part of the positive going wave that reaches the junction is propagated on to the right while part is reflected back to the left. Likewise part of the backward traveling wave is propagated on to the left while part is reflected back to the right. Thus, if we solve for $u_{k+1}^+(t)$ and $u_k^-(t+\tau_k)$ in terms of $u_{k+1}^-(t)$ and $u_k^+(t-\tau_k)$ we will be able to see how the forward and reverse traveling waves propagate in the overall system. Solving Eq. (3.37b) for $u_k^-(t+\tau_k)$ and substituting the result into Eq. (3.37a) yields

$$u_{k+1}^+(t) = \left(\frac{2A_{k+1}}{A_{k+1}+A_k}\right) u_k^+(t-\tau_k) + \left(\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right) u_{k+1}^-(t) \tag{3.38a}$$

Subtracting Eq. (3.37b) from Eq. (3.37a) gives

$$u_k^-(t+\tau_k) = -\left(\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right) u_k^+(t-\tau_k) + \left(\frac{2A_k}{A_{k+1}+A_k}\right) u_{k+1}^-(t) \tag{3.38b}$$

It can be seen from Eq. (3.38a) that the quantity

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{3.39}$$

is the amount of $u_{k+1}^-(t)$ that is reflected at the junction. Thus, the quantity r_k is called the reflection coefficient for the k^{th} junction. It is easily shown that since the areas are all positive (see Problem 3.4),

$$-1 \leqslant r_k \leqslant 1 \tag{3.40}$$

Using this definition of r_k, Eqs. (3.38) can be expressed as

$$u_{k+1}^+(t) = (1+r_k)u_k^+(t-\tau_k) + r_k u_{k+1}^-(t) \tag{3.41a}$$

$$u_k^-(t+\tau_k) = -r_k u_k^+(t-\tau_k) + (1-r_k)u_{k+1}^-(t)\ . \tag{3.41b}$$

Equations of this form were first used for speech synthesis by Kelly and Lochbaum [32]. It is useful to depict these equations graphically as in Figure 3.34. In this figure, signal flow-graph conventions[6] are used to represent the multiplications and additions of Eqs. (3.41). Clearly, each junction of a system such as that depicted in Fig. 3.32 can be represented by a system such as Fig. 3.34, as long as our interest is only in values of pressure and volume velocity at the input and output of the tubes. This is not restrictive since we are primarily interested only in the relationship between the output of the last tube and the input of the first tube. Thus, a 5 tube model such as Fig. 3.32, would have 5 sets of forward and backward delays and 4 junctions, each characterized by a reflection coefficient. To complete the representation of wave propagation in such a system we must consider boundary conditions at the "lips" and "glottis" of the system.

3.3.2 Boundary conditions

Let us assume that there are N sections indexed from 1 to N starting at the glottis. Then the boundary condition at the lips will relate pressure, $p_N(l_N,t)$, and volume velocity, $u_N(l_N,t)$, at the output of the N^{th} tube to the

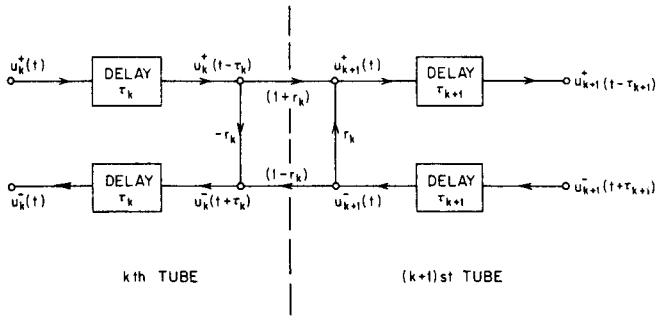


Fig. 3.34 Signal-flow representation of the junction between two lossless tubes.

[6]See Ref. [33] for an introduction to the use of signal flow graphs in signal processing.

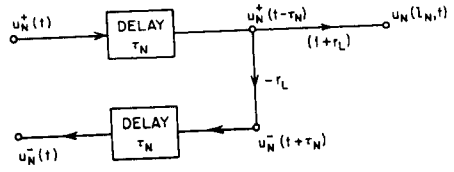


Fig. 3.35 Termination at lip end of a concatenation of lossless tubes.

radiated pressure and volume velocity. If we use the frequency-domain relations of Section 3.2.4 we obtain a relation of the form

$$P_N(l_N, \Omega) = Z_L \cdot U_N(l_N, \Omega) \tag{3.42}$$

If we assume for the moment that Z_L is real, then we obtain the time domain relation

$$\frac{\rho c}{A_N}\left[u_N^+(t-\tau_N) + u_N^-(t+\tau_N)\right] = Z_L\left[u_N^+(t-\tau_N) - u_N^-(t+\tau_N)\right] \tag{3.43}$$

(If Z_L is complex Eq. (3.43) would be replaced by a differential equation relating $p_N(l_N,t)$ and $u_N(l_N,t)$.) Solving for $u_N^-(t+\tau_N)$ we obtain

$$u_N^-(t+\tau_N) = -r_L u_N^+(t-\tau_N) \tag{3.44}$$

where the reflection coefficient at the lips is

$$r_L = \left(\frac{\rho c/A_N - Z_L}{\rho c/A_N + Z_L}\right) \tag{3.45}$$

The output volume velocity at the lips is

$$\begin{aligned} u_N(l_N,t) &= u_N^+(t-\tau_N) - u_N^-(t+\tau_N) \\ &= (1+r_L)u_N^+(t-\tau_N) \end{aligned} \tag{3.46}$$

The effect of this termination as represented by Eqs. (3.44) and (3.46) is depicted in Fig. 3.35. Note that if Z_L is complex, it can be shown that Eq. (3.45) remains valid, but, of course, r_L will then be complex also, and it would be necessary to replace Eq. (3.44) by its frequency-domain equivalent. Alternatively, $u_N^-(t+\tau_N)$ and $u_N^+(t-\tau_N)$ could be related by a differential equation. (See Problem 3.5.)

The frequency domain relations, assuming that the excitation source is linearly separable from the vocal tract, are given in Section 3.2.7. Applying this assumption to the pressure and volume velocity at the input to the first tube we get

$$U_1(0, \Omega) = U_G(\Omega) - P_1(0, \Omega)/Z_G \tag{3.47}$$

Assuming again that Z_G is real,

$$u_1^+(t) - u_1^-(t) = u_G(t) - \frac{\rho c}{A_1}\left(\frac{u_1^+(t) + u_1^-(t)}{Z_G}\right) \tag{3.48}$$

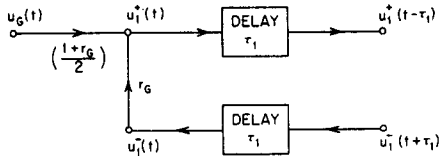


Fig. 3.36 Termination at glottal end of a concatenation of lossless tubes.

Solving for $u_1^+(t)$ we obtain (see Problem 3.6)

$$u_1^+(t) = \frac{(1+r_G)}{2}u_G(t) + r_G u_1^-(t) \tag{3.49}$$

where the glottal reflection coefficient is

$$r_G = \left(\frac{Z_G - \frac{\rho c}{A_1}}{Z_G + \frac{\rho c}{A_1}}\right) \tag{3.50}$$

Equation 3.49 can be depicted as in Fig. 3.36. As in the case of the radiation termination, if Z_G is complex, then Eq. (3.50) still holds. However, r_G would then be complex and Eq. (3.49) would be replaced by its frequency domain equivalent or $u_1^+(t)$ would be related to $u_G(t)$ and $u_1^-(t)$ by a differential equation. Normally the impedances Z_G and Z_L are taken to be real for simplicity.

As an example, the complete diagram representing wave propagation in a two tube model is shown in Fig. 3.37. The volume velocity at the lips is defined as $u_L(t) = u_2(l_2,t)$. Writing the equations for this system in the frequency domain, the frequency response of the system can be shown to be

$$\begin{aligned} V_a(\Omega) &= \frac{U_L(\Omega)}{U_G(\Omega)} \\ &= \frac{0.5(1+r_G)(1+r_L)(1+r_1)e^{-j\Omega(\tau_1+\tau_2)}}{1 + r_1 r_G e^{-j\Omega 2\tau_1} + r_1 r_L e^{-j\Omega 2\tau_2} + r_L r_G e^{-j\Omega 2(\tau_1+\tau_2)}} \end{aligned} \tag{3.51}$$

(See Problem 3.7.) Several features of $V_a(\Omega)$ are worth pointing out. First, note the factor $e^{-j\Omega(\tau_1+\tau_2)}$ in the numerator. This represents simply the tota

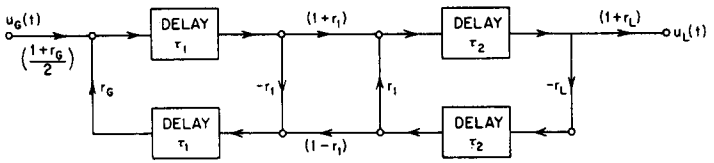


Fig. 3.37 Complete flow diagram of a two-tube model.

propagation delay in the system from glottis to lips. The system function of the system is found by replacing $j\Omega$ by s in Eq. (3.51), with the result

$$V_a(s) = \frac{0.5(1+r_G)(1+r_L)(1+r_1)e^{-s(\tau_1+\tau_2)}}{1 + r_1r_Ge^{-s2\tau_1} + r_1r_Le^{-s2\tau_2} + r_Lr_Ge^{-s2(\tau_1+\tau_2)}} \tag{3.52}$$

The poles of $V_a(s)$ are the complex resonance frequencies of the system. We see that there will be an infinite number of poles because of the exponential dependence upon s. Fant [1] and Flanagan [2] show that through proper choice of section lengths and cross-sectional areas, realistic formant frequency distributions can be obtained for vowels. (Also see Problem 3.8.)

