Full text available at: http://dx.doi.org/10.1561/2000000001

Introduction to Digital Speech Processing

Introduction to Digital Speech Processing

Lawrence R. Rabiner

Rutgers University and University of California
Santa Barbara
USA

rabiner@ece.ucsb.edu

Ronald W. Schafer

Hewlett-Packard Laboratories
Palo Alto, CA
USA

now

the essence of knowledge

Boston – Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing, Foundations and Trends® in Signal Processing, vol 1, no 1–2, pp 1–194, 2007

ISBN: 978-1-60198-070-0

Foundations and Trends® in Signal Processing
Volume 1 Issue 1–2, 2007

Editorial Board

Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
- Classification and detection
- Estimation and regression
- Tree-structured methods

Information for Librarians
Foundations and Trends® in Signal Processing, 2007, Volume 1, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in
Signal Processing
Vol. 1, Nos. 1–2 (2007) 1–194

DOI: 10.1561/2000000001

now
the essence of knowledge

Introduction to Digital Speech Processing

Lawrence R. Rabiner[1] and Ronald W. Schafer[2]

[1] *Rutgers University and University of California, Santa Barbara, USA, rabiner@ece.ucsb.edu*
[2] *Hewlett-Packard Laboratories, Palo Alto, CA, USA*

Abstract

Since even before the time of Alexander Graham Bell's revolutionary invention, engineers and scientists have studied the phenomenon of speech communication with an eye on creating more efficient and effective systems of human-to-human and human-to-machine communication. Starting in the 1960s, digital signal processing (DSP), assumed a central role in speech studies, and today DSP is the key to realizing the fruits of the knowledge that has been gained through decades of research. Concomitant advances in integrated circuit technology and computer architecture have aligned to create a technological environment with virtually limitless opportunities for innovation in speech communication applications. In this text, we highlight the central role of DSP techniques in modern speech communication research and applications. We present a comprehensive overview of digital speech processing that ranges from the basic nature of the speech signal, through a variety of methods of representing speech in digital form, to applications in voice communication and automatic synthesis and recognition of speech. The breadth of this subject does not allow us to discuss any

aspect of speech processing to great depth; hence our goal is to provide a useful introduction to the wide range of important concepts that comprise the field of digital speech processing. A more comprehensive treatment will appear in the forthcoming book, *Theory and Application of Digital Speech Processing* [101].

Contents

1 Introduction **1**

1.1 The Speech Chain 2
1.2 Applications of Digital Speech Processing 7
1.3 Our Goal for this Text 14

2 The Speech Signal **17**

2.1 Phonetic Representation of Speech 17
2.2 Models for Speech Production 19
2.3 More Refined Models 23

3 Hearing and Auditory Perception **25**

3.1 The Human Ear 25
3.2 Perception of Loudness 27
3.3 Critical Bands 28
3.4 Pitch Perception 29
3.5 Auditory Masking 31
3.6 Complete Model of Auditory Processing 32

4 Short-Time Analysis of Speech **33**

4.1 Short-Time Energy and Zero-Crossing Rate 37
4.2 Short-Time Autocorrelation Function (STACF) 40
4.3 Short-Time Fourier Transform (STFT) 42
4.4 Sampling the STFT in Time and Frequency 44

4.5 The Speech Spectrogram 46
4.6 Relation of STFT to STACF 49
4.7 Short-Time Fourier Synthesis 51
4.8 Short-Time Analysis is Fundamental to our Thinking 53

5 Homomorphic Speech Analysis 55

5.1 Definition of the Cepstrum and Complex Cepstrum 55
5.2 The Short-Time Cepstrum 58
5.3 Computation of the Cepstrum 58
5.4 Short-Time Homomorphic Filtering of Speech 63
5.5 Application to Pitch Detection 65
5.6 Applications to Pattern Recognition 67
5.7 The Role of the Cepstrum 72

6 Linear Predictive Analysis 75

6.1 Linear Prediction and the Speech Model 75
6.2 Computing the Prediction Coefficients 79
6.3 The Levinson–Durbin Recursion 84
6.4 LPC Spectrum 87
6.5 Equivalent Representations 91
6.6 The Role of Linear Prediction 96

7 Digital Speech Coding 97

7.1 Sampling and Quantization of Speech (PCM) 97
7.2 Digital Speech Coding 105
7.3 Closed-Loop Coders 108
7.4 Open-Loop Coders 127
7.5 Frequency-Domain Coders 134
7.6 Evaluation of Coders 136

8 Text-to-Speech Synthesis Methods 139

8.1 Text Analysis 140
8.2 Evolution of Speech Synthesis Systems 145

8.3 Unit Selection Methods 152
8.4 TTS Applications 159
8.5 TTS Future Needs 160

9 Automatic Speech Recognition (ASR) 163

9.1 The Problem of Automatic Speech Recognition 163
9.2 Building a Speech Recognition System 165
9.3 The Decision Processes in ASR 168
9.4 Representative Recognition Performance 181
9.5 Challenges in ASR Technology 183

Conclusion 185

Acknowledgments 187

References 189

Supplemental References 197

1

Introduction

The fundamental purpose of speech is communication, i.e., the transmission of messages. According to Shannon's information theory [116], a message represented as a sequence of discrete symbols can be quantified by its *information content* in bits, and the rate of transmission of information is measured in bits/second (bps). In speech production, as well as in many human-engineered electronic communication systems, the information to be transmitted is encoded in the form of a continuously varying (analog) waveform that can be transmitted, recorded, manipulated, and ultimately decoded by a human listener. In the case of speech, the fundamental analog form of the message is an acoustic waveform, which we call the *speech signal*. Speech signals, as illustrated in Figure 1.1, can be converted to an electrical waveform by a microphone, further manipulated by both analog and digital signal processing, and then converted back to acoustic form by a loudspeaker, a telephone handset or headphone, as desired. This form of speech processing is, of course, the basis for Bell's telephone invention as well as today's multitude of devices for recording, transmitting, and manipulating speech and audio signals. Although Bell made his invention without knowing the fundamentals of information theory, these ideas

