Theory and Application of Digital Speech Processing
by
L. R. Rabiner and R. W. Schafer

Preliminary Edition
Chapter 1
June 3, 2009

Chapter 1

Introduction to Digital Speech Processing

1.1 Digital Speech Processing

Digital signal processing (DSP) technology has been extensively applied to problems in a variety of fields including speech and image processing, video processing, radar, sonar, etc. Digital processing of speech signals (DPSS) has an extensive theoretical and experimental base that has been developed in common with DSP over the past 75 years. As a result, a great deal of speech research has been performed and the resulting systems are highly sophisticated and widely used. Further, there exists a highly advanced implementation technology in the form of VLSI (very large scale implementation) technology that exists and is well matched to the computational demands of DPSS. Finally, there are abundant applications that are in widespread use commercially.

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.

1.2 The Speech Signal

The purpose of speech is communication, i.e., the transmission of messages. 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 [3]. 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, that 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 20 and 50. For example, English can be represented by a set of around 40 phonemes.

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 (i.e., 64 possibilities) 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 of 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 can be made to the speech signal (e.g., enhancing the sound) 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 will see that representations of the speech signal (rather than the message content) can 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.

1.3 The Speech Stack

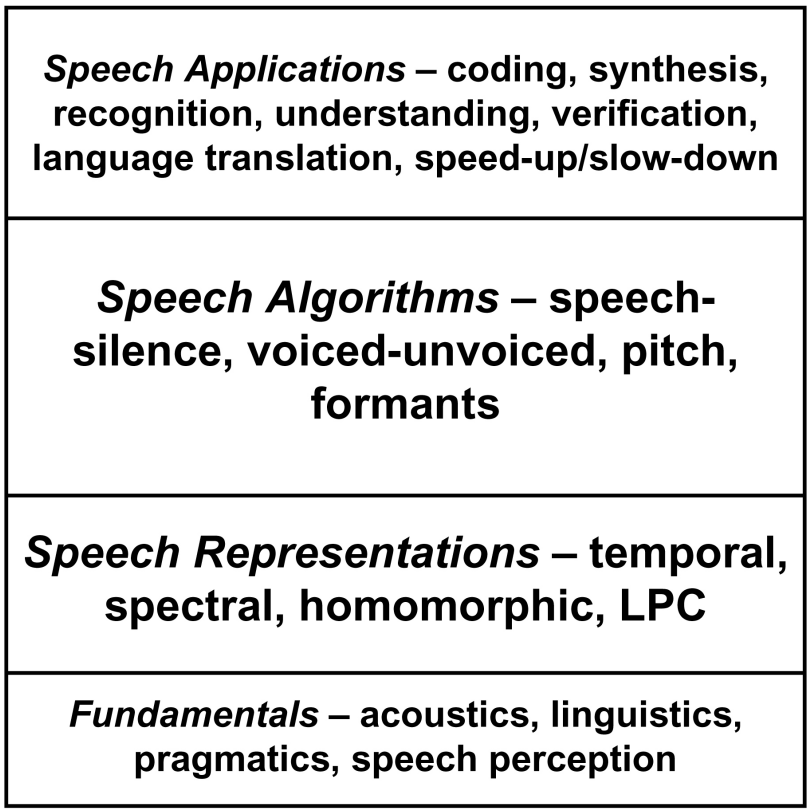

Figure 1.1: The Speech Stack–From Fundamentals to Applications

Figure 1.1 shows an hierarchical view of the basic processes involved in digital processing of speech signals. At the bottom layer of the stack are the fundamentals of speech science in the form of acoustics (speech production), linguistics (the sound codes of speech), pragmatics (the understanding of the task or scenario that the speech represents), and perception (of sounds, syllables, words, sentences and ultimately meaning). These fundamental processes form the theoretical basis for the signal processing that is performed to convert a speech signal into a form that is more useful for getting at the information embedded in the speech signal.

The second layer in the speech stack includes all the basic representations of the speech signal. These representations are in the form of:

- a temporal representation (including the speech waveform itself),

- a spectral representation (Fourier magnitudes and phases),
- a homomorphic representation (the so-called *cepstral* domain), and finally
- a modeling domain such as linear predictive coding (LPC).

It will be shown, throughout this book, that each of these representations has various strengths and weaknesses and, as such, all are used widely and extensively in modern speech processing systems.

The third layer in the stack is the one that converts the various speech representations into algorithms that estimate important properties of the speech signal; e.g. whether a section of the signal waveform should be classified as:

- speech or silence (background signal),
- voiced speech or unvoiced speech or background signal

If the section of speech is classified as voiced speech, various speech algorithms (known collectively as pitch detection methods) facilitate the determination of the pitch period (or pitch frequency), and a different set of algorithms (known collectively as formant estimation methods) can be used to estimate the set of vocal tract resonances or formants for all speech sounds.

The fourth and top layer in the stack is the set of end-user applications of speech processing. This layer represents the payoff for the technology and consists of a range of applications including speech coding, speech synthesis, speech recognition and understanding, speaker verification and recognition, language translation, speech enhancement systems, speech speed-up and slow-down systems, etc. In the following sections we give an overview of several of these application areas, and in the later chapters of this book we give detailed descriptions of the main three application areas, namely speech and audio coding, speech synthesis and speech recognition and natural language understanding.

1.4 Organization of the Book

The organization of this book closely follows the layers in the speech stack. We begin, in Chapter 2, with a review of the most important digital signal processing concepts that are utilized throughout this book for the various speech representations and measurements. We place special emphasis on three topics, namely:

1. conversion from the time domain to the frequency domain (via discrete Fourier transform methods),
2. understanding the impact of sampling in the frequency domain (i.e., aliasing in the time domain), and
3. understanding the impact of sampling (both down and up sampling) in the time domain, and the resulting aliasing or imaging in the frequency domain.