3.3.3 Relationship to digital filters

The form of $V_a(s)$ for the two tube model suggests that lossless tube models have many properties in common with digital filters. To see this, let us consider a system composed of N lossless tubes each of length $\Delta x = l/N$,

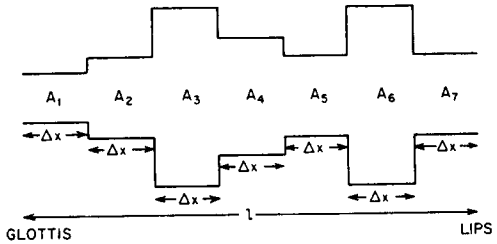


Fig. 3.38 Concatenation of (N=7) lossless tubes of equal length.

where l is the overall length of the vocal tract. Such a system is depicted in Figure 3.38 for $N = 7$. Wave propagation in this system can be represented as in Fig. 3.34 with all the delays being equal to $\tau = \Delta x/c$, the time to propagate the length of one tube. It is instructive to begin by considering the response of the system to a unit impulse source, $u_G(t) = \delta(t)$. The impulse propagates down the series of tubes, being partially reflected and partially propagated at the junctions. A detailed consideration of this process will confirm that the impulse response (i.e., the volume velocity at the lips due to an impulse at the glottis) will be of the form

$$v_a(t) = \alpha_0\delta(t-N\tau) + \sum_{k=1}^{\infty} \alpha_k\delta(t-N\tau-2k\tau) \tag{3.53}$$

Clearly, the soonest that an impulse can reach the output is $N\tau$ sec. Then successive impulses due to reflections at the junctions reach the output at multiples of 2τ seconds later. The quantity 2τ is the time required to propagate both ways in one section. The system function of such a system will be of the form

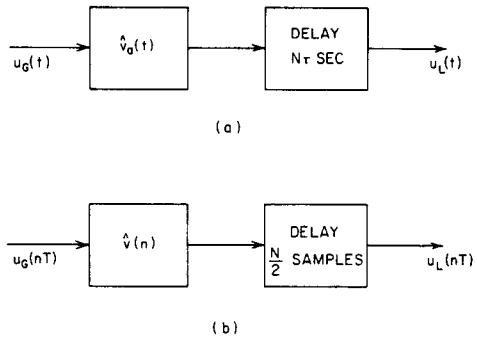


Fig. 3.39 (a) Block diagram representation of lossless acoustic tube model; (b) equivalent discrete-time system.

$$V_a(s) = \sum_{k=0}^{\infty} \alpha_k e^{-s(N+2k)\tau}$$

$$= e^{-sN\tau} \sum_{k=0}^{\infty} \alpha_k e^{-s2\tau k} \tag{3.54}$$

The factor $e^{-sN\tau}$ corresponds to the delay time required to propagate through all N sections. The quantity

$$\hat{V}_a(s) = \sum_{k=0}^{\infty} \alpha_k e^{-sk2\tau} \tag{3.55}$$

is the system function of a linear system whose impulse response is simply $\hat{v}_a(t) = v_a(t+N\tau)$. This part represents the resonance properties of the system. Figure 3.39a is a block diagram representation of the lossless tube model showing the separation of the system $\hat{v}_a(t)$ from the delay. The frequency response $\hat{V}_a(\Omega)$ is

$$\hat{V}_a(\Omega) = \sum_{k=0}^{\infty} \alpha_k e^{-j\Omega k2\tau} \tag{3.56}$$

It is easily shown that

$$\hat{V}_a(\Omega + \frac{2\pi}{2\tau}) = \hat{V}_a(\Omega) \tag{3.57}$$

This is, of course, very reminiscent of the frequency response of a discrete-time system. In fact, if the input to the system (i.e., the excitation) is bandlimited to frequencies below $\pi/(2\tau)$, then we can sample the input with period $T = 2\tau$ and filter the sampled signal with a digital filter whose impulse response is

$$\hat{v}(n) = \alpha_n \quad n \geq 0$$

$$= 0 \quad n < 0 \tag{3.58}$$

For a sampling period of $T = 2\tau$, the delay of $N\tau$ sec corresponds to a shift of

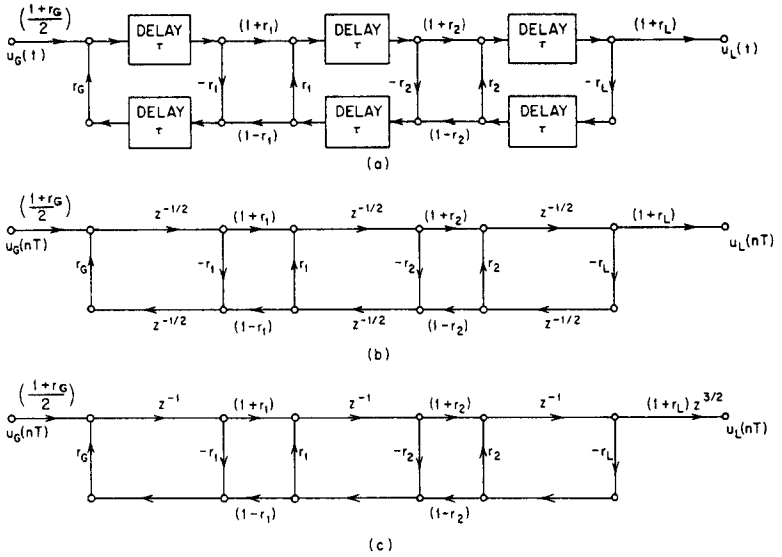


Fig. 3.40 (a) Signal flow graph for lossless tube model of the vocal tract; (b) equivalent discrete-time system; (c) equivalent discrete-time system using only whole delays in ladder part.

$N/2$ samples. Thus, the equivalent discrete time system for bandlimited inputs is shown in Fig. 3.39b. Note that if N is even, $N/2$ is an integer and the delay can be implemented by simply shifting the output sequence of the first system. If N is odd, however, an interpolation would be required to obtain samples of the output of Fig. 3.39a. This delay would most likely be ignored or avoided in some way (see below) since it is of no consequence in most applications of speech models.

The z-transform of $\hat{v}(n)$ is simply $\hat{V}_a(s)$ with e^{sT} replaced by z. Thus,

$$\hat{V}(z) = \sum_{k=0}^{\infty} \alpha_k z^{-k} \tag{3.59}$$

A signal flow graph for the equivalent discrete-time system can be obtained from the flow graph of the analog system in an analogous way. Specifically, each node variable in the analog system is replaced by the corresponding sequence of samples. Also each τ sec delay is replaced by a 1/2 sample delay, since $\tau = T/2$. An example is depicted in Figure 3.40. Note in particular that the propagation delay is represented in Fig. 3.40b by a transmittance of $z^{-1/2}$.

The 1/2 sample delays in Fig. 3.40b imply an interpolation half-way between sample values. Such interpolation is impossible to implement exactly. A more desirable configuration can be obtained by observing that the structure of Fig. 3.40b has the form of a ladder, with the delay elements only in the

upper and lower paths. Signals propagate to the right in the upper path and to the left in the lower path. We can see that the delay around any closed path in Fig. 3.40b will be preserved if the delays in the lower branches are literally moved up to the corresponding branches directly above. The overall delay from input to output will then be wrong but this is of minor significance in practice and theoretically can be compensated by the insertion of the correct amount of advance (in general $z^{N/2}$).[7] Figure 3.40c shows how this is done for the three tube example. The advantage of this form is that difference equations can be written for this system and these difference equations can be used iteratively to compute samples of the output from samples of the input.

Digital networks [33] such as Fig. 3.40c can be used to compute samples of a synthetic speech signal from samples of an appropriate excitation signal [32]. In such applications, the structure of the network representation determines the complexity of the operations required to compute each output sample. Each branch whose transmittance is not unity requires a multiplication. We see that each junction requires 4 multiplications and 2 additions. Generalizing from Fig. 3.40c, we see that $4N$ multiplications and $2N$ additions are required to implement an N-tube model. Since multiplications often are the most time consuming operation, it is of interest to consider other structures (literally, other organizations of the computations) which may require fewer multiplications. These can easily be derived by considering a typical junction as depicted in Fig. 3.41a. The difference equations represented by this diagram are

$$u^+(n) = (1+r)\,w^+(n) + ru^-(n) \tag{3.60a}$$

$$w^-(n) = -\,rw^+(n) + (1-r)\,u^-(n) \tag{3.60b}$$

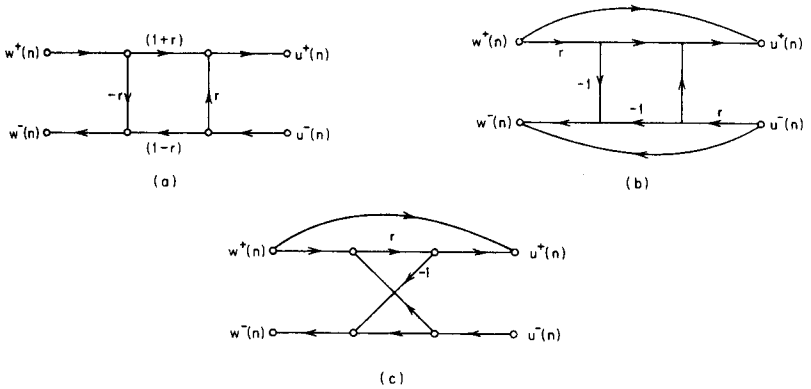


Fig. 3.41 (a) 4 multiplier representation of lossless tube junction; (b) 2 multiplier configuration; (c) 1 multiplier configuration.