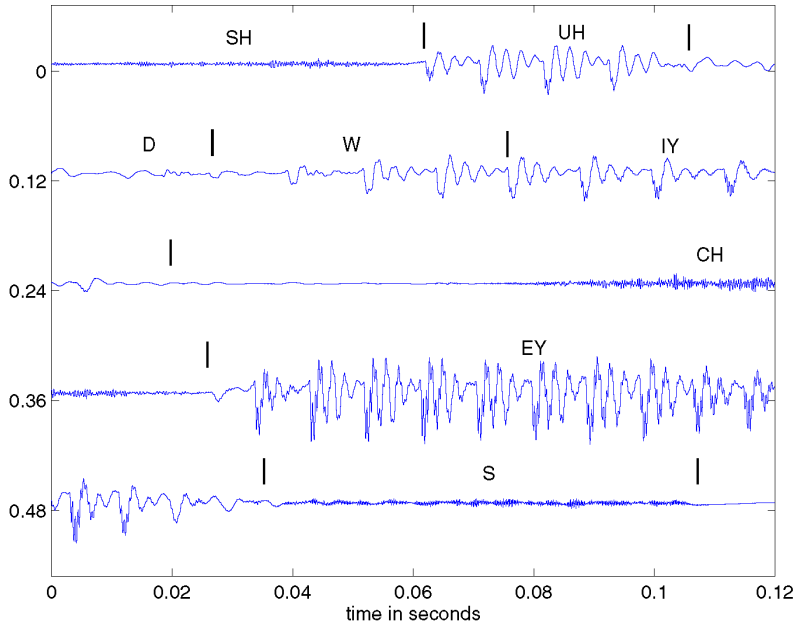

Fig. 1.1 A speech waveform with phonetic labels for the text message "Should we chase."

have assumed great importance in the design of sophisticated modern communications systems. Therefore, even though our main focus will be mostly on the speech waveform and its representation in the form of parametric models, it is nevertheless useful to begin with a discussion of how information is encoded in the speech waveform.

1.1 The Speech Chain

Figure 1.2 shows the complete process of producing and perceiving speech from the formulation of a message in the brain of a talker, to the creation of the speech signal, and finally to the understanding of the message by a listener. In their classic introduction to speech science, Denes and Pinson aptly referred to this process as the "speech chain" [29]. The process starts in the upper left as a message represented somehow in the brain of the speaker. The message information can be thought of as having a number of different representations during the process of speech production (the upper path in Figure 1.2).

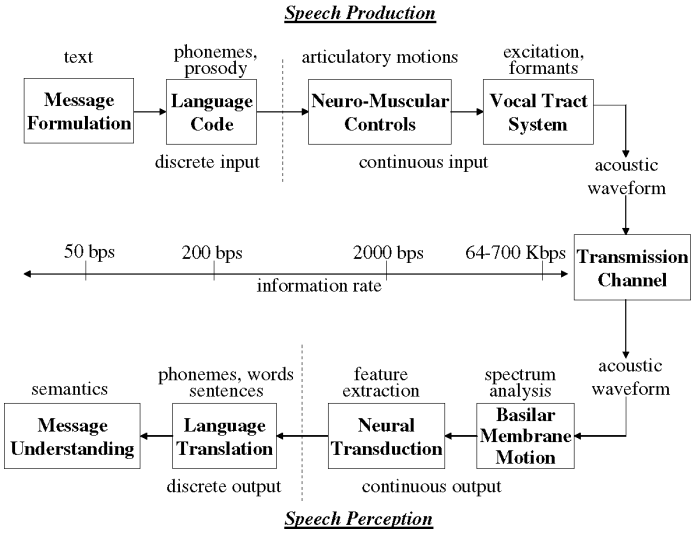

Fig. 1.2 The Speech Chain: from message, to speech signal, to understanding.

For example the message could be represented initially as English text. In order to "speak" the message, the talker implicitly converts the text into a symbolic representation of the sequence of sounds corresponding to the spoken version of the text. This step, called the language code generator in Figure 1.2, converts text symbols to phonetic symbols (along with stress and durational information) that describe the basic sounds of a spoken version of the message and the manner (i.e., the speed and emphasis) in which the sounds are intended to be produced. As an example, the segments of the waveform of Figure 1.1 are labeled with phonetic symbols using a computer-keyboard-friendly code called ARPAbet.[1] Thus, the text "should we chase" is represented phonetically (in ARPAbet symbols) as [SH UH D — W IY — CH EY S]. (See Chapter 2 for more discussion of phonetic transcription.) The third step in the speech production process is the conversion to "neuro-muscular controls," i.e., the set of control signals that direct the neuro-muscular system to move the speech articulators, namely the tongue, lips, teeth,

[1] The International Phonetic Association (IPA) provides a set of rules for phonetic transcription using an equivalent set of specialized symbols. The ARPAbet code does not require special fonts and is thus more convenient for computer applications.

jaw and velum, in a manner that is consistent with the sounds of the desired spoken message and with the desired degree of emphasis. The end result of the neuro-muscular controls step is a set of articulatory motions (continuous control) that cause the vocal tract articulators to move in a prescribed manner in order to create the desired sounds. Finally the last step in the Speech Production process is the "vocal tract system" that physically creates the necessary sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform, such as the one shown in Figure 1.1, that encodes the information in the desired message into the speech signal.

To determine the rate of information flow during speech production, assume that there are about 32 symbols (letters) in the language (in English there are 26 letters, but if we include simple punctuation we get a count closer to $32 = 2^5$ symbols). Furthermore, the rate of speaking for most people is about 10 symbols per second (somewhat on the high side, but still acceptable for a rough information rate estimate). Hence, assuming independent letters as a simple approximation, we estimate the base information rate of the text message as about 50 bps (5 bits per symbol times 10 symbols per second). At the second stage of the process, where the text representation is converted into phonemes and prosody (e.g., pitch and stress) markers, the information rate is estimated to increase by a factor of 4 to about 200 bps. For example, the ARBAbet phonetic symbol set used to label the speech sounds in Figure 1.1 contains approximately $64 = 2^6$ symbols, or about 6 bits/phoneme (again a rough approximation assuming independence of phonemes). In Figure 1.1, there are 8 phonemes in approximately 600 ms. This leads to an estimate of $8 \times 6/0.6 = 80$ bps. Additional information required to describe prosodic features of the signal (e.g., duration, pitch, loudness) could easily add 100 bps to the total information rate for a message encoded as a speech signal.