Following our review of the basics of digital signal processing (DSP) technology, we move on to a discussion of the fundamentals of speech production and perception in Chapters 3 and 4. We begin with a discussion of the acoustics of speech production. We derive a series of acoustic-phonetic models for the various sounds of speech and show how linguistics and pragmatics interact with the acoustics of speech production to create the speech signal along with its linguistic interpretation. We complete our discussion of the fundamental processes underlying speech with a look at speech perception processes, from the ways in which sound is processed in the ear to the transduction of sound to neural signals in the auditory neural pathways leading to the brain. We briefly discuss one possible way of embedding our knowledge of speech perception into an auditory model that can be utilized in speech recognition applications. In Chapter 5 we complete our discussion of the fundamentals of speech processing with a discussion of the issues involved in sound propagation in the human vocal tract. We show that a uniform tube approximation to the vocal tract has resonances representative of a neutral vowel (sort of like the 'uh' sound in English). Further, we analyze the transmission properties of a sequence of uniform tubes of varying lengths and cross-sectional areas, and show that we can represent a range of different sounds using such an arrangement of tubes. We show how the transmission properties of a series of concatenated tubes can be represented by an appropriate 'terminal-analog' digital system with a specified excitation function and a specified system response corresponding to the differing tube lengths and areas, along with the radiation characteristic for transmission of sound at the lips.

We devote the next four chapters of the book to digital speech representations, with separate chapters on each of the four major representations. We begin, in Chapter 6, with the temporal model of speech and show how we can estimate basic properties of speech from simple time-domain measurements. In Chapter 7 we show how the concept of short-time Fourier analysis can be applied to speech waveforms in a simple and consistent manner such that a complete, transparent, analysis-synthesis system can be realized. We show that there are two interpretations of short-time Fourier analysis-synthesis systems and that both can be utilized in various applications, depending on the nature of the information that is required for further processing. In Chapter 8 we describe a homomorphic representation of speech where we utilize the property that a convolutional signal (such as speech) can be transformed into a set of additive signals by utilizing transforms and logarithms at suitable points in the processing. We show that a speech signal can be well represented as the convolution of a glottal excitation signal and a vocal tract system that represents the speech sound being produced. Finally, Chapter 9 deals with the theory and practice of linear predictive analysis which is a representation of speech that models the current speech sample as a linear combination of the previous p speech samples, and finds the coefficients of the best combining method (a linear predictor) that optimally matches the speech signal over a given time duration.

Chapter 10 deals with using the signal processing representations presented in the preceding chapters and shows how to measure (or estimate) various properties and attributes of the speech signal. Hence we show how the measurements of short-time (log) energy, short-time zero crossing rates, and short-time auto-

correlation can be utilized to estimate basic speech attributes such as whether a (finite) section of signal represents speech or silence (background signal), whether a speech signal represents a segment of voiced or unvoiced speech, the pitch period (or pitch frequency) associated with a section of voiced speech, and finally, the formants (vocal tract resonances) associated with a section of speech. For estimating pitch period (or frequency) we show how each of the four digital speech representations of Chapters 6-9 can be used as the basis of an effective algorithm . Similarly we show how to estimate formants based on measurements from two of the four digital speech representations.

Chapters 11 to 14 deal with the major speech, natural language understanding, and audio applications. These applications are the payoffs of understanding speech and audio technology and they represent decades of research on how best to integrate various representations and measurements to give the best performance for each application. Our goal in discussing speech applications is to give the reader a sense of how such applications are built and how well they perform at various bit rates and for various task scenarios. In particular, Chapter 11 deals with speech coding systems (both open-loop and closed-loop systems); Chapter 12 deals with audio coding systems based on minimizing the perceptual error of coding using well understood masking criteria; Chapter 13 deals with building tex-to-speech (TTS) synthesis systems suitable for use in a speech dialog system, and Chapter 14 deals with speech recognition and natural language understanding systems and their application to a range of task-oriented scenarios. Entire books have been written on each of these application areas and the interested reader is referred to the many excellent texts that are readily available. (Some representative examples of textbooks in the area of speech processing are given in references [4]- [17]; in the area of speech coding some representative examples are [18]- [25]; in the area of speech synthesis some representative examples are [26]- [33]; in the area of speech and speaker recognition some representative examples are [34]- [41]; in the area of speech enhancement some representative examples are [42]- [43]; and in the area of audio coding some representative examples are [44]- [45]).

1.4.1 The Technology Triangle

In order to fully appreciate almost any technology (including speech processing), it is essential to realize that there are three levels of understanding that must be reached, namely the theoretical level (theory), the conceptual level (concepts) and the working level (practice). These concepts are illustrated in Figure 1.2. At the theoretical level the reader must understand the basic mathematics of the speech representations, the derivations of various properties of speech associated with each representation, and the basic signal processing mathematics that relates the speech signal to the real world via sampling, aliasing, filtering, etc. At the conceptual level the reader must get a good understanding of how the theory is applied to make various speech measurements and to estimate various attributes of the speech signal. Finally it is essential to be able to convert theory and conceptual understanding to practice; that is to be able to implement speech processing systems in working computer code (most often as a program written in MATLAB, C or C++) or as specialized code running on real-time

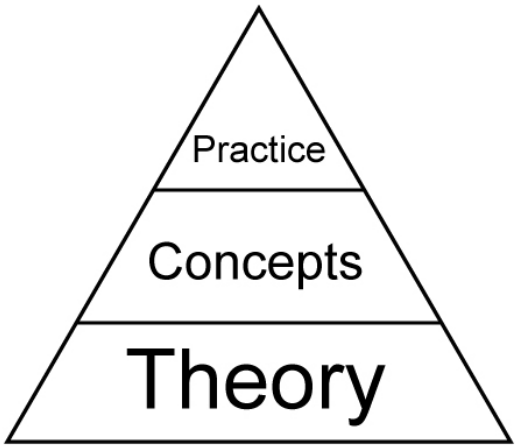

Figure 1.2: The Technology Triangle–Theory, Concepts and Practice

signal processing chips (ASICs, Field Programmable Gate Arrays (FPGAs), or DSP chips). For every topic in speech processing that is covered in this book, we will endeavor to provide as much understanding as possible at the theory and concepts level, and we will provide exercises that enable the reader to gain expertise at the practice level (usually via MATLAB exercises).