[7]Note that we could also move all the delay to the lower branches. In this case, the delay through the system could be corrected by inserting a *delay* of $N/2$ samples.

These equations can be written as

$$u^+(n) = w^+(n) + rw^+(n) + ru^-(n) \tag{3.61a}$$

$$w^-(n) = -rw^+(n) - ru^-(n) + u^-(n) \tag{3.61b}$$

Noting that the terms $rw^+(n)$ and $ru^-(n)$ occur in both equations, 2 out of the 4 multiplications in Eqs. (3.60) can be eliminated as shown in Fig. 3.41b. Note that this configuration requires 2 multiplications and 4 additions. Still another implementation follows from grouping terms involving r as in

$$u^+(n) = w^+(n) + r[w^+(n)+u^-(n)] \tag{3.62a}$$

$$w^-(n) = u^-(n) - r[w^+(n)+u^-(n)] \tag{3.62b}$$

Now, since the term $r[w^+(n)+u^-(n)]$ occurs in both equations, this configuration requires only 1 multiplication and 3 additions as shown in Fig. 3.41c. This form of the lossless tube model was first obtained by Itakura and Saito [34]. When using the lossless tube model for speech synthesis, the choice of computational structure depends on the speed with which multiplications and additions can be done, and the ease of controlling the computation.

3.3.4 Transfer function of the lossless tube model

To complete our discussion of lossless tube discrete-time models for speech production it is instructive to derive a general expression for the transfer function in terms of the reflection coefficients. Equations of the type that we shall derive have been obtained before by Atal and Hanauer [35], Markel and Gray [36], and Wakita [37] in the context of linear predictive analysis of speech. We shall return to a consideration of lossless tube models and their relation to linear predictive analysis in Chapter 8. Our main concern at this point is the general form of the transfer function and the variety of other models suggested by the lossless tube model.

Let us begin by noting that we seek the transfer function

$$V(z) = \frac{U_L(z)}{U_G(z)} \tag{3.63}$$

To find $V(z)$, it is most convenient to express $U_G(z)$ in terms of $U_L(z)$ and then solve for the ratio above. To do this, let us consider Figure 3.42 which depicts a junction in the lossless tube model. The z-transform equations for this junction are

$$U_{k+1}^+(z) = (1+r_k)z^{-1/2}U_k^+(z) + r_kU_{k+1}^-(z) \tag{3.64a}$$

$$U_k^-(z) = -r_kz^{-1}U_k^+(z) + (1-r_k)z^{-1/2}U_{k+1}^-(z) \tag{3.64b}$$

Solving for $U_k^+(z)$ and $U_k^-(z)$ we obtain

$$U_k^+(z) = \frac{z^{1/2}}{1+r_k}U_{k+1}^+(z) - \frac{r_kz^{1/2}}{1+r_k}U_{k+1}^-(z) \tag{3.65a}$$

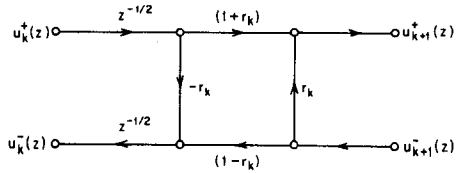


Fig. 3.42 Flow graph representing relationship among z-transforms at a junction.

$$U_k^-(z) = \frac{-r_kz^{-1/2}}{1+r_k}U_{k+1}^+(z) + \frac{z^{-1/2}}{1+r_k}U_{k+1}^-(z) \tag{3.65b}$$

Equations (3.65) permit us to work backwards from the output of the lossless tube model to obtain $U_G(z)$ in terms of $U_L(z)$.

To make the result more compact it is helpful to represent the boundary condition at the lips in the same manner as all the junctions in the system. Toward this end, we define $U_{N+1}(z)$ to be the z-transform of the input to a fictitious $(N+1)^{st}$ tube that is infinitely long so that there is no negative-going wave in the $(N+1)^{st}$ tube. An equivalent point of view is that the $(N+1)^{st}$ tube is terminated in its characteristic impedance. In any case, $U_{N+1}^+(z) = U_L(z)$ and $U_{N+1}^-(z) = 0$. Then from Eqs. (3.39) and (3.45) we see that if $A_{N+1} = \rho c/Z_L$, we can define $r_N = r_L$.

Now, Eqs. (3.65) can be expressed in matrix form as

$$\mathbf{U}_k = \mathbf{Q}_k\mathbf{U}_{k+1} \tag{3.66}$$

where

$$\mathbf{U}_k = \begin{bmatrix} U_k^+(z) \\ U_k^-(z) \end{bmatrix} \tag{3.67}$$

and

$$\mathbf{Q}_k = \begin{bmatrix} \dfrac{z^{1/2}}{1+r_k} & \dfrac{-r_kz^{1/2}}{1+r_k} \\ \dfrac{-r_kz^{-1/2}}{1+r_k} & \dfrac{z^{-1/2}}{1+r_k} \end{bmatrix} \tag{3.68}$$

By repeatedly applying Eq. (3.66), it can be easily shown that the variables at the input to the first tube can be expressed in terms of the variables at the output by the matrix product

$$\begin{aligned} \mathbf{U}_1 &= \mathbf{Q}_1 \cdot \mathbf{Q}_2 \cdots \mathbf{Q}_N\mathbf{U}_{N+1} \\ &= \prod_{k=1}^{N} \mathbf{Q}_k \cdot \mathbf{U}_{N+1} \end{aligned} \tag{3.69}$$

From Fig. 3.36 it can be seen that the boundary condition at the glottis can be expressed as

$$U_G(z) = \frac{2}{(1+r_G)}\, U_1^+(z) - \frac{2r_G}{1+r_G}\, U_1^-(z) \tag{3.70}$$

which can also be expressed as

$$U_G(z) = \left[\frac{2}{1+r_G}\,,\; -\frac{2r_G}{1+r_G}\right] \mathbf{U}_1 \tag{3.71}$$

Thus, since

$$\mathbf{U}_{N+1} = \begin{bmatrix} U_L(z) \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} U_L(z) \tag{3.72}$$

we can at last write

$$\frac{U_G(z)}{U_L(z)} = \left[\frac{2}{1+r_G}\,,\; -\frac{2r_G}{1+r_G}\right] \prod_{k=1}^{N} \mathbf{Q}_k \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3.73}$$

which is equal to $1/V(z)$.

To examine the properties of $V(z)$, it is helpful to first express $\mathbf{Q}_k$ as

$$\mathbf{Q}_k = z^{1/2} \begin{bmatrix} \dfrac{1}{1+r_k} & \dfrac{-r_k}{1+r_k} \\ \dfrac{-r_k z^{-1}}{1+r_k} & \dfrac{z^{-1}}{1+r_k} \end{bmatrix}$$

$$= z^{1/2}\, \hat{\mathbf{Q}}_k \tag{3.74}$$

Thus, Eq. (3.73) can be expressed as

$$\frac{1}{V(z)} = z^{N/2} \left[\frac{2}{1+r_G}\,,\; -\frac{2r_G}{1+r_G}\right] \prod_{k=1}^{N} \hat{\mathbf{Q}}_k \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3.75}$$

First, we note that since the elements of the matrices $\hat{\mathbf{Q}}_k$ are either constant or proportional to z^{-1}, the complete matrix product will reduce to a polynomial in the variable z^{-1} of order N. For example it can be shown (see Problem 3.9) that for $N = 2$,

$$\frac{1}{V(z)} = \frac{2(1+r_1r_2z^{-1}+r_1r_Gz^{-1}+r_2r_Gz^{-2})z}{(1+r_G)(1+r_1)(1+r_2)} \tag{3.76}$$

or

$$V(z) = \frac{0.5(1+r_G)(1+r_1)(1+r_2)z^{-1}}{1+(r_1r_2+r_1r_G)z^{-1}+r_2r_Gz^{-2}} \tag{3.77}$$

In general, it can be seen from Eqs. (3.74) and (3.75) that for a lossless tube model, the transfer function can always be expressed as

$$V(z) = \frac{0.5(1+r_G)\prod_{k=1}^{N}(1+r_k)z^{-N/2}}{D(z)} \tag{3.78a}$$

where $D(z)$ is a polynomial in z^{-1} given by the matrix

$$D(z) = [1, -r_G] \begin{bmatrix} 1 & -r_1 \\ -r_1z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_N \\ -r_Nz^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3.78b}$$

It can be seen from Eq. (3.78b) that $D(z)$ will have the form

$$D(z) = 1 - \sum_{k=1}^{N} \alpha_k z^{-k} \tag{3.79}$$

In other words, the transfer function of a lossless tube model has a delay corresponding to the number of sections of the model and it has no zeros – only poles. These poles, of course, define the resonances or formants of the lossless tube model.