The information representations for the first two stages in the speech chain are discrete so we can readily estimate the rate of information flow with some simple assumptions. For the next stage in the speech production part of the speech chain, the representation becomes continuous (in the form of control signals for articulatory motion). If they could be measured, we could estimate the spectral bandwidth of these

control signals and appropriately sample and quantize these signals to obtain equivalent digital signals for which the data rate could be estimated. The articulators move relatively slowly compared to the time variation of the resulting acoustic waveform. Estimates of bandwidth and required accuracy suggest that the total data rate of the sampled articulatory control signals is about 2000 bps [34]. Thus, the original text message is represented by a set of continuously varying signals whose digital representation requires a much higher data rate than the information rate that we estimated for transmission of the message as a speech signal.[2] Finally, as we will see later, the data rate of the digitized speech waveform at the end of the speech production part of the speech chain can be anywhere from 64,000 to more than 700,000 bps. We arrive at such numbers by examining the sampling rate and quantization required to represent the speech signal with a desired perceptual fidelity. For example, "telephone quality" requires that a bandwidth of 0–4 kHz be preserved, implying a sampling rate of 8000 samples/s. Each sample can be quantized with 8 bits on a log scale, resulting in a bit rate of 64,000 bps. This representation is highly intelligible (i.e., humans can readily extract the message from it) but to most listeners, it will sound different from the original speech signal uttered by the talker. On the other hand, the speech waveform can be represented with "CD quality" using a sampling rate of 44,100 samples/s with 16 bit samples, or a data rate of 705,600 bps. In this case, the reproduced acoustic signal will be virtually indistinguishable from the original speech signal.

As we move from text to speech waveform through the speech chain, the result is an encoding of the message that can be effectively transmitted by acoustic wave propagation and robustly decoded by the hearing mechanism of a listener. The above analysis of data rates shows that as we move from text to sampled speech waveform, the data rate can increase by a factor of 10,000. Part of this extra information represents characteristics of the talker such as emotional state, speech mannerisms, accent, etc., but much of it is due to the inefficiency

[2] Note that we introduce the term data rate for digital representations to distinguish from the inherent information content of the message represented by the speech signal.

of simply sampling and finely quantizing analog signals. Thus, motivated by an awareness of the low intrinsic information rate of speech, a central theme of much of digital speech processing is to obtain a digital representation with lower data rate than that of the sampled waveform.

The complete speech chain consists of a speech production/generation model, of the type discussed above, as well as a speech perception/recognition model, as shown progressing to the left in the bottom half of Figure 1.2. The speech perception model shows the series of steps from capturing speech at the ear to understanding the message encoded in the speech signal. The first step is the effective conversion of the acoustic waveform to a spectral representation. This is done within the inner ear by the basilar membrane, which acts as a non-uniform spectrum analyzer by spatially separating the spectral components of the incoming speech signal and thereby analyzing them by what amounts to a non-uniform filter bank. The next step in the speech perception process is a neural transduction of the spectral features into a set of sound features (or distinctive features as they are referred to in the area of linguistics) that can be decoded and processed by the brain. The next step in the process is a conversion of the sound features into the set of phonemes, words, and sentences associated with the in-coming message by a language translation process in the human brain. Finally, the last step in the speech perception model is the conversion of the phonemes, words and sentences of the message into an understanding of the meaning of the basic message in order to be able to respond to or take some appropriate action. Our fundamental understanding of the processes in most of the speech perception modules in Figure 1.2 is rudimentary at best, but it is generally agreed that some physical correlate of each of the steps in the speech perception model occur within the human brain, and thus the entire model is useful for thinking about the processes that occur.

There is one additional process shown in the diagram of the complete speech chain in Figure 1.2 that we have not discussed — namely the transmission channel between the speech generation and speech perception parts of the model. In its simplest embodiment, this transmission channel consists of just the acoustic wave connection between

a speaker and a listener who are in a common space. It is essential to include this transmission channel in our model for the speech chain since it includes real world noise and channel distortions that make speech and message understanding more difficult in real communication environments. More interestingly for our purpose here — it is in this domain that we find the applications of digital speech processing.

1.2 Applications of Digital Speech Processing

The first step in most applications of digital speech processing is to convert the acoustic waveform to a sequence of numbers. Most modern A-to-D converters operate by sampling at a very high rate, applying a digital lowpass filter with cutoff set to preserve a prescribed bandwidth, and then reducing the sampling rate to the desired sampling rate, which can be as low as twice the cutoff frequency of the sharp-cutoff digital filter. This discrete-time representation is the starting point for most applications. From this point, other representations are obtained by digital processing. For the most part, these alternative representations are based on incorporating knowledge about the workings of the speech chain as depicted in Figure 1.2. As we will see, it is possible to incorporate aspects of both the speech production and speech perception process into the digital representation and processing. It is not an oversimplification to assert that digital speech processing is grounded in a set of techniques that have the goal of pushing the data rate of the speech representation to the left along either the upper or lower path in Figure 1.2.

The remainder of this chapter is devoted to a brief summary of the applications of digital speech processing, i.e., the systems that people interact with daily. Our discussion will confirm the importance of the digital representation in all application areas.

1.2.1 Speech Coding

Perhaps the most widespread applications of digital speech processing technology occur in the areas of digital transmission and storage

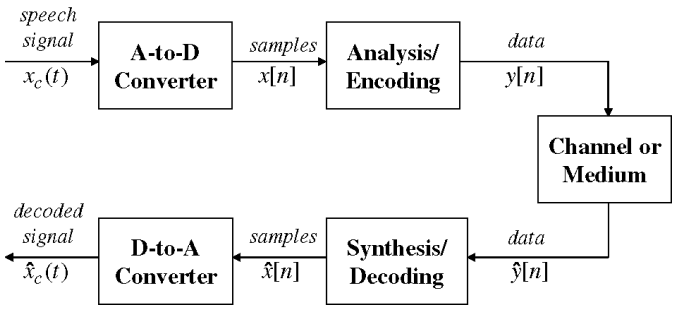

Fig. 1.3 Speech coding block diagram — encoder and decoder.

of speech signals. In these areas the centrality of the digital representation is obvious, since the goal is to compress the digital waveform representation of speech into a lower bit-rate representation. It is common to refer to this activity as "speech coding" or "speech compression."

Figure 1.3 shows a block diagram of a generic speech encoding/decoding (or compression) system. In the upper part of the figure, the A-to-D converter converts the analog speech signal $x_c(t)$ to a sampled waveform representation $x[n]$. The digital signal $x[n]$ is analyzed and coded by digital computation algorithms to produce a new digital signal $y[n]$ that can be transmitted over a digital communication channel or stored in a digital storage medium as $\hat{y}[n]$. As we will see, there are a myriad of ways to do the encoding so as to reduce the data rate over that of the sampled and quantized speech waveform $x[n]$. Because the digital representation at this point is often not directly related to the sampled speech waveform, $y[n]$ and $\hat{y}[n]$ are appropriately referred to as *data signals* that represent the speech signal. The lower path in Figure 1.3 shows the decoder associated with the speech coder. The received data signal $\hat{y}[n]$ is decoded using the inverse of the analysis processing, giving the sequence of samples $\hat{x}[n]$ which is then converted (using a D-to-A Converter) back to an analog signal $\hat{x}_c(t)$ for human listening. The decoder is often called a *synthesizer* because it must reconstitute the speech waveform from data that may bear no direct relationship to the waveform.