1.5 Speech Applications

In this section we present a brief overview of the key speech applications as this is generally the motivation for the reader to understand speech processing technology. Speech applications also represent the real world payoff for speech technology as these are the systems that users interact with on a day-to-day basis.

1.5.1 Speech Coding

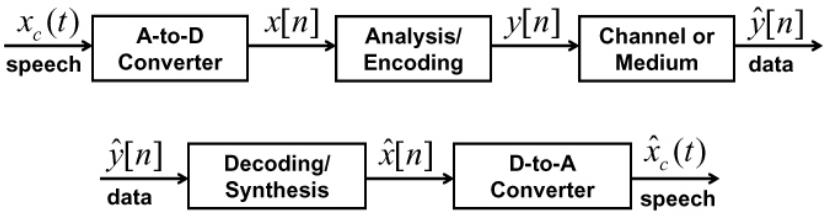

Figure 1.3: Speech Coding Block Diagram–Encoder and Decoder

Figure 1.3 shows a block diagram of a generic speech coding system. The

upper path in the figure shows a speech encoder beginning with an analog-to-digital (A-to-D) converter that converts an analog speech signal, $x_c(t)$, to a digital representation, $x[n]$[1]. Depending on the details of the speech coder, the digital signal $x[n]$ is analyzed and encoded giving the processed digital signal $y[n]$ which is then transmitted over a channel (or other medium) giving the final encoded signal $\hat{y}[n]$ (which is often referred to as the data signal of the coder).

The lower path in Figure 1.3 shows the decoder associated with the speech coder. The data signal $\hat{y}[n]$ is synthesized (or decoded) using the inverse of the analysis processing, giving the signal $\hat{x}[n]$ which is then converted back to an analog signal using a digital-to-analog converter (D-to-A), giving the processed signal $\hat{x}_c(t)$.

The goal of speech coding systems is to enable efficient transmission and storage of speech for a broad range of applications including narrowband and broadband wired telephony, cellular and other wireless 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 storage of speech for telephone answering machines, interactive voice response (IVR) systems, and pre-recorded messages.

In order to understand the effectiveness of speech coding systems, it is essential to listen to a range of speech utterances that have been encoded and decoded using systems of the type shown in Figure 1.3. To that end, the enclosed CDROM (included with this book and also available at the book website) contains examples of coding of narrowband speech (3.2 kHz bandwidth as might be used over ordinary telephone lines), coding of wideband speech (7.5 kHz bandwidth as might be used in AM Radio), and finally coding of audio signals (20 kHz bandwidth as used on music CDs). The narrowband speech coding demos include the following samples:

- 64 Kbps Pulse Code Modulation (PCM)
- 32 Kbps Adaptive Differential Pulse Code Modulation (ADPCM)
- 16 Kbps Low Delay Code-Excited Linear Prediction (LDCELP)
- 8 Kbps Code-Excited Linear Prediction (CELP)
- 4.8 Kbps Federal Standard 1016 (FS1016)
- 2.4 Kbps Linear Predictive Coding 10E (LPC10E)

When listening to these coded narrowband speech utterances, the reader should get a feeling for when the speech quality begins to change (between 8 and 16 Kbps) and when the speech quality begins to deteriorate (between 4.8 and 8 Kbps). The reader will hopefully gain an understanding of why this happens when we discuss the details of each of these speech coders in Chapter 11.

[1]Throughout this book we use the parenthesis notation to denote a continuous variable, e.g., $x_c(t)$, and we use the square bracket notation to denote a discrete time variable, e.g., $x[n]$.

The demonstrations of wideband speech coding (also contained on the CDROM and at the book website) utilize both a male talker and a female talker and include the following samples:

- 3.2 KHz bandwidth, uncoded speech sample
- 7 KHz bandwidth, uncoded speech sample
- 7 KHz bandwidth, coded at 64 Kbps
- 7 KHz bandwidth, coded at 32 Kbps
- 7 KHz bandwidth, coded at 16 Kbps

When listening to the samples in the wideband speech coding demo, the reader should immediately notice the quality difference between 3.2 kHz bandwidth speech (narrowband) and 7 kHz bandwidth speech (wideband) in the form of a dramatic increase in presence and naturalness. Although the intelligibility of speech is adequately preserved in narrowband speech (which is used for both wireless and wireline telephony), the quality is only preserved in wideband speech. Further, by listening to the coded wideband samples at 64, 32 and 16 Kbps rates, it is clear that there is no major degradation in quality even at the 16 Kbps rate.

The third set of demos illustrates audio coding of CD quality music using the well known MP3 audio coder. The original CD music was recorded at a sampling rate of 44.1 kHz, in stereo, with 16 bits per sample quantization for an overall bit rate of 1.4 Mbps. The MP3-coded material is played at an overall bitrate of 128 Kbps. The demo presentation includes two versions of each selection, one at the 1.4 Mbps original CD rate, and one at 128 Kbps representing the MP3-coded version. The order of the presentations is random for each selection. The reader should decide whether he or she can determine which of each pair of recordings is the CD original and which is the MP3-coded version. The individual selections are the following:

- female vocal
- trumpet selection
- baroque
- guitar

For those readers who took this test seriously, the material actually presented in the demo was in the following order:

- female vocal–CD original, MP3-coded
- trumpet selection–CD original, MP3-coded
- baroque–CD original, MP3-coded
- guitar–MP3-coded, CD original

It is a very difficult task to determine which of the pair of music files is the original CD selection and which is MP3-coded since the resulting coded audio quality is extremely close to the CD original quality. More discussion of speech and audio coding systems will be provided in Chapters 11 and 12.