In the special case $r_G = 1$ ($Z_G = \infty$), the polynomial $D(z)$ can be found using a recursion formula that can be derived from Eq. (3.78b). If we begin by evaluating the matrix product from the left, we will always be multiplying a 1×2 row matrix by a 2×2 matrix until finally we multiply by the 2×1 column vector on the right in Eq. (3.78b). The desired recursion formula becomes evident after evaluating the first few matrix products. Let us define

$$\mathbf{P}_1 = [1, -1] \begin{bmatrix} 1 & -r_1 \\ -r_1z^{-1} & z^{-1} \end{bmatrix} = \left[(1+r_1z^{-1}),\; -(r_1+z^{-1})\right] \tag{3.80}$$

If we define

$$D_1(z) = 1 + r_1z^{-1} \tag{3.81}$$

then it is easily shown that

$$\mathbf{P}_1 = [D_1(z),\; -z^{-1}D_1(z^{-1})] \tag{3.82}$$

Similarly, the row matrix $\mathbf{P}_2$ is defined as

$$\mathbf{P}_2 = \mathbf{P}_1 \begin{bmatrix} 1 & -r_2 \\ -r_2z^{-1} & z^{-1} \end{bmatrix} \tag{3.83}$$

If the indicated multiplication is carried out it is easily shown that

$$\mathbf{P}_2 = [D_2(z),\; -z^{-2}D_2(z^{-1})] \tag{3.84}$$

where

$$D_2(z) = D_1(z) + r_2z^{-2}D_1(z^{-1}) \tag{3.85}$$

By induction it can be shown that

$$\mathbf{P}_k = \mathbf{P}_{k-1}\begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix}$$

$$= [D_k(z),\ -z^{-k}D_k(z^{-1})] \tag{3.86}$$

where

$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) \tag{3.87}$$

Finally, the desired polynomial $D(z)$ is

$$D(z) = \mathbf{P}_N \begin{bmatrix} 1 \\ 0 \end{bmatrix} = D_N(z) \tag{3.88}$$

Thus, we can see that it is not necessary to carry out all the matrix multiplies but we can simply evaluate the recursion

$$D_0(z) = 1 \tag{3.89a}$$

$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) \quad k = 1, 2, \ldots, N \tag{3.89b}$$

$$D(z) = D_N(z) \tag{3.89c}$$

The effectiveness of the lossless tube model can be demonstrated by computing the transfer function for the area function data used to compute Figures 3.23-3.26. To do this we must decide upon the termination at the lips and the number of sections to use. In our derivations, we have represented the radiation load as a tube of area A_{N+1} which has no reflected wave. The value of A_{N+1} is chosen to give the desired reflection coefficient at the output. This is the only source of loss in the system (if $r_G = 1$), and thus it is to be expected that the choice of A_{N+1} will control the bandwidths of the resonances of $V(z)$. For example, $A_{N+1} = \infty$ gives $r_N = r_L = 1$, the reflection coefficient for an acoustic short circuit. This, of course, is the completely lossless case. Usually A_{N+1} would be chosen to give a reflection coefficient at the lips which produces reasonable bandwidths for the resonances. An example is presented below.

The choice of number of sections depends upon the sampling rate chosen to represent the speech signal. Recall that the frequency response of the lossless tube model is periodic; and thus, the model can only approximate the vocal tract behavior in a band of frequencies $|F| < 1/(2T)$, where T is the sampling period. We have seen that this requires $T = 2\tau$, where τ is the one-way propagation time in a single section. If there are N sections, for a total length, l, then $\tau = l/(cN)$. Since the order of the denominator polynomial is N, there can be at most $N/2$ complex conjugate poles to provide resonances in the band $|F| < 1/(2T)$. Using the above value for τ with $l = 17.5$ cm and $c = 35000$ cm/sec, we see that

$$\frac{1}{2T} = \frac{1}{4\tau} = \frac{Nc}{4l} = \frac{N}{2}\,(1000)\ \text{Hz} \tag{3.90}$$

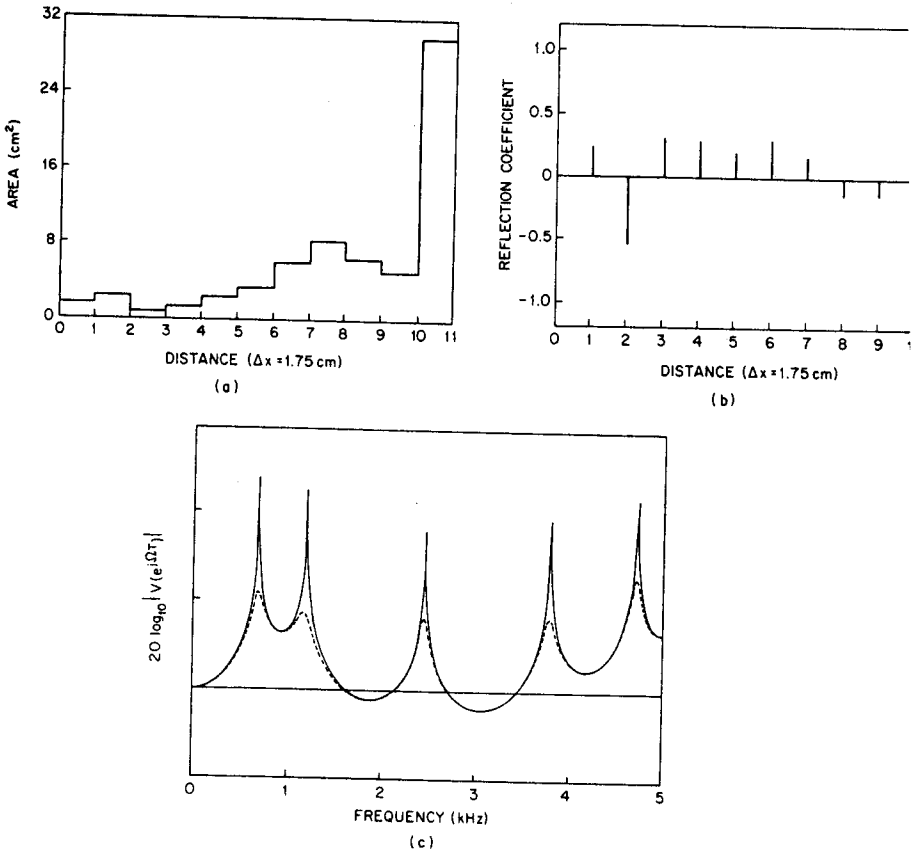


Fig. 3.43 (a) Area function for 10 section lossless tube terminated with reflectionless section of area 30 cm^2; (b) reflection coefficients for 10 section tube; (c) frequency response of 10 section tube; dotted curve corresponds to conditions of (b); solid curve corresponds to short-circuit termination. (Note area data of (a) estimated from data given by Fant [1] for the Russian vowel /a/.)

This implies that there will be about $N/2$ resonances (formants) per 1000 Hz of frequency for a vocal tract of total length 17.5 cm. For example, if $1/T = 10000$ Hz, then the baseband is 5000 Hz. This implies that N should be 10. A glance at Figures 3.21 through 3.26 confirm that vocal tract resonances seem to occur with a density of about one formant per 1000 Hz. Shorter overall vocal tract lengths will have fewer resonances per kiloHertz and vice versa.

Figure 3.43 shows an example for $N = 10$ and $1/T = 10$ kHz. Figure 3.43a shows the area function data of Fig. 3.23 sampled to give a 10 tube

approximation for the vowel /a/. Figure 3.43b shows the resulting set of 10 reflection coefficients for $A_{11} = 30$ cm^2. This gives a reflection coefficient at the lips of $r_N = 0.714$. Note that the largest reflection coefficients occur where the relative change in area is greatest. Figure 3.43c shows the frequency response curves for $r_N = 1$ and $r_N = .714$ (dotted curve). A comparison of the dotted curve of Figure 3.43c to Fig. 3.23 confirms that with appropriate loss at the lip boundary, the frequency response of the lossless tube model is very much like that of the more detailed model.

3.4 Digital Models for Speech Signals

We have seen in Section 3.2 that it is possible to derive rather detailed mathematical representations of the acoustics of speech production. Our purpose in surveying this theory is to call attention to the basic features of the speech signal and to show how these features are related to the physics of speech production. We have seen that sound is generated in 3 ways, and that each mode results in a distinctive type of output. We have also seen that the vocal tract imposes its resonances upon the excitation so as to produce the different sounds of speech. This is the essence of what we have learned so far.

An important idea should now be emerging from this lengthy discussion of models. It is simply that a valid approach to representation of speech signals is in terms of a "terminal analog" model such as depicted before in Fig. 3.31; that is, a linear system whose output has the desired speech-like properties when controlled by a set of parameters that are somehow related to the process of speech production. The model is thus equivalent to the physical model at its terminals (output) but its internal structure does not mimic the physics of speech production. In particular, we are interested in discrete-time terminal analog models for representing sampled speech signals.

To produce a speech-like signal the mode of excitation and the resonance properties of the linear system must change with time. The nature of this time variation can be seen in Section 3.1. In particular, waveform plots such as Fig. 3.3a show that the properties of the speech signal change relatively slowly with time. For many speech sounds it is reasonable to assume that the general properties of the excitation and vocal tract remain fixed for periods of 10-20 msec. Thus, a terminal analog model involves a slowly time-varying linear system excited by an excitation signal whose basic nature changes from quasi-periodic pulses for voiced speech to random noise for unvoiced speech.

The lossless tube discrete-time model of the previous section serves as an example of what we mean. The essential features of that model are depicted in Fig. 3.44a. Recall that the vocal tract system was characterized by a set of areas or, equivalently, reflection coefficients. Systems of the form of Fig. 3.40c can thus be used to compute the speech output given an appropriate input. We showed that the relationship between the input and output could be represented by a transfer function, $V(z)$, of the form

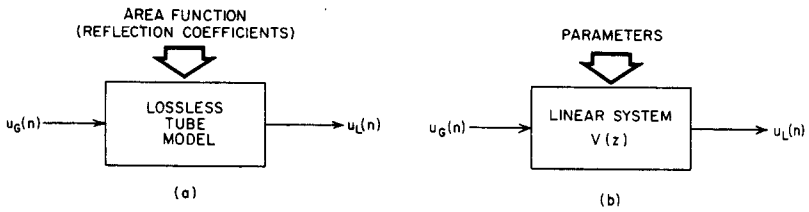


Fig. 3.44 (a) Block diagram representation of the lossless tube model; (b) terminal analog model.