With carefully designed error protection coding of the digital representation, the transmitted ($y[n]$) and received ($\hat{y}[n]$) data can be essentially identical. This is the quintessential feature of digital coding. In theory, perfect transmission of the coded digital representation is possible even under very noisy channel conditions, and in the case of digital storage, it is possible to store a perfect copy of the digital representation in perpetuity if sufficient care is taken to update the storage medium as storage technology advances. This means that the speech signal can be reconstructed to within the accuracy of the original coding for as long as the digital representation is retained. In either case, the goal of the speech coder is to start with samples of the speech signal and reduce (compress) the data rate required to represent the speech signal while maintaining a desired perceptual fidelity. The compressed representation can be more efficiently transmitted or stored, or the bits saved can be devoted to error protection.

Speech coders enable a broad range of applications including narrowband and broadband wired telephony, cellular communications, voice over internet protocol (VoIP) (which utilizes the internet as a real-time communications medium), secure voice for privacy and encryption (for national security applications), extremely narrowband communications channels (such as battlefield applications using high frequency (HF) radio), and for storage of speech for telephone answering machines, interactive voice response (IVR) systems, and prerecorded messages. Speech coders often utilize many aspects of both the speech production and speech perception processes, and hence may not be useful for more general audio signals such as music. Coders that are based on incorporating only aspects of sound perception generally do not achieve as much compression as those based on speech production, but they are more general and can be used for all types of audio signals. These coders are widely deployed in MP3 and AAC players and for audio in digital television systems [120].

1.2.2 Text-to-Speech Synthesis

For many years, scientists and engineers have studied the speech production process with the goal of building a system that can start with

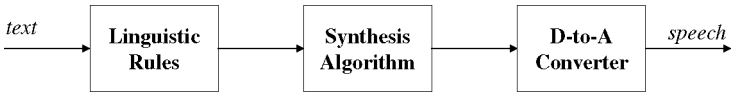

Fig. 1.4 Text-to-speech synthesis system block diagram.

text and produce speech automatically. In a sense, a text-to-speech synthesizer such as depicted in Figure 1.4 is a digital simulation of the entire upper part of the speech chain diagram. The input to the system is ordinary text such as an email message or an article from a newspaper or magazine. The first block in the text-to-speech synthesis system, labeled linguistic rules, has the job of converting the printed text input into a set of sounds that the machine must synthesize. The conversion from text to sounds involves a set of linguistic rules that must determine the appropriate set of sounds (perhaps including things like emphasis, pauses, rates of speaking, etc.) so that the resulting synthetic speech will express the words and intent of the text message in what passes for a natural voice that can be decoded accurately by human speech perception. This is more difficult than simply looking up the words in a pronouncing dictionary because the linguistic rules must determine how to pronounce acronyms, how to pronounce ambiguous words like *read*, *bass*, *object*, how to pronounce abbreviations like St. (street or Saint), Dr. (Doctor or drive), and how to properly pronounce proper names, specialized terms, etc. Once the proper pronunciation of the text has been determined, the role of the synthesis algorithm is to create the appropriate sound sequence to represent the text message in the form of speech. In essence, the synthesis algorithm must simulate the action of the vocal tract system in creating the sounds of speech. There are many procedures for assembling the speech sounds and compiling them into a proper sentence, but the most promising one today is called "unit selection and concatenation." In this method, the computer stores multiple versions of each of the basic units of speech (phones, half phones, syllables, etc.), and then decides which sequence of speech units sounds best for the particular text message that is being produced. The basic digital representation is not generally the sampled speech wave. Instead, some sort of compressed representation is normally used to

save memory and, more importantly, to allow convenient manipulation of durations and blending of adjacent sounds. Thus, the speech synthesis algorithm would include an appropriate decoder, as discussed in Section 1.2.1, whose output is converted to an analog representation via the D-to-A converter.

Text-to-speech synthesis systems are an essential component of modern human–machine communications systems and are used to do things like read email messages over a telephone, provide voice output from GPS systems in automobiles, provide the voices for talking agents for completion of transactions over the internet, handle call center help desks and customer care applications, serve as the voice for providing information from handheld devices such as foreign language phrasebooks, dictionaries, crossword puzzle helpers, and as the voice of announcement machines that provide information such as stock quotes, airline schedules, updates on arrivals and departures of flights, etc. Another important application is in reading machines for the blind, where an optical character recognition system provides the text input to a speech synthesis system.

1.2.3 Speech Recognition and Other Pattern Matching Problems

Another large class of digital speech processing applications is concerned with the automatic extraction of information from the speech signal. Most such systems involve some sort of pattern matching. Figure 1.5 shows a block diagram of a generic approach to pattern matching problems in speech processing. Such problems include the following: speech recognition, where the object is to extract the message from the speech signal; speaker recognition, where the goal is to identify who is speaking; speaker verification, where the goal is to verify a speaker's claimed identity from analysis of their speech

Fig. 1.5 Block diagram of general pattern matching system for speech signals.

signal; word spotting, which involves monitoring a speech signal for the occurrence of specified words or phrases; and automatic indexing of speech recordings based on recognition (or spotting) of spoken keywords.

The first block in the pattern matching system converts the analog speech waveform to digital form using an A-to-D converter. The feature analysis module converts the sampled speech signal to a set of feature vectors. Often, the same analysis techniques that are used in speech coding are also used to derive the feature vectors. The final block in the system, namely the pattern matching block, dynamically time aligns the set of feature vectors representing the speech signal with a concatenated set of stored patterns, and chooses the identity associated with the pattern which is the closest match to the time-aligned set of feature vectors of the speech signal. The symbolic output consists of a set of recognized words, in the case of speech recognition, or the identity of the best matching talker, in the case of speaker recognition, or a decision as to whether to accept or reject the identity claim of a speaker in the case of speaker verification.

Although the block diagram of Figure 1.5 represents a wide range of speech pattern matching problems, the biggest use has been in the area of recognition and understanding of speech in support of human–machine communication by voice. The major areas where such a system finds applications include command and control of computer software, voice dictation to create letters, memos, and other documents, natural language voice dialogues with machines to enable help desks and call centers, and for agent services such as calendar entry and update, address list modification and entry, etc.