1.5.2 Speech Synthesis

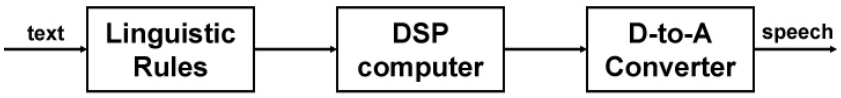

Figure 1.4: Speech Synthesis System Block Diagram

Figure 1.4 shows a block diagram of a canonic system for speech synthesis from text. The input to the system is ordinary text (think of an email message, an article from a newspaper or magazine, etc.). The first block in the synthesis system 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 have to determine the appropriate set of sounds (including factors like emphasis, pauses, rates of speaking, etc.) so that the resulting synthetic speech expresses (hopefully in a natural voice) the words and intent of the text message. Hence the linguistic rules component must figure out how to pronounce acronyms, how to determine the pronunciation of ambiguous words like read, bass, object, etc., how to determine the pronunciation of 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 DSP computer is to create the appropriate sound sequence to match the text message in the form of speech. There are many procedures for assembling the speech sounds and compiling them into a proper sentence, but the most popular one today is called "unit selection and concatenation" whereby the computer stores multiple versions of each of the basic units of speech (phones, half phones, syllables, etc.), and then optimally decides which sequence of speech units sounds best for the particular text message that is being produced (the process of unit selection). The last block in the speech synthesis system is the conversion from a digital to an analog signal via the D-to-A converter.

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 telematics feedback 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, etc., and serve as the voice of announcement machines that provide information such as stock quotes, airline schedules, updates on arrivals and departures of flights, etc.

Much like speech coding, the only proper way to evaluate speech synthesis systems is by listening to their outputs for a range of input text messages. The demo CDROM includes a set of examples of speech synthesis systems based on unit selection and concatenation. The particular examples included on the CDROM that comes with this book are the following:

- Soliloquy from Hamlet
- Gettysburg Address
- Third Grade Story

Although there are clear sections where the reader can readily determine that a machine created the speech associated with the text message (rather than a human talker), the quality of speech synthesis by machine has steadily risen over the past decade and is rapidly approaching natural speech quality, at least for certain application domains.

1.5.3 Speech Recognition and Other Pattern Matching Problems

Figure 1.5: Block Diagram of General Pattern Matching System

Figure 1.5 shows a block diagram of a generic approach to pattern matching problems (including speech recognition, speaker recognition, speaker verification, word spotting, and automatic indexing of speech recordings). The first block in the system converts the analog speech waveform to digital form using an A-to-D converter. The feature analysis module converts the digital speech signal to a set of feature vectors (based on one of the short-time analysis methods discussed in the chapters on speech representations) which efficiently characterize the features (over time) of the speech being matched by the system. The final block in the system, namely the pattern matching block, dynamically time aligns the feature vector of 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 dynamically aligned feature vector of the speech signal. The identity 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, etc.

Although the block diagram of Figure 1.5 is generic and can be used for a wide range of pattern matching problems, the predominant 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 (C&C) applications, e.g., control of the look and feel of a computer desktop, control of fonts and styles in word processing, simple control commands for spreadsheets, control of presentation graphics, etc., voice dictation to create letters, memos, and other documents, natural language voice dialogues with machines to enable help desks and call centers, voice dialing for cellphones and from PDAs and other small devices, and for agent services such as calendar entry and update, address list modification and entry, etc.

The best way to gain an understanding as to the state of the art in speech recognition and natural language understanding is via video demos of how such systems work for some simple task scenarios. The enclosed CDROM includes two such demos, labeled Air Traffic Demo and Broadcast News Demo. These demos illustrate the performance and speed of modern speech recognition and understanding systems.

1.5.4 Other Speech Applications

Although we will not discuss any other speech applications in detail, it is worthwhile mentioning some of the key areas of interest other than speech coding, speech synthesis and speech recognition.

The area of *speaker verification* is important for providing secure access to premises, information and virtual spaces. Speech is a natural biometric feature that is transportable and available always as a means of verifying a claimed identity. Speaker verification research has shown how to build systems that can perform adequately for moderate security tasks; e.g., order of 0.5% "equal error rate" systems, where the probability of rejecting a valid speaker is the same as the probability of accepting an invalid speaker (an imposter) and both probabilities are about 0.5%.

The area of *speaker recognition* is also an important application area for speech processing systems. Speaker recognition technology is intended for use in legal and forensic applications, that is to identify a speaker from a database of known speakers, for use in criminal investigations, etc. Here the task is considerably more difficult than speaker verification since there is a non-zero probability that the speaker to be recognized is not one of the speakers in the database (potentially leading to one type of error), or that the speaker is in the database but has disguised his or her voice sufficiently that he or she cannot be reliably recognized (a second type of error). The goal of most speaker recognition systems is to minimize the second type of error, while holding the first type of error at some acceptable level.

The area of *speech enhancement* is an application area that is widely used in noisy environments. The goals of speech enhancement systems are to reduce noise, to eliminate echo and other spectral distortions, to align voices with video segments, to change voice qualities, to speed-up or slow-down prerecorded speech (e.g., for talking books, rapid review of material, careful scrutinizing of spoken material, etc.), and to generally make the speech signal more useful for further processing as part of a speech coder, synthesizer, or recognizer. The natural goal of speech enhancement systems is to make the speech more intelligible and more natural; in reality the best that one can achieve is less perceptually

annoying speech that essentially maintains the intelligibility of the noisy speech.

Another major speech application that has become very popular recently is the area of *language translation.* The goal of language translation systems is to convert spoken words in one language to another language in order to facilitate natural language voice dialogues between people speaking different languages, i.e., tourists, business people. Language translation technology requires speech synthesis systems that work well in both languages, along with speech recognition (and generally natural language understanding) that also works well for both languages; hence it is a very difficult task and one for which only limited progress has been made.