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \tag{3.91}$$

where G and $\{\alpha_k\}$ depend upon the area function. (Note that the fixed delay in Eq. (3.78a) has been dropped.) Insofar as the output is concerned, any system having this transfer function will produce the same output in response to a given input. (This is not strictly true for time-varying systems, but differences can be minimized by careful implementation.) Thus, discrete-time terminal analog models take the general form of Fig. 3.44b. This leads to a consideration of alternative implementations of the vocal tract filter.

In addition to the vocal tract response a complete terminal analog model includes a representation of the changing excitation function and the effects of sound radiation at the lips. In the remainder of this section we shall examine each of the model components separately, and then combine them into a complete model.

3.4.1 Vocal tract

The resonances (formants) of speech correspond to the poles of the transfer function $V(z)$. An all-pole model is a very good representation of vocal tract effects for a majority of speech sounds; however, the acoustic theory tells us that nasals and fricatives require both resonances and anti-resonances (poles and zeros). In these cases, we may include zeros in the transfer function or we may reason with Atal [35] that effect of a zero of the transfer function can be achieved by including more poles. (See Problem 3.10.) In most cases this approach is to be preferred.

Since the coefficients of the denominator of $V(z)$ in Eq. (3.91) are real, the roots of the denominator polynomial will be either real or occur in complex conjugate pairs. A typical complex resonant frequency of the vocal tract is

$$s_k, s_k^* = -\sigma_k \pm j2\pi F_k \tag{3.92}$$

The corresponding complex conjugate poles in the discrete-time representation would be

$$\begin{aligned} z_k, z_k^* &= e^{-\sigma_k T} e^{\pm j2\pi F_k T} \\ &= e^{-\sigma_k T}\cos(2\pi F_k T) \pm j e^{-\sigma_k T}\sin(2\pi F_k T) \end{aligned} \tag{3.93}$$

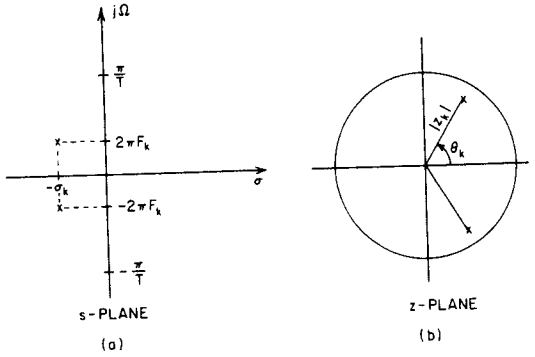


Fig. 3.45 (a) s-plane; and (b) z-plane representations of a vocal tract resonance.

The bandwidth of the vocal tract resonance is approximately $2\sigma_k$ and the center frequency is $2\pi F_k$ [26]. In the z-plane, the radius from the origin to the pole determines the bandwidth, i.e.,

$$|z_k| = e^{-\sigma_k T} \tag{3.94a}$$

and the z-plane angle is

$$\theta_k = 2\pi F_k T \tag{3.94b}$$

Thus if the denominator of $V(z)$ is factored, the corresponding analog formant frequencies and bandwidths can be found using Eqs. (3.94). As shown in Figure 3.45 the complex natural frequencies of the human vocal tract are all in the left half of the s-plane since it is a stable system. Thus, $\sigma_k > 0$, and therefore $|z_k| < 1$; i.e., all of the corresponding poles of the discrete-time model must be inside the unit circle as required for stability. Figure 3.45 depicts typical complex resonant frequencies in both the s-plane and the z-plane.

In Section 3.3 we showed how a lossless tube model leads to a transfer function of the form of Eq. (3.91). It can be shown [35,36] that as long as the areas of the tube model are positive, all the poles of the corresponding $V(z)$ will be inside the unit circle. Conversely, it can be shown that given a transfer function, $V(z)$, as in Eq. (3.91), a lossless tube model can be found [35,36]. Thus, one way to implement a given transfer function is to use a ladder structure as in Fig. 3.40c, possibly incorporating one of the junction forms of Fig. 3.41. Another approach is to use one of the standard digital filter implementation structures given in Chapter 2. For example we could use a direct form implementation of $V(z)$ as depicted in Fig. 3.46a. Alternatively, we can represent $V(z)$ as a cascade of second order systems (resonators); i.e.,

$$V(z) = \prod_{k=1}^{M} V_k(z) \tag{3.95}$$

where M is the largest integer in $((N+1)/2)$, and

$$V_k(z) = \frac{(1-2|z_k|\cos(2\pi F_k T) + |z_k|^2)}{(1-2|z_k|\cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2})} \tag{3.96}$$

The numerator of $V_k(z)$ is chosen so that the product will have the same gain as the lossless tube model. Note that at zero frequency ($z = 1$), $V_k(1) = 1$. A cascade model is depicted in Fig. 3.46b. Problem 3.11 shows a novel way of eliminating multiplications in cascade models. Still another approach to implementing the system $V(z)$ is to make a partial fraction expansion of $V(z)$ and thus obtain a parallel form model. This approach is explored in Problem 3.12.

It is interesting to note that cascade and parallel models were first considered as analog models. In this context there is a serious limitation, since analog second order systems (resonators) have frequency responses that die away with frequency. This led Fant [1] to derive "higher pole correction" factors that were cascaded with the analog formant resonators to achieve proper high frequency spectral balance. When digital simulations began to be used, Gold and Rabiner [38] observed that digital resonators had, by virtue of their inherent periodicity, the correct high frequency behavior. We have, of course, already seen this in the context of the lossless tube model. Thus no "higher pole correction" network is required in digital simulations.

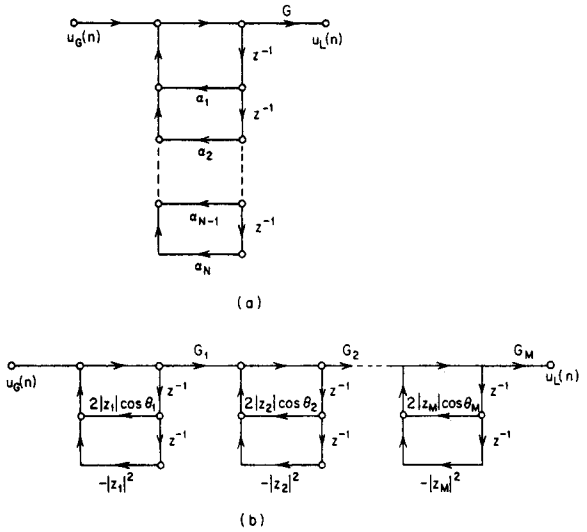


Fig. 3.46 (a) Direct form implementation of all-pole transfer function; (b) cascade implementation of all-pole transfer function ($G_k = 1 - 2|z_k|\cos\theta_k + |z_k|^2$).

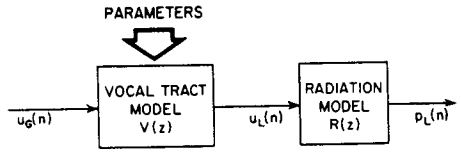


Fig. 3.47 Terminal analog model including radiation effects.

3.4.2 Radiation

So far we have considered the transfer function $V(z)$ which relates volume velocity at the source to volume velocity at the lips. If we wish to obtain a model for pressure at the lips (as is usually the case), then the effects of radiation must be included. We saw in Section 3.2.4 that in the analog model, the pressure and volume velocity are related by Eqs. (3.29). We desire a similar z-transform relation of the form

$$P_L(z) = R(z)\,U_L(z) \tag{3.97}$$

It can be seen from the discussion of Section 3.2.4 and from Fig. 3.20 that pressure is related to volume velocity by a highpass filtering operation. In fact, at low frequencies it can be argued that the pressure is approximately the derivative of the volume velocity. Thus, to obtain a discrete-time representation of this relationship we must use a digitization technique that avoids aliasing. For example, by using the bilinear transform method of digital filter design [33] it can be shown (see Problem 3.13) that a reasonable approximation to the radiation effects is obtained with

$$R(z) = R_0(1-z^{-1}) \tag{3.98}$$

i.e., a first backward difference. (A more accurate approximation is also considered in Problem 3.13.) The crude "differentiation" effect of the first difference is consistent with the approximate differentiation at low frequencies that is commonly assumed.

This radiation "load" can be cascaded with the vocal tract model as in Fig. 3.47. $V(z)$ can be implemented in any convenient way and the required parameters will, of course, be appropriate for the chosen configuration; e.g., area function for the lossless tube model or formant frequencies and bandwidths for the cascade model.

3.4.3 Excitation

To complete our terminal analog model, we must discuss means for generating an appropriate input to the vocal tract radiation system. Recalling that the majority of speech sounds can be classed as either voiced or voiceless, we see that in general terms what is required is a source that can produce either a quasi-periodic pulse waveform or a random noise waveform.

In the case of voiced speech, the excitation waveform must appear somewhat like the upper waveform in Fig. 3.30. A convenient way to represent the generation of the glottal wave is shown in Fig. 3.48. The impulse train generator produces a sequence of unit impulses which are spaced by the desired fundamental period. This signal in turn excites a linear system whose impulse response $g(n)$ has the desired glottal wave shape. A gain control, A_v, controls the intensity of the voiced excitation.