Pattern recognition applications often occur in conjunction with other digital speech processing applications. For example, one of the preeminent uses of speech technology is in portable communication devices. Speech coding at bit rates on the order of 8 Kbps enables normal voice conversations in cell phones. Spoken name speech recognition in cellphones enables voice dialing capability that can automatically dial the number associated with the recognized name. Names from directories with upwards of several hundred names can readily be recognized and dialed using simple speech recognition technology.

Another major speech application that has long been a dream of speech researchers is *automatic language translation.* The goal of language translation systems is to convert spoken words in one language to spoken words in another language so as to facilitate natural language voice dialogues between people speaking different languages. Language translation technology requires speech synthesis systems that work in both languages, along with speech recognition (and generally natural language understanding) that also works for both languages; hence it is a very difficult task and one for which only limited progress has been made. When such systems exist, it will be possible for people speaking different languages to communicate at data rates on the order of that of printed text reading!

1.2.4 Other Speech Applications

The range of speech communication applications is illustrated in Figure 1.6. As seen in this figure, the techniques of digital speech processing are a key ingredient of a wide range of applications that include the three areas of transmission/storage, speech synthesis, and speech recognition as well as many others such as speaker identification, speech signal quality enhancement, and aids for the hearing- or visually-impaired.

The block diagram in Figure 1.7 represents any system where time signals such as speech are processed by the techniques of DSP. This figure simply depicts the notion that once the speech signal is sampled, it can be manipulated in virtually limitless ways by DSP techniques. Here again, manipulations and modifications of the speech signal are

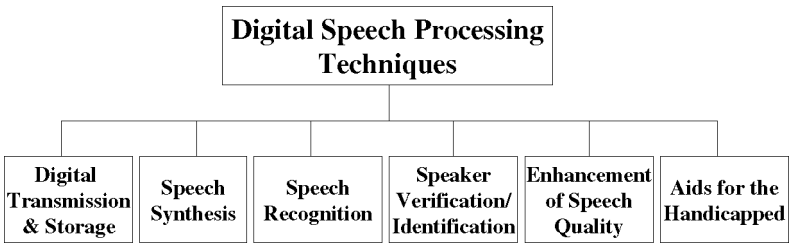

Fig. 1.6 Range of speech communication applications.

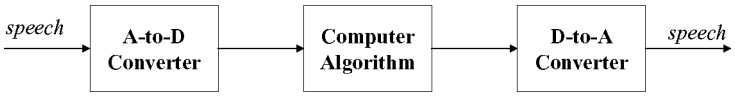

Fig. 1.7 General block diagram for application of digital signal processing to speech signals.

usually achieved by transforming the speech signal into an alternative representation (that is motivated by our understanding of speech production and speech perception), operating on that representation by further digital computation, and then transforming back to the waveform domain, using a D-to-A converter.

One important application area is *speech enhancement*, where the goal is to remove or suppress noise or echo or reverberation picked up by a microphone along with the desired speech signal. In human-to-human communication, the goal of speech enhancement systems is to make the speech more intelligible and more natural; however, in reality the best that has been achieved so far is less perceptually annoying speech that essentially maintains, but does not improve, the intelligibility of the noisy speech. Success *has* been achieved, however, in making distorted speech signals more useful for further processing as part of a speech coder, synthesizer, or recognizer. An excellent reference in this area is the recent textbook by Loizou [72].

Other examples of manipulation of the speech signal include timescale modification to align voices with video segments, to modify voice qualities, and to speed-up or slow-down prerecorded speech (e.g., for talking books, rapid review of voice mail messages, or careful scrutinizing of spoken material).

1.3 Our Goal for this Text

We have discussed the speech signal and how it encodes information for human communication. We have given a brief overview of the way in which digital speech processing is being applied today, and we have hinted at some of the possibilities that exist for the future. These and many more examples all rely on the basic principles of digital speech processing, which we will discuss in the remainder of this text. We make no pretense of exhaustive coverage. The subject is too broad and

too deep. Our goal is only to provide an up-to-date introduction to this fascinating field. We will not be able to go into great depth, and we will not be able to cover all the possible applications of digital speech processing techniques. Instead our focus is on the fundamentals of digital speech processing and their application to coding, synthesis, and recognition. This means that some of the latest algorithmic innovations and applications will not be discussed — not because they are not interesting, but simply because there are so many fundamental tried-and-true techniques that remain at the core of digital speech processing. We hope that this text will stimulate readers to investigate the subject in greater depth using the extensive set of references provided.

References

[1] J. B. Allen and L. R. Rabiner, "A unified theory of short-time spectrum analysis and synthesis," *Proceedings of IEEE*, vol. 65, no. 11, pp. 1558–1564, November 1977.

[2] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Transactions on Communications*, vol. COM-30, no. 4, pp. 600–614, April 1982.

[3] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, pp. 561–580, 1971.

[4] B. S. Atal and J. Remde, "A new model of LPC exitation for producing natural-sounding speech at low bit rates," *Proceedings of IEEE ICASSP*, pp. 614–617, 1982.

[5] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell System Technical Journal*, vol. 49, pp. 1973–1986, October 1970.

[6] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criterion," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, pp. 247–254, June 1979.

[7] B. S. Atal and M. R. Schroeder, "Improved quantizer for adaptive predictive coding of speech signals at low bit rates," *Proceedings of ICASSP*, pp. 535–538, April 1980.

[8] T. B. Barnwell III, "Recursive windowing for generating autocorrelation analysis for LPC analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 5, pp. 1062–1066, October 1981.

[9] T. B. Barnwell III, K. Nayebi, and C. H. Richardson, *Speech Coding, A Computer Laboratory Textbook*. John Wiley and Sons, 1996.

[10] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[11] L. E. Baum, T. Petri, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals in Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[12] W. R. Bennett, "Spectra of quantized signals," *Bell System Technical Journal*, vol. 27, pp. 446–472, July 1948.

[13] M. Berouti, H. Garten, P. Kabal, and P. Mermelstein, "Efficient computation and encoding of the multipulse excitation for LPC," *Proceedings of ICASSP*, pp. 384–387, March 1984.

[14] M. Beutnagel, A. Conkie, and A. K. Syrdal, "Diphone synthesis using unit selection," Third Speech Synthesis Workshop, Jenolan Caes, Australia, November 1998.