1.6 Speech Enabled Devices

Figure 1.6: A Range of Speech-Enabled Devices

Figure 1.6 shows 6 fairly common devices that utilize speech and audio processing technology. Internet audio devices enable downloads of coded audio signals to a range of audio players which faithfully reproduce the signal at bit rates substantially below the uncoded rate of 1.4 Mbps. The most general format for the audio is MP3 at 128 Kbps, but devices also can utilize the more recent AAC (Advanced Audio Coding) format at bit rates of from 64-128 Kbps, or the WMA (Windows Media Audio) format at bit rates of from 64-128 Kbps. The second device, shown at the middle of the top row of Figure 1.6 uses speech technology to control the actions of the digital camera via simple speech recognition commands. The third device shown at the top of Figure 1.6 is a PDA (Personal Digital Assistant) which receives coded audio and video streams, thereby enabling people to listen to radio broadcasts and view video broadcasts

of news, sporting events, television shows, and whatever other entertainment they deem interesting and appropriate on a small portable device.

Perhaps the preeminent usage of speech technology in handheld devices is the speech coding that enables normal voice conversations in cell phones (based on speech coding at bit rates on the order of 8 Kbps), as well as the voice dialing capability that retrieves names and automatically dials the associated numbers by recognizing simple voice commands. Name directories with upwards of several hundred names can readily be recognized and dialed using simple speech recognition technology.

Other key applications of speech technology include hearing aids (both the amplification type as well as the more sophisticated compression technology that does auditory analysis comparable to a human cochlea and feeds the resulting signals to the inner hair cells) and MP3 audio players that store up to several thousand coded songs (in MP3 or AAC format) on a hard disk and enable the music to be played virtually anywhere.

Figure 1.7: Apple iPod Player

Figure 1.7 shows the Apple iPod music player that has a hard disk with up to 80 GB of memory that can store about 24,000 songs, or the equivalent of about 1200 CDs in compressed format. More than 120 million iPods have been sold by the end of 2007.

Figure 1.8 shows the architecture of a modern cellular phone. One of the most important parts of the cellphone is the DSP chip that is used for speech coding and for speech recognition of user names and numbers. More than 1.2 billion such cellphones were sold throughout the world in 2007.

1.7 The Speech Chain [10]

The goal of this book is to provide an understanding of the *nature of the speech signal,* and how digital signal processing techniques, along with communication technologies and information theory methods can be applied to help enable the various application scenarios described in the preceding section. We will see that the main emphasis of speech signal processing is the process of converting

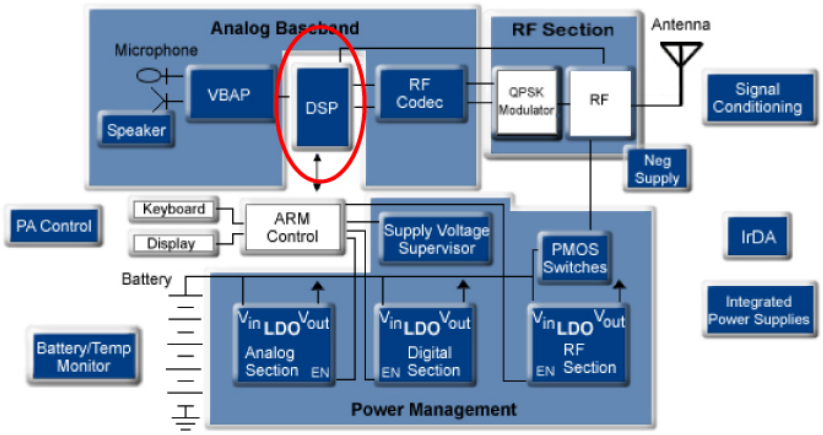

Figure 1.8: Cellular Phone Architecture

one type of speech signal representation to another, so as to uncover various mathematical and physical properties of the speech signal, and to do the appropriate processing to aid in solving both fundamental and deep problems of interest. To that end, we now examine a generic model of speech production (or speech generation).

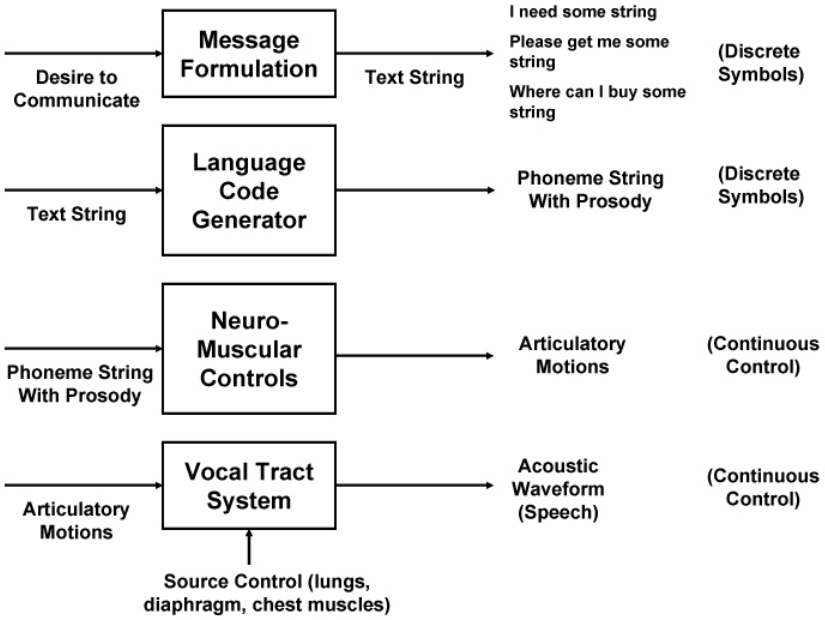

Figure 1.9: The Steps in the Speech Production Chain

Figure 1.9 shows the steps that occur in the process of converting a set of thoughts that constitute a message into a speech signal that can be heard by a listener. The first step in the chain is called *message formulation* and is basically the user's conscious desire to communicate a message in the form of an idea, a wish, or a request and to express the resulting message as an appropriate sequence of sounds. We represent this message formulation step as producing a text string of discrete symbols corresponding to the message that is to be communicated. For example, the end result of a conscious desire to obtain some string for an experiment could be any of the sentences: "I need some string", or "Please get me some string" or "Where can I buy some string".