The choice of the form of $g(n)$ is probably not critical as long as its Fourier transform has the right properties. Rosenberg [39], in a study of the effect of glottal pulse shape on speech quality, found that the natural glottal pulse waveform could be replaced by a synthetic pulse waveform of the form

$$\begin{aligned} g(n) &= \frac{1}{2}\,[1-\cos(\pi n/N_1)] && 0 \leqslant n \leqslant N_1\,. \\ &= \cos(\pi(n-N_1)/2N_2) && N_1 \leqslant n \leqslant N_1+N_2 \\ &= 0 && \textit{otherwise} \end{aligned} \tag{3.99}$$

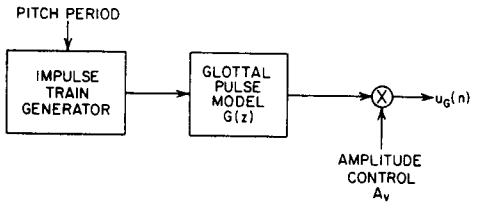


Fig. 3.48 Generation of the excitation signal for voiced speech.

This wave shape is very similar in appearance to the pulses in Fig. 3.30. Figure 3.49 shows the pulse waveform and its Fourier transform magnitude for typical values of N_1 and N_2. It can be seen that, as would be expected, the effect of the glottal pulse in the frequency domain is to introduce a lowpass filtering effect.

Since $g(n)$ in Eq. (3.99) has finite length, its z-transform, $G(z)$, has only zeros. An all-pole model is often more desirable. Good success has also been achieved using a two-pole model for $G(z)$ [36].

For voiceless sounds the excitation model is much simpler. All that is required is a source of random noise and a gain parameter to control the intensity of the unvoiced excitation. For discrete-time models, a random number generator provides a source of flat-spectrum noise. The probability distribution of the noise samples does not appear to be critical.

3.4.4 The complete model

Putting all the ingredients together we obtain the model of Figure 3.50. By switching between the voiced and unvoiced excitation generators we can

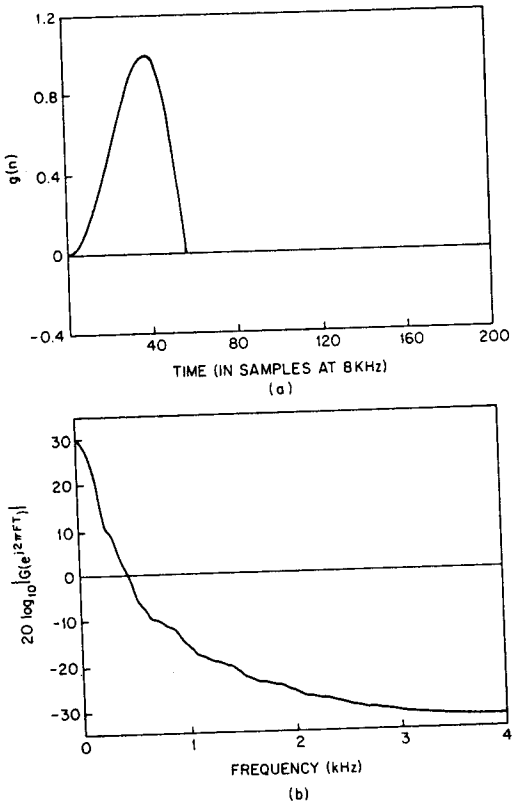


Fig. 3.49 (a) Rosenberg approximation to glottal pulse; (b) corresponding Fourier transform.

model the changing mode of excitation. The vocal tract can be modeled in a wide variety of ways as we have discussed. In some cases it is convenient to combine the glottal pulse and radiation models into a single system. In fact, we shall see that in the case of linear predictive analysis it is convenient to combine the glottal pulse, radiation and vocal tract components all together and represent them as a single transfer function

$$H(z) = G(z)V(z)R(z) \tag{3.100}$$

of the all-pole type. In other words Figure 3.50 is only a general representation. There is much latitude for modification.

A natural question at this point concerns the limitations of such a model. Certainly the model is far from the partial differential equations with which we began. Fortunately none of the deficiencies of this model severely limits its applicability. First, there is the question of time variation of the parameters.

In continuant sounds such as vowels, the parameters change very slowly and the model works very well. With transient sounds such as stops, the model is not as good but still adequate. It should be emphasized that our use of transfer functions and frequency response functions implicitly assumes that we can represent the speech signal on a "short-time" basis. That is, the parameters of the model are assumed to be constant over time intervals typically 10-20 msec long. The transfer function $V(z)$, then, really serves to define the structure of a model whose parameters vary slowly with time. We shall repeatedly invoke this principle of quasi-stationarity in subsequent chapters. A second limitation is the lack of provision for zeros as required theoretically for nasals and fricatives. This is definitely a limitation for nasals, but not too severe for fricatives. Zeros can be included in the model if desired. Third, the simple dichotomy of voiced-unvoiced excitation is inadequate for voiced fricatives. Simply adding the voiced and unvoiced excitations is inadequate since frication is correlated with the peaks of the glottal flow. A more sophisticated model for voiced fricatives has been developed [40] and can be employed when needed. Finally, a relatively minor concern is that the model of Fig. 3.50 requires that the glottal pulses be spaced by an integer multiple of the sampling period, T. Winham and Steiglitz [41] have considered ways of eliminating this limitation in situations requiring precise pitch control.

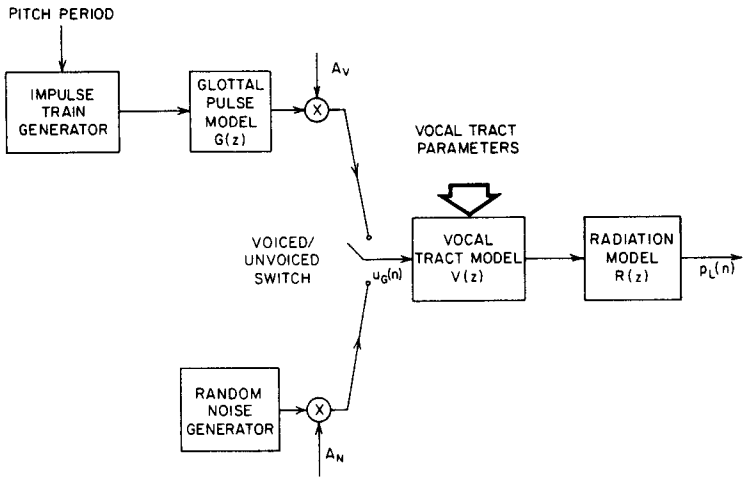


Fig. 3.50 General discrete-time model for speech production.

3.5 Summary

This chapter has focused upon three main areas: the sounds of speech, the physics of speech production, and discrete-time models for speech production. Our review of acoustic phonetics and the acoustic theory of speech production has

been lengthy but far from complete. Our purpose has been to provide adequate knowledge about the general properties of speech signals so as to motivate and suggest models that are useful for speech processing.

The models discussed in Sections 3.3 and 3.4 will be the basis for our discussion in the remainder of this book. We shall think of these models in two ways. One point of view is called speech analysis; the other is called speech synthesis. In speech analysis we are interested in techniques for estimating the parameters of the model from a natural speech signal that is assumed to be the output of the model. In speech synthesis, we wish to use the model to create a synthetic speech signal by controlling the model with suitable parameters. These two points of view will become intermingled in many cases and will arise in many problem areas. Underlying all our subsequent discussion will be models of the type discussed in this chapter. Having reviewed the subject of digital signal processing in Chapter 2 and the acoustic theory of speech production here, we are now ready to begin to see how digital signal processing techniques can be applied in processing speech signals.

REFERENCES

1. G. Fant, *Acoustic Theory of Speech Production,* Mouton, The Hague, 1970.
2. J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* 2nd Ed., Springer-Verlag, New York, 1972.
3. H. Fletcher, *Speech and Hearing in Communcation,* original edition, D. Van Nostrand Co., New York, 1953. Reprinted by Robert E. Krieger Pub. Co. Inc., New York, 1972.
4. T. Chiba and M. Kajiyama, *The Vowel, Its Nature and Structure,* Phonetic Society of Japan, 1958.
5. I. Lehiste, Ed., *Readings in Acoustic Phonetics,* MIT Press, Cambridge, Mass., 1967.
6. J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic Voices for Computers," *IEEE Spectrum,* Vol. 7, No. 10, pp. 22-45, October 1970.
7. W. Koenig, H. K. Dunn, and L. Y. Lacy, "The Sound Spectrograph," *J. Acoust. Soc. Am.,* Vol. 17, pp. 19-49, July 1946.
8. R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech,* D. Van Nostrand Co., New York, 1947. Republished by Dover Publications, Inc., 1966.
9. R. Jakobson, C. G. M. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates,* M.I.T. Press, Cambridge, Mass., 1963.