[15] M. Beutnatel and A. Conkie, "Interaction of units in a unit selection database," *Proceedings of Eurospeech '99,* Budapest, Hungary, September 1999.

[16] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of times series for echos: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, (M. Rosenblatt, ed.), New York: John Wiley and Sons, Inc., 1963.

[17] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchornized and noise-robust audio recordings during realtime MRI scans," *Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, October 2006.

[18] C. S. Burrus and R. A. Gopinath, *Introduction to Wavelets and Wavelet Transforms*. Prentice-Hall Inc., 1998.

[19] J. P. Campbell Jr., V. C. Welch, and T. E. Tremain, "An expandable error-pretected 4800 bps CELP coder," *Proceedings of ICASSP*, vol. 2, pp. 735–738, May 1989.

[20] F. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveform concatenation," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 2015–2018, 1986.

[21] J. H. Chung and R. W. Schafer, "Performance evaluation of analysis-by-synthesis homomorphic vocoders," *Proceedings of IEEE ICASSP*, vol. 2, pp. 117–120, March 1992.

[22] C. H. Coker, "A model of articulatory dynamics and control," *Proceedings of IEEE*, vol. 64, pp. 452–459, 1976.

[23] R. V. Cox, S. L. Gay, Y. Shoham, S. Quackenbush, N. Seshadri, and N. Jayant, "New directions in subband coding," *IEEE Journal of Selected Areas in Communications*, vol. 6, no. 2, pp. 391–409, February 1988.

[24] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Prentice-Hall Inc., 1983.

[25] R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital coding of speech in subbands," *Bell System Technical Journal*, vol. 55, no. 8, pp. 1069–1085, October 1976.

[26] C. C. Cutler, "Differential quantization of communication signals," U.S. Patent 2,605,361, July 29, 1952.

[27] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, August 1980.

[28] F. deJager, "Delta modulation — a new method of PCM transmission using the 1-unit code," *Philips Research Reports*, pp. 442–466, December 1952.

[29] P. B. Denes and E. N. Pinson, *The speech chain*. W. H. Freeman Company, 2nd Edition, 1993.

[30] H. Dudley, "The vocoder," *Bell Labs Record*, vol. 17, pp. 122–126, 1939.

[31] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Netherlands: Kluwer Academic Publishers, 1997.

[32] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton & Co., 1960; Walter de Gruyter, 1970.

[33] J. D. Ferguson, "Hidden Markov Analysis: An Introduction," *Hidden Markov Models for Speech*, Princeton: Institute for Defense Analyses, 1980.

[34] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Springer-Verlag, 1972.

[35] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum*, vol. 7, pp. 22–45, October 1970.

[36] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell System Technical Journal*, vol. 54, no. 3, pp. 485–506, March 1975.

[37] H. Fletcher and W. J. Munson, "Loudness, its definition, measurement and calculation," *Journal of Acoustical Society of America*, vol. 5, no. 2, pp. 82–108, October 1933.

[38] G. D. Forney, "The Viterbi algorithm," *IEEE Proceedings*, vol. 61, pp. 268–278, March 1973.

[39] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics Speech, and Signal Processing*, vol. ASSP-29, no. 2, pp. 254–272, April 1981.

[40] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP-26, no. 1, pp. 52–59, February 1986.

[41] O. Ghitza, "Audiotry nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, (S. Furui and M. Sondhi, eds.), pp. 453–485, NY: Marcel Dekker, 1991.

[42] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech corpus for research and development," *Proceedings of ICASSP 1992*, pp. 517–520, 1992.

[43] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch period of speech in the time domain," *Journal of Acoustical Society of America*, vol. 46, no. 2, pt. 2, pp. 442–448, August 1969.

[44] A. L. Gorin, B. A. Parker, R. M. Sachs, and J. G. Wilpon, "How may I help you?," *Proceedings of the Interactive Voice Technology for Telecommunications Applications (IVTTA)*, pp. 57–60, 1996.

[45] R. M. Gray, "Vector quantization," *IEEE Signal Processing Magazine*, pp. 4–28, April 1984.

[46] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[47] J. A. Greefkes and K. Riemens, "Code modulation with digitally controlled companding for speech transmission," *Philips Technical Review*, pp. 335–353, 1970.

[48] H. Hermansky, "Auditory modeling in automatic recognition of speech," in *Proceedings of First European Conference on Signal Analysis and Prediction*, pp. 17–21, Prague, Czech Republic, 1997.

[49] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice-Hall Inc., 2001.

[50] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proceedings of ICASSP-96,* Atlanta, vol. 1, pp. 373–376, 1996.

[51] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[52] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of Acoustical Society of America*, vol. 57, pp. 535(a), p. s35(A).

[53] F. Itakura and S. Saito, "Analysis-synthesis telephony based upon the maximum likelihood method," *Proceedings of 6th International of Congress on Acoustics*, pp. C17–C20, 1968.

[54] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electronics and Communications in Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.

[55] F. Itakura and T. Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum," in *Proceedings of ICASSP87*, pp. 1257–1260, Dallas TX, April 1987.

[56] N. S. Jayant, "Adaptive delta modulation with a one-bit memory," *Bell System Technical Journal*, pp. 321–342, March 1970.

[57] N. S. Jayant, "Adaptive quantization with one word memory," *Bell System Technical Journal*, pp. 1119–1144, September 1973.

[58] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Prentice-Hall, 1984.

[59] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge: MIT Press, 1997.

[60] F. Jelinek, R. L. Mercer, and S. Roucos, "Principles of lexical language modeling for speech recognition," in *Advances in Speech Signal Processing*, (S. Furui and M. M. Sondhi, eds.), pp. 651–699, Marcel Dekker, 1991.

[61] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT&T Technology Journal*, vol. 64, no. 6, pp. 1235–1249, 1985.

[62] B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Transactions in Information Theory*, vol. 32, no. 2, pp. 307–309, 1986.

[63] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speehc recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-35, no. 7, pp. 947–954, July 1987.

[64] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, 1980.

[65] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of the Acoustical Society of America*, vol. 82, pp. 737–793, September 1987.

[66] W. Koenig, H. K. Dunn, and L. Y. Lacey, "The sound spectrograph," *Journal of the Acoustical Society of America*, vol. 18, pp. 19–49, 1946.

[67] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation: A nove approach to effective and efficient multipulse coding of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 1054–1063, October 1986.

[68] R. G. Leonard, "A database for speaker-independent digit recognition," *Proceedings of ICASSP 1984*, pp. 42.11.1–42.11.4, 1984.