The second step in the speech production model is the conversion of the desired text string to a sound or phoneme string with prosody markers (a new set of discrete symbols). We call this step the language code generator step since it converts text symbols to phonetic symbols (along with stress and durational information) which describe the basic sounds of the intended message and the manner (i.e., the speed and emphasis) in which the sounds are intended to be reproduced. The third step in the speech production process is the utilization of neuro-muscular controls, i.e., the set of control signals that direct the neuro-muscular system to appropriately move the speech articulators, namely the tongue, lips, teeth, 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 enable 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 creates the desired speech signal corresponding to the desired message. It does this by creating the appropriate sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform (the speech signal) that is understandable in the environment in which it is spoken. The entire speech production process is often called the speech chain because of the organized sequence of events that occur from message formulation to speech signal creation, as outlined above [10].

An example of the waveform of a couple of time slices of a speech signal is shown in Figure 1.10. The upper trace shows an example of a voiced speech signal where the vocal cords are vibrating during the production of the speech signal, as seen by the quasi-periodic signal that results in the interval after the background signal ends. The quasi-periodicity of the voiced sound is also seen by the repetition of the signal at intervals which are labelled as the pitch period or about 32 msec for this example. The lower trace of Figure 1.10 shows an example of an unvoiced speech signal which we identify by the noise-like sound that is produced, with no trace of periodicity in the signal, but instead a random behavior of the speech samples.

The complete speech chain consists of a speech production/generation model, of the type discussed above, as well as a speech perception model, as shown in Figure 1.11. The speech perception model shows the steps from capturing speech at the ear to understanding the message underlying the speech signal. The first step is the conversion of the acoustic waveform of the speech to a spectral representation as performed by the basilar membrane within the inner

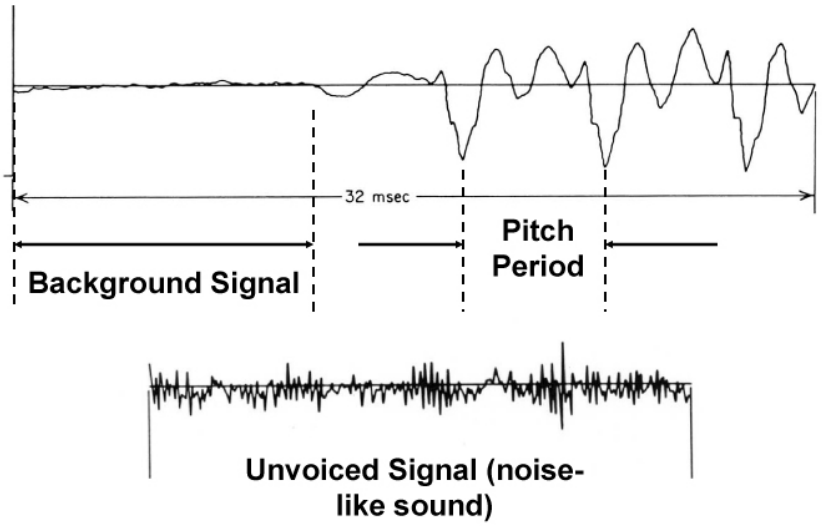

Figure 1.10: Examples of the Speech Signal; Voiced Speech (top), unvoiced speech (bottom)

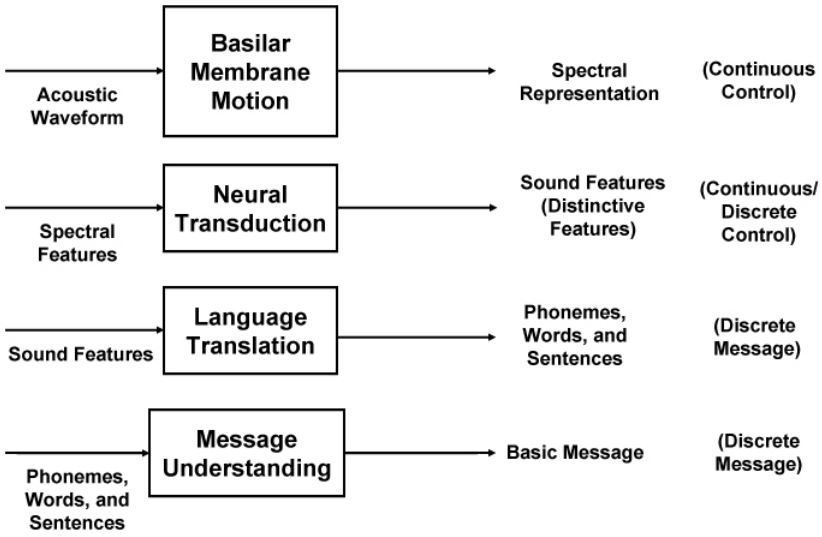

Figure 1.11: The Steps in the Speech Perception Chain

ear. We will see later in Chapter 4 that the basilar membrane is fundamentally a non-uniform spectrum analyzer that effectively spatially separates the spectral components of the incoming speech signal and analyzes them using 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) which represent a fundamental base set of sound features 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 is rudimentary at best, but it is generally agreed upon that some physical correlate of each of the steps in the speech perception model occurs within the human brain, and thus the entire model is useful for thinking about the processes that occur.