10. N. Chomsky and M. Halle, *The Sound Pattern of English,* Harper & Row, Publishers, New York, 1968.
11. G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *J. Acoust. Soc. Am.,* Vol. 24, No. 2, pp. 175-184, March 1952.
12. A. Holbrook and G. Fairbanks, "Diphthong Formants and Their Movements," *J. of Speech and Hearing Research,* Vol. 5, No. 1, pp. 38-58, March 1962.
13. O. Fujimura, "Analysis of Nasal Consonants," *J. Acoust. Soc. Am.,* Vol. 34, No. 12, pp. 1865-1875, December 1962.
14. J. M. Heinz and K. N. Stevens, "On the Properties of Voiceless Fricative Consonants," *J. Acoust. Soc. Am.,* Vol. 33, No. 5, pp. 589-596, May 1961.
15. P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic Loci and Transitional Cues for Consonants," *J. Acoust. Soc. Am.,* Vol. 27, No. 4, pp. 769-773, July 1955.
16. L. L. Beranek, *Acoustics,* McGraw-Hill Book Co., New York, 1954.
17. P. M. Morse and K. U. Ingard, *Theoretical Acoustics,* McGraw-Hill Book Co., New York, 1968.
18. M. R. Portnoff, "A Quasi-One-Dimensional Digital Simulation for the Time-Varying Vocal Tract," M. S. Thesis, Dept. of Elect. Engr., MIT, Cambridge, Mass., June 1973.
19. M. R. Portnoff and R. W. Schafer, "Mathematical Considerations in Digital Simulations of the Vocal Tract," *J. Acoust. Soc. Am.,* Vol. 53, No. 1 (Abstract), p. 294, January 1973.
20. M. M. Sondhi, "Model for Wave Propagation in a Lossy Vocal Tract," *J. Acoust. Soc. Am.,* Vol. 55, No. 5, pp. 1070-1075, May 1974.
21. J. S. Perkell, *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study,* MIT Press, Cambridge, Mass., 1969.
22. M. M. Sondhi and B. Gopinath, "Determination of Vocal-Tract Shape from Impulse Response at the Lips," *J. Acoust. Soc. Am.,* Vol. 49, No. 6 (Part 2), pp. 1847-1873, June 1971.
23. B. S. Atal, "Towards Determining Articulator Positions from the Speech Signal," *Proc. Speech Comm. Seminar,* Stockholm, Sweden, pp. 1-9, 1974.
24. R. B. Adler, L. J. Chu, and R. M. Fano, *Electromagnetic Energy Transmission and Radiation,* John Wiley and Sons, Inc., New York, 1963.
25. D. T. Paris and F. K. Hurd, *Basic Electromagnetic Theory,* McGraw-Hill Book Co., New York, 1969.
26. A. M. Bose and K. N. Stevens, *Introductory Network Theory,* Harper and Row, New York, 1965.

27. H. K. Dunn, "Methods of Measuring Vowel Formant Bandwidths," *J. Acoust. Soc. Am.*, Vol. 33, pp. 1737-1746, 1961.

28. J. L. Flanagan and L. L. Landgraf, "Self Oscillating Source for Vocal-Tract Synthesizers," *IEEE Trans. Audio and Electroacoustics*, Vol. AU-16, pp. 57-64, March 1968.

29. J. L. Flanagan and L. Cherry, "Excitation of Vocal-Tract Synthesizer," *J. Acoust. Soc. Am.*, Vol. 45, No. 3, pp. 764-769, March 1969.

30. K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," *Bell Syst. Tech. J.*, Vol. 50, No. 6, pp. 1233-1268, July-August 1972.

31. J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract," *Bell Sys. Tech J.*, Vol. 54, No. 3, pp. 485-506, March 1975.

32. J. L. Kelly, Jr. and C. Lochbaum, "Speech Synthesis," *Proc. Stockholm Speech Communications Seminar*, R.I.T., Stockholm, Sweden, September 1962.

33. A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.

34. F. Itakura and S. Saito, "Digital Filtering Techniques for Speech Analysis and Synthesis," *7th Int. Cong. on Acoustics*, Budapest, Paper 25 C1, 1971.

35. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, Vol. 50, No. 2 (Part 2), pp. 637-655, August 1971.

36. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

37. H. Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," *IEEE Trans. Audio and Electroacoustics*, Vol. AU-21, No. 5, pp. 417-427, October 1973.

38. B. Gold and L. R. Rabiner, "Analysis of Digital and Analog Formant Synthesizers," *IEEE Trans. Audio and Electroacoustics*, Vol. AU-16, pp. 81-94, March 1968.

39. A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *J. Acoust. Soc. Am.*, Vol. 49, No. 2, pp. 583-590, February 1971.

40. L. R. Rabiner, "Digital Formant Synthesizer for Speech Synthesis Studies," *J. Acoust. Soc. Am.*, Vol. 43, No. 4, pp. 822-828, April 1968.

41. G. Winham and K. Steiglitz, "Input Generators for Digital Sound Synthesis," *J. Acoust. Soc. Am.*, Vol. 47, No. 2, pp. 665-666, February 1970.

PROBLEMS

3.1 The waveform plot of Fig. P3.1 shows a 500 msec section (100 msec/line) of a speech waveform.

(a) Indicate the regions of voiced speech, unvoiced speech, and silence (background noise).

(b) For the voiced regions estimate the pitch period on a period-by-period basis and plot the pitch period versus time for this section of speech. (Let the period be indicated as zero during unvoiced and silence intervals.)

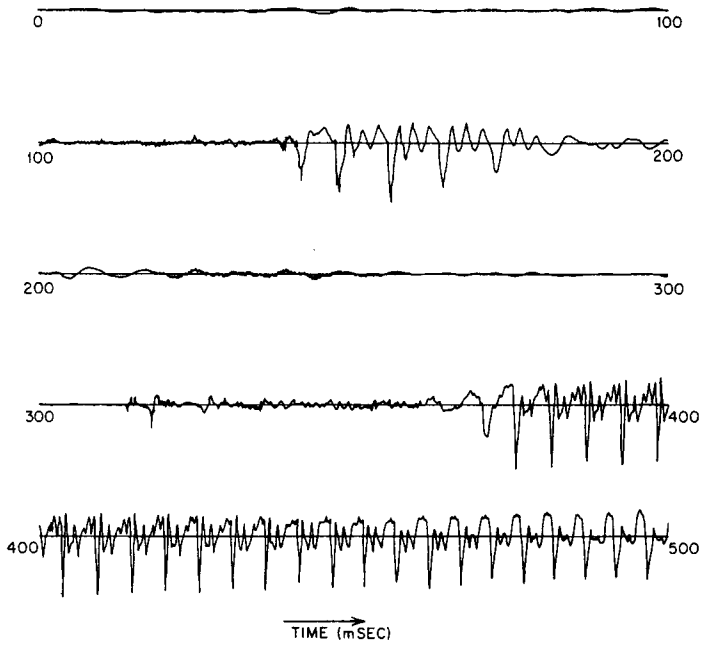


Fig. P3.1

3.2 The waveform plot of Fig. P3.2 is for the word "cattle." Note that each line of the plot corresponds to 100 msec of the signal.

(a) Indicate the boundaries between the phonemes; i.e. give the times corresponding to the boundaries /*c*/*a*/*tt*/*le*/.

(b) Indicate the point where the voice pitch frequency is (i) the highest; and (ii) the lowest. What are the approximate pitch frequencies at these points?

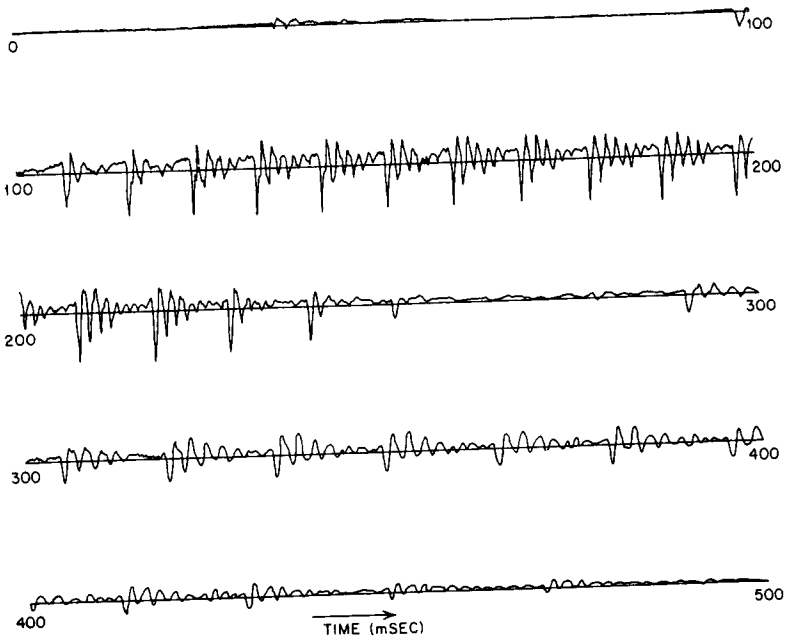


Fig. P3.2

(c) Is the speaker most probably a male, female, or a child? How do you know?

3.3 By substitution, show that Eqs. (3.3) are solutions to the partial differential equations of Eqs. (3.2).

3.4 Note that the reflection coefficients for the junction of two lossless acoustic tubes of areas A_k and A_{k+1} can be written as either

$$r_k = \frac{\frac{A_{k+1}}{A_k} - 1}{\frac{A_{k+1}}{A_k} + 1}$$

or

$$r_k = \frac{1 - \frac{A_k}{A_{k+1}}}{1 + \frac{A_k}{A_{k+1}}}$$

Show that since both A_k and A_{k+1} are positive,

$$-1 \leq r_k \leq 1$$

3.5 In determining the effect of the radiation load termination on a lossless tube model, it was assumed that Z_L was real and constant. A more realistic model is given by Eq. (3.29b).

(a) Beginning with the boundary condition

$$P_N(l_N, \Omega) = Z_L \cdot U_N(l_N, \Omega)$$

find a relation between the Fourier transforms of $u_N^-(t+\tau_N)$ and $u_N^+(t-\tau_N)$.