[69] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[70] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. COM-28, pp. 84–95, January 1980.

[71] S. P. Lloyd, "Least square quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, March 1982.

[72] P. C. Loizou, *Speech Enhancement, Theory and Practice*. CRC Press, 2007.

[73] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France, May 1982.

[74] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of IEEE*, vol. 63, pp. 561–580, 1975.

[75] J. Makhoul, V. Viswanathan, R. Schwarz, and A. W. F. Huggins, "A mixed source model for speech compression and synthesis," *Journal of the Acoustical Society of America*, vol. 64, pp. 1577–1581, December 1978.

[76] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time-scaling of pitch signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 121–133, 1979.

[77] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-20, no. 5, pp. 367–377, December 1972.

[78] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[79] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. IT-6, pp. 7–12, March 1960.
[80] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.
[81] M. Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, pp. 269–312, 1997.
[82] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5–6, 1990.
[83] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, February 1967.
[84] P. Noll, "A comparative study of various schemes for speech encoding," *Bell System Technical Journal*, vol. 54, no. 9, pp. 1597–1614, November 1975.
[85] A. V. Oppenehim, "Superposition in a class of nonlinear systems," PhD dissertation, MIT, 1964. Also: MIT Research Lab. of Electronics, Cambridge, Massachusetts, Technical Report No. 432, 1965.
[86] A. V. Oppenheim, "A speech analysis-synthesis system based on homomorphic filtering," *Journal of the Acoustical Society of America*, vol. 45, no. 2, pp. 293–309, February 1969.
[87] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, pp. 57–62, August 1970.
[88] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-16, pp. 221–228, June 1968.
[89] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice-Hall Inc., 1999.
[90] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham Jr., "Nonlinear filtering of multiplied and convolved signals," *Proceedings of IEEE*, vol. 56, no. 8, pp. 1264–1291, August 1968.
[91] M. D. Paez and T. H. Glisson, "Minimum mean-squared error quantization in speech," *IEEE Transactions on Communications*, vol. Com-20, pp. 225–230, April 1972.
[92] D. S. Pallett et al., "The 1994 benchmark tests for the ARPA spoken language program," *Proceedings of 1995 ARPA Human Language Technology Workshop*, pp. 5–36, 1995.
[93] M. R. Portnoff, "A quasi-one-dimensional simulation for the time-varying vocal tract," MS Thesis, MIT, Department of Electrical Engineering, 1973.
[94] T. F. Quatieri, *Discrete-time speech signal processing*. Prentice Hall, 2002.
[95] L. R. Rabiner, "A model for synthesizing speech by rule," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-17, no. 1, pp. 7–13, March 1969.
[96] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, no. 1, pp. 24–33, February 1977.
[97] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *IEEE Proceedings*, vol. 77, no. 2, pp. 257–286, 1989.

[98] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE Signal Processing Magazine*, 1985.

[99] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall Inc., 1993.

[100] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, February 1975.

[101] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Speech Processing*. Prentice-Hall Inc., 2009. (In preparation).

[102] L. R. Rabiner, R. W. Schafer, and J. L. Flanagan, "Computer synthesis of speech by concatenation of formant coded words," *Bell System Technical Journal*, vol. 50, no. 5, pp. 1541–1558, May–June 1971.

[103] D. W. Robinson and R. S. Dadson, "A re-determination of the equal-loudness contours for pure tones," *British Journal of Applied Physics*, vol. 7, pp. 166–181, 1956.

[104] R. C. Rose and T. P. Barnwell III, "The self excited vocoder — an alternate approach to toll quality at 4800 bps," *Proceedings of ICASSP '86*, vol. 11, pp. 453–456, April 1986.

[105] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 822–828, February 1971.

[106] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *IEEE Proceedings*, vol. 88, no. 8, pp. 1270–1278, 2000.

[107] M. B. Sachs, C. C. Blackburn, and E. D. Young, "Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus," *Journal of Phonetics*, vol. 16, pp. 37–53, 1988.

[108] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 679–682, 1988.

[109] R. W. Schafer, "Echo removal by discrete generalized linear filtering," PhD dissertation, MIT, 1968. Also: MIT Research Laboratory of Electronics, Cambridge, Massachusetts, Technical Report No. 466, 1969.

[110] R. W. Schafer, "Homomorphic systems and cepstrum analysis of speech," *Springer Handbook of Speech Processing and Communication,* Springer, 2007.

[111] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *Journal of the Acoustical Society of America*, vol. 47, no. 2, pp. 458–465, February 1970.

[112] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," *Proceedings of IEEE ICASSP*, pp. 937–940, 1985.

[113] M. R. Schroeder and E. E. David, "A vocoder for transmitting 10 kc/s speech over a 3.5 kc/s channel," *Acustica*, vol. 10, pp. 35–43, 1960.

[114] J. H. Schroeter, "Basic principles of speech synthesis," *Springer Handbook of Speech Processing,* Springer-Verlag, 2006.

[115] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[116] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.
[117] G. A. Sitton, C. S. Burrus, J. W. Fox, and S. Treitel, "Factoring very-high-degree polynomials," *IEEE Signal Processing Magazine*, vol. 20, no. 6, pp. 27–42, November 2003.
[118] B. Smith, "Instantaneous companding of quantized signals," *Bell System Technical Journal*, vol. 36, no. 3, pp. 653–709, May 1957.
[119] F. K. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 15–24, January 1993.
[120] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. Wiley Interscience, 2007.
[121] K. N. Stevens, *Acoustic Phonetics*. MIT Press, 1998.
[122] S. S. Stevens and J. Volkman, "The relation of pitch to frequency," *American Journal of Psychology*, vol. 53, p. 329, 1940.
[123] L. C. Stewart, R. M. Gray, and Y. Linde, "The design of trellis waveform coders," *IEEE transactions on Communications*, vol. COM-30, pp. 702–710, April 1982.
[124] T. G. Stockham Jr., T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proceedings of IEEE*, vol. 63, pp. 678–692, April 1975.
[125] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press, 1996.
[126] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 1414–1422, October 1987.
[127] J. M. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustical, Speech, and Signal Processing*, vol. ASSP-25, no. 2, pp. 170–177, April 1977.
[128] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbits/s," *IEEE Transactions on Communications*, vol. COM-23, no. 12, pp. 1466–1474, December 1975.
[129] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Prentice-Hall Inc., 1993.
[130] R. Viswanathan, W. Russell, and J. Makhoul, "Voice-excited LPC coders for 9.6 kbps speech transmission," vol. 4, pp. 558–561, April 1979.
[131] V. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-23, no. 3, pp. 309–321, June 1975.
[132] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding Aalgorithm," *IEEE Transactions on Information Theory*, vol. IT-13, pp. 260–269, April 1967.
[133] W. Ward, "Evaluation of the CMU ATIS system," *Proceedings of DARPA Speech and Natural Language Workshop*, pp. 101–105, February 1991.
[134] E. Zwicker and H. Fastl, *Psycho-acoustics*. Springer-Verlag, 2nd Edition, 1990.