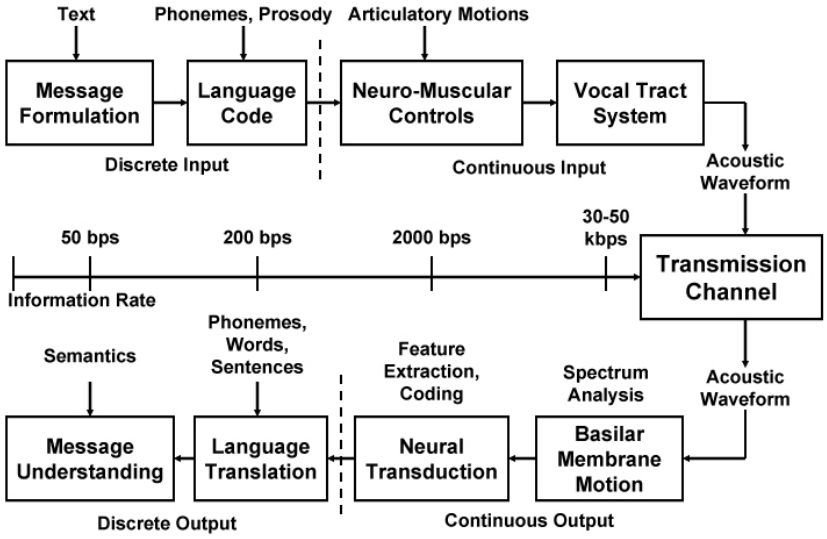

Figure 1.12: The Speech Chain–From Idea to Speech to Understanding

It is instructive to show the complete speech chain from the formulation of an idea to the creation of speech and finally to the understanding of the idea on the part of the listener as a complete chain, as shown in Figure 1.12, and to estimate the information rates at various steps in the process. Initially, at the idea stage, the information is embedded in the text message and we have already shown how to make a crude estimate of the information rate of a text message. Assume 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 symbols), and that the rate of speaking for most people is about 10 symbols per second (somewhat on the high side, but still acceptable for a crude information rate estimate). We can estimate the base information rate of the text message

as about 50 bits per second (5 bits per symbol times 10 symbols per second).[2] At the second step of the process, where we actually convert the text representation into phonemes and prosody markers, the information rate is estimated to increase by a factor of 4 to about 200 bits per second. The information representation for the first two processes in the speech chain is discrete so we can readily estimate the information rate with some simple assumptions. For the next two steps in the speech generation part of the speech chain, the information becomes continuous in the form of control signals for articulatory motion and ultimately the speech waveform, so we need to estimate the spectral bandwidth of the control signals and appropriately sample and quantize these signals to estimate the information rate. Using some realistic estimates of bandwidth and required signal quantization accuracy, we estimate that the information rate of the articulatory control signals is about 2000 bits per second (bps), and that the information rate of the speech waveform, at the end of the speech chain, is somewhere between 30,000 and 50,000 bps (depending on the bandwidth that is utilized and the precision with which speech is represented). The interesting aspect of the analysis of information rates is that over a range of about 1000-to-1 (from 50 bps text rates to 50,000 bps waveform rates), fundamentally the same information content is preserved at each level of the speech chain. Hence the goal of many speech applications (e.g., speech coding, speech recognition, speech synthesis, etc.) is to convert the 50,000 bps speech signal (basically all that we have to work with in most real world systems) back to the 50 bps textual representation (this is fundamentally the speech recognition process), or to the 200-2000 bps range for speech coding applications, without seriously degrading the speech quality and intelligibility. These are hard problems and have been areas of intensive research in speech processing for more than 50 years.

There is one additional process shown in the diagram of the complete speech chain that we have not discussed yet, namely the inclusion of a transmission channel between the speech generation and speech perception parts of the model. This transmission channel might consist of just the free space air connection between two speakers who are in a common space, or it could consist of a wireline or wireless channel between two talkers who are connected by a wireline or cellular communications channel. In any case, it is essential to include this transmission channel in our model for the speech chain as it includes real world noise and channel distortions that make speech and message understanding more difficult processes in real communication environments.

The process of converting a message idea into a speech waveform and then back into an understood message is often more difficult than depicted in the speech chain. The following example illustrates this problem with a simple message scenario. We assume that the basic message you want to create has the goal of determining if your office mate has had lunch already. Hence the following scenario might occur for this simple message context:

- **Goal:** Find out if your office mate has had lunch already.
- **Text:** "Did you eat yet"

[2]Earlier we estimated the base information rate of the phonetic message as about 60 bits per second. This somewaht higher rate is due to the fact that there are more basic sounds (phonemes) in English than letters of the alphabet.

- **Phonemes:** /dId yu it yɛt/
- **Articulator Dynamics:** /dI jə it jɛt/

Although the perfectly articulated sentence, as seen by the sequence of phonemes shown above, is completely understandable, the reality of human speech production is that the sounds co-articulate (blend into each other, especially at word boundaries) as much as possible, leading to the set of actual sounds shown in the line labeled articulator dynamics above, where the phoneme /d/ at the end of 'did' and the phoneme /y/ at the beginning of 'you' co-articulate to produce the /j/ sound, and the vowel /u/ of 'you' becomes the highly reduced vowel /ə/; furthermore the beginning sound of 'yet' becomes a /j/ sound rather than the fully articulated /y/ sound, as expected from the text. The amazing aspect of human speech perception is that almost every native speaker of English will understand the highly co-articulated sequence shown above, irrelevant of the degree of co-articulation in producing the speech signal. All of these linguistic variations in sounds make speech processing for speech understanding a rather difficult process, as we will see in Chapter 14.

1.8 Speech Science

Although the primary emphasis in this book is on signal processing analyses of speech signals, a complete understanding of speech requires forays into the following areas and sub-areas of speech science:

- **Linguistics:** the science of language
- **Phonetics:** the study of speech sounds, their production, transmission, and reception, and their analysis, classification and transcription
- **Phonemes:** the smallest set of units considered to be the basic set of distinctive sounds of a language (there are between 20 and 60 phoneme units for most languages)
- **Phonemics:** the study of phonemes and phonemic systems
- **Syntax:** analysis and description of the grammatical structure of a body of textual material
- **Semantics:** analysis and description of the meaning of a body of textual material and its relationship to a task description of the language

We will have numerous occasions, throughout this book, to examine each of these areas of Speech Science to see how they affect algorithms for speech processing for various applications. Further, in Chapter 3, we will extensively look at the acoustic-phonetic properties of speech as a basis for understanding how speech is created and how various speech sounds are represented in the time and spectral domains.

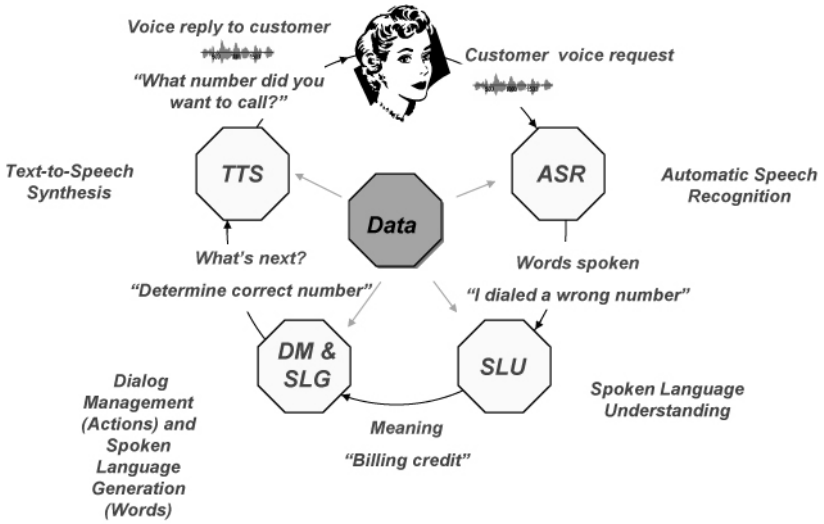

Figure 1.13: The Speech Dialogue Circle for Maintaining a Voice Dialogue With a Machine

1.9 The Speech Dialogue Circle

Speech recognition and understanding (both by humans and by machines) are error prone processes due to factors such as slurred speech, highly co-articulated words (as in the example above), strong accents, fatigue, noisy backgrounds, strong interference from other talkers, etc. Hence humans have naturally developed a process for communications (both with other humans and more recently with machines) of extracting as much information as possible from the spoken inputs and utilizing a dialogue circle that enables speech understanding to be built up in a series of 'turns' around the speech dialogue circle, as illustrated in Figure 1.13. The dialogue circle consists of the steps that both a human and a machine would naturally go through in order to maintain communication with another human (or a machine) without necessarily asking for all the speech to be repeated whenever it wasn't completely understood the first time. Hence at the start of the cycle, the human (called a customer in the figure) makes a voice request for information or help, e.g., by speaking the phrase "I dialed a wrong number" in response to a welcome greeting. For human-machine communications, as illustrated in Figure 1.13, (as well as for human-human communications), the first step in the dialogue circle is to recognize the spoken input and convert it to the textual word sequence of the underlying message. This task is done by an automatic speech recognizer (ASR) system (or by a human listener) that chooses the most likely sentence that was spoken. The next stage in the dialogue circle has to determine the 'meaning' of the spoken word sequence, namely to 'understand' (rather than just recognize) the speech. This job is performed by a spoken language understanding module that utilizes semantics to determine the most likely meaning, in the context of a telephony

help desk, of the spoken input sentence. In this case the system determines that the user most likely wants a billing credit for the mis-dialed number, although it is highly likely that the user might want some assistance in dialing the correct number, or in looking up the correct number, etc. In any case the system now must take some action and that is the job of the box labeled dialogue management and spoken language generation. In this case, the system decides that the most appropriate course of action is to determine the correct number that was to be dialed, and thus it generates the sentence "What number did you want to call?" and sends it to a text-to-speech (TTS) module that generates a synthetic speech utterance (of extremely high quality) that asks the customer the desired phone number to be dialed. The customer then responds appropriately to the request for information and the dialogue circle begin a new loop. Typically, to complete a complex transaction, the dialogue circle is traversed anywhere from 2 to 10 times, with each complete loop increasing the state of knowledge about what is desired and getting the customer closer to the desired transaction. This same dialogue circle works in human-human communication where each cycle represents either a clarification of information from a previous cycle or addition of new information that enables the dialogue to move forward.

1.10 The Information Rate of Speech

We will be working extensively with a range of representations of the speech signal. Although we have already discussed ways of estimating the basic information rates of speech, it is essential to understand that this rate strongly depends on the signal representation. There are two fairly extreme ways of measuring the basic information rate of a speech signal, namely from a Shannon view of information and from a signal processing and communication point of view. The Shannon view looks at the source of the signal, namely the message content, and decides how much information is present in the message. The signal processing or communication view looks at the speech waveform and determines the information rate of the communications channel that allows the signal to pass without adding any distortion or noise. In this section we look at the speech signal from these two points of view and contrast the resulting information rates.

1.10.1 The Shannon View of Information

From the Shannon view of information, we need to look at the orthographic text message and try to determine how much information is conveyed in the text symbols. If we take the simplistic point of view, namely that all text symbols are equally likely, then we can count the number of unique text symbols that we encounter, determine how many bits are required to represent all the text symbols, determine the rate of text symbols for a typical speaker, and then we can get the total information rate of the speech in message form. The English language has a 26 letter alphabet with about 26 additional punctuation and other diacritical symbols appearing fairly often in text documents. For simplicity we assume a symbol library of about 64 symbols; hence a 6-bit representation of the symbol set is more than adequate and we assume each of the 64

symbols is equally likely (a highly imprecise calculation since the letters of the alphabet are far more likely to occur than punctuation or other symbols, and further some letters are far more likely to occur than others). For the average rate of talking we get about 10 letters per second (this corresponds roughly with a speaking rate of about 120 words per minute). Hence the total information rate of a text message is the product of 6 bits/symbol times the rate of 10 symbols per second, for a total information rate of about 60 bps, a result which we have seen earlier in this chapter.

1.10.2 The Signal