(b) From the frequency domain relation found in (a) and Eq. (3.29b), show that $u_N^-(t+\tau_N)$ and $u_N^+(t-\tau_N)$ satisfy the ordinary differential equation

$$L_r\left[R_r+\frac{\rho c}{A_N}\right]\frac{du_N^-(t+\tau_N)}{dt} + \frac{\rho c}{A_N} R_r u_N^-(t+\tau_N)$$
$$= L_r\left[R_r - \frac{\rho c}{A_N}\right]\frac{du_N^+(t-\tau_N)}{dt} - \frac{\rho c}{A_N} R_r u_N^+(t-\tau_N)$$

3.6 By substitution of Eq. (3.50) into Eq. (3.49), show that Eq. (3.48) and Eq. (3.49) are equivalent.

3.7 Consider the two-tube model of Fig. 3.37. Write the frequency domain equations for this model and show that the transfer function between the input and output volume velocities is given by Eq. (3.51).

3.8 Consider an ideal lossless tube model for the production of vowels consisting of 2 sections as shown in Fig. P3.8. Assume that the terminations at the glottis and lips are completely lossless. For the above conditions the system function of the lossless tube model will be obtained from Eq. (3.52) by substituting $r_G = r_L = 1$ and

$$r_1 = \frac{A_2 - A_1}{A_2 + A_1}.$$

(a) Show that the poles of the system are on the $j\Omega$ axis and are located at values of Ω satisfying the equations

$$\cos \Omega(\tau_1+\tau_2) + r_1 \cos \Omega(\tau_2-\tau_1) = 0$$

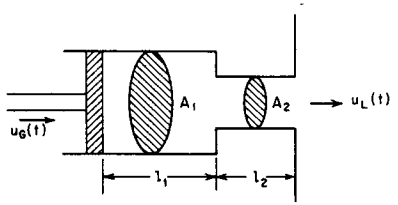


Fig. P3.8

or equivalently

$$\frac{A_1}{A_2}\tan(\Omega\tau_2) = \cot(\Omega\tau_1)$$

where $\tau_1 = l_1/c$, $\tau_2 = l_2/c$, and c is the velocity of sound.

(b) The values of Ω that satisfy the equations derived in (a) are the formant frequencies of the lossless tube model. By judicious choice of the parameters l_1, l_2, A_1, and A_2 we can approximate the vocal tract configurations of vowels, and by solving the above equations obtain the formant frequencies for the model. The following table gives parameters for several vowel configurations [2]. Solve for the formant frequencies for each case. (Note that the nonlinear equations must be solved graphically or iteratively.) Use $c = 35000$ cm/sec.

Vowel	l_1	A_1	l_2	A_2
/i/	9 cm	8 cm^2	6 cm	1 cm^2
/ae/	4 cm	1 cm^2	13 cm	8 cm^2
/a/	9 cm	1 cm^2	8 cm	7 cm^2
/Λ/	17 cm	6 cm^2	0	6 cm^2

3.9 By substituting the appropriate matrices $\hat{\mathbf{Q}}_1$ and $\hat{\mathbf{Q}}_2$ into Eq. (3.75) show that the transfer function of a two-tube discrete-time vocal tract model is given by Eq. (3.77).

3.10 Show that if $|a| < 1$,

$$1 - az^{-1} = \frac{1}{\sum_{n=0}^{\infty} a^n z^{-n}}$$

and thus, that a zero can be approximated as closely as desired by multiple poles.

3.11 The transfer function of a digital formant resonator is of the form

$$V_k(z) = \frac{1 - 2|z_k|\cos\theta_k + |z_k|^2}{1 - 2|z_k|\cos\theta_k z^{-1} + |z_k|^2 z^{-2}}$$

where $|z_k| = e^{-\sigma_k T}$ and $\theta_k = 2\pi F_k T$.

(a) Plot the locations of the poles of $V_k(z)$ in the z-plane. Also plot the corresponding analog poles in the s-plane.

(b) Write the difference equation relating the output, $y_k(n)$, of $V_k(z)$ to its input, $x_k(n)$.

(c) Draw a digital network implementation of the digital formant network with three multipliers.

(d) By rearranging the terms in the difference equation obtained in (b), draw a digital network implementation of the digital formant network that only requires two multiplications.

3.12 Consider the system function for a discrete-time vocal tract model

$$V(z) = \frac{G}{\prod_{k=1}^{N}(1 - z_k z^{-1})}$$

(a) Show that $V(z)$ can be expressed as the partial-fraction expansion

$$V(z) = \sum_{k=1}^{M}\left[\frac{G_k}{1 - z_k z^{-1}} + \frac{G_k^*}{1 - z_k^* z^{-1}}\right]$$

where M is the largest integer contained in $(N+1)/2$, and it is assumed that all the poles of $V(z)$ are complex. Give an expression for the G_k's in the above expression.

(b) Combine terms in the above partial fraction expansion to show that

$$V(z) = \sum_{k=1}^{M}\frac{B_k - C_k z^{-1}}{1 - 2|z_k|\cos\theta_k z^{-1} + |z_k|^2 z^{-2}}$$

where $z_k = |z_k|e^{j\theta_k}$. Give expressions for B_k and C_k in terms of G_k and z_k. This expression is the *parallel form* representation of $V(z)$.

(c) Draw the digital network diagram for the parallel form implementation of $V(z)$ for $M = 3$.

(d) For a given all-pole system function $V(z)$ which implementation would require the most multiplications – the parallel form or the cascade form as suggested in Problem 3.11?

3.13 The relationship between pressure and volume velocity at the lips is given by

$$P(l,s) = Z_L(s)\,U(l,s)$$

where $P(l,s)$ and $U(l,s)$ are the Laplace transforms of $p(l,t)$ and $u(l,t)$ respectively, and

$$Z_L(s) = \frac{sR_rL_r}{R_r + sL_r}$$

where

$$R_r = \frac{128}{9\pi^2} \quad \text{and} \quad L_r = \frac{8a}{3\pi c}$$

and c is the velocity of sound and a is the radius of the lip opening. In a discrete-time model, we desire a corresponding relationship of the form (Eq. (3.97))

$$P_L(z) = R(z)\,U_L(z)$$

where $P_L(z)$ and $U_L(z)$ are z-transforms of $p_L(n)$ and $u_L(n)$, the sampled versions of the bandlimited pressure and volume velocity.

One approach to obtaining $R(z)$ is to use the bilinear transformation, [33] i.e.,

$$R(z) = Z_L(s)\Big|_{s=\frac{2}{T}\left[\frac{1-z^{-1}}{1+z^{-1}}\right]}$$

(a) For $Z_L(s)$ as given above determine $R(z)$.
(b) Write the corresponding difference equation that relates $p_L(n)$ and $u_L(n)$.
(c) Give the locations of the pole and zero of $R(z)$.
(d) If $c = 35000$ cm/sec, $T = 10^{-4}$ sec^{-1}, and $0.5 \text{ cm} < a < 1.3$ cm, what is the range of pole values?
(e) A simple approximation to $R(z)$ obtained above is obtained by neglecting the pole; i.e.,

$$\hat{R}(z) = R_0(1-z^{-1})$$

For $a = 1$ cm and $T = 10^{-4}$ find R_0 such that $\hat{R}(-1) = Z_L(\infty) = R(-1)$.
(f) Sketch the frequency responses $Z_L(\Omega)$, $R(e^{j\Omega T})$, and $\hat{R}(e^{j\Omega T})$ as a function of Ω for $a = 1$ cm and $T = 10^{-4}$ for $0 \leqslant \Omega \leqslant \pi/T$.

3.14 A simple approximate model for a glottal pulse is given in Fig. P3.14a.
(a) Find the z-transform, $G_1(z)$, of the above sequence. (Hint: Note that $g_1(n)$ can be expressed as the convolution of the sequence

$$p(n) = 1 \quad 0 \leqslant n \leqslant N-1$$
$$= 0 \quad \textit{otherwise}$$

with itself).

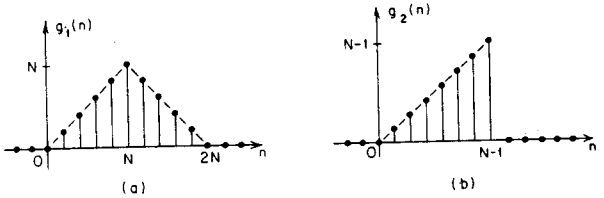


Fig. P3.14

(b) Plot the poles and zeros of $G_1(z)$ in the z-plane for $N = 10$.
(c) Sketch the magnitude of the Fourier transform of $g_1(n)$ as a function of ω.

Now consider the glottal pulse model $g_2(n)$ as given in Fig. P3.14b.
(d) Show that the z-transform, $G_2(z)$, is given by

$$G_2(z) = z^{-1}\sum_{n=0}^{N-2}(n+1)z^{-n}$$
$$= z^{-1}\left[\frac{1 - Nz^{-(N-1)} + (N-1)z^{-N}}{(1-z^{-1})^2}\right]$$

(Hint: Use the fact that the z-transform of $nx(n)$ is $-z\dfrac{dX(z)}{dz}$.)
(e) Show that in general $G_2(z)$ must have at least one zero outside the unit circle. Find the zeros of $G_2(z)$ for $N = 4$.

3.15 A commonly used approximation to the glottal pulse is

$$g(n) = na^n \quad n \geqslant 0$$
$$= 0 \quad n < 0$$

(a) Find the z-transform of $g(n)$.
(b) Sketch the Fourier transform, $G(e^{j\omega})$, as a function of ω.
(c) Show how a should be chosen so that

$$20\log_{10}|G(e^{j0})| - 20\log_{10}|G(e^{j\pi})| = 60 \text{ dB}.$$