Supplemental References

The specific references that comprise the Bibliography of this text are representative of the literature of the field of digital speech processing. In addition, we provide the following list of journals and books as a guide for further study. Listing the books in chronological order of publication provides some perspective on the evolution of the field.

Speech Processing Journals

- *IEEE Transactions on Signal Processing.* Main publication of IEEE Signal Processing Society.
- *IEEE Transactions on Speech and Audio.* Publication of IEEE Signal Processing Society that is focused on speech and audio processing.
- *Journal of the Acoustical Society of America.* General publication of the American Institute of Physics. Papers on speech and hearing as well as other areas of acoustics.
- *Speech Communication.* Published by Elsevier. A publication of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA).

General Speech Processing References

- *Speech Analysis, Synthesis and Perception*, J. L. Flanagan, Springer-Verlag, Second Edition, Berlin, 1972.
- *Linear Prediction of Speech*, J. D. Markel and A. H. Gray, Jr., Springer Verlag, Berlin, 1976.
- *Digital Processing of Speech Signals*, L. R. Rabiner and R. W. Schafer, Prentice-Hall Inc., 1978.
- *Speech Analysis*, R. W. Schafer and J. D. Markel (eds.), IEEE Press Selected Reprint Series, 1979.
- *Speech Communication, Human and Machine*, D. O'Shaughnessy, Addison-Wesley, 1987.
- *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Marcel Dekker Inc., New York, 1991.
- *Discrete-Time Processing of Speech Signals*, J. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, Wiley-IEEE Press, Classic Reissue, 1999.
- *Acoustic Phonetics*, K. N. Stevens, MIT Press, 1998.
- *Speech and Audio Signal Processing*, B. Gold and N. Morgan, John Wiley and Sons, 2000.
- *Digital Speech Processing, Synthesis and Recognition*, S. Furui, Second Edition, Marcel Dekker Inc., New York, 2001.
- *Discrete-Time Speech Signal Processing*, T. F. Quatieri, Prentice Hall Inc., 2002.
- *Speech Processing, A Dynamic and Optimization-Oriented Approach*, L. Deng and D. O'Shaughnessy, Marcel Dekker, 2003.
- *Springer Handbook of Speech Processing and Speech Communication*, J. Benesty, M. M. Sondhi and Y Huang (eds.), Springer, 2008.
- *Theory and Application of Digital Speech Processing*, L. R. Rabiner and R. W. Schafer, Prentice Hall Inc., 2009.

Speech Coding References

- *Digital Coding of Waveforms*, N. S. Jayant and P. Noll, Prentice Hall Inc., 1984.
- *Practical Approaches to Speech Coding*, P. E. Papamichalis, Prentice Hall Inc., 1987.
- *Vector Quantization and Signal Compression*, A. Gersho and R. M. Gray, Kluwer Academic Publishers, 1992.
- *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Elsevier, 1995.
- *Speech Coding, A Computer Laboratory Textbook*, T. P. Barnwell and K. Nayebi, John Wiley and Sons, 1996.

- *A Practical Handbook of Speech Coders*, R. Goldberg and L. Riek, CRC Press, 2000.
- *Speech Coding Algorithms*, W. C. Chu, John Wiley and Sons, 2003.
- *Digital Speech: Coding for Low Bit Rate Communication Systems*, Second Edition, A. M. Kondoz, John Wiley and Sons, 2004.

Speech Synthesis

- *From Text to Speech*, J. Allen, S. Hunnicutt and D. Klatt, Cambridge University Press, 1987.
- *Acoustics of American English*, J. P. Olive, A. Greenwood and J. Coleman, Springer-Verlag, 1993.
- *Computing Prosody*, Y. Sagisaka, N. Campbell and N. Higuchi, Springer-Verlag, 1996.
- *Progress in Speech Synthesis*, J. VanSanten, R. W. Sproat, J. P. Olive and J. Hirschberg (eds.), Springer-Verlag, 1996.
- *An Introduction to Text-to-Speech Synthesis*, T. Dutoit, Kluwer Academic Publishers, 1997.
- *Speech Processing and Synthesis Toolboxes*, D. Childers, John Wiley and Sons, 1999.
- *Text To Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan (eds.), Prentice Hall Inc., 2004.
- *Text-to-Speech Synthesis*, P. Taylor, Cambridge University Press, 2008.

Speech Recognition and Natural Language Processing

- *Fundamentals of Speech Recognition*, L. R. Rabiner and B. H. Juang, Prentice Hall Inc., 1993.
- *Connectionist Speech Recognition-A Hybrid Approach*, H. A. Bourlard and N. Morgan, Kluwer Academic Publishers, 1994.
- *Automatic Speech and Speaker Recognition*, C. H. Lee, F. K. Soong and K. K. Paliwal (eds.), Kluwer Academic Publisher, 1996.
- *Statistical Methods for Speech Recognition*, F. Jelinek, MIT Press, 1998.
- *Foundations of Statistical Natural Language Processing*, C. D. Manning and H. Schutze, MIT Press, 1999.
- *Spoken Language Processing*, X. Huang, A. Acero and H.-W. Hon, Prentice Hall Inc., 2000.
- *Speech and Language Processing*, D. Jurafsky and J. H. Martin, Prentice Hall Inc., 2000.
- *Mathematical Models for Speech Technology*, S. E. Levinson, John Wiley and Sons, 2005.

Speech Enhancement

- *Digital Speech Transmission, Enhancement, Coding and Error Concealment*, P. Vary and R. Martin, John Wiley and Sons, Ltd., 2006.
- *Speech Enhancement, Theory and Practice*, P. C. Loizou, CRC Press, 2007.

Audio Processing

- *Applications of Digital Signal Processing to Audio and Acoustics*, H. Kahrs and K. Brandenburg (eds.), Kluwer Academic Publishers, 1998.
- *Audio Signal Processing and Coding*, A. Spanias, T. Painter and V. Atti, John Wiley and Sons, 2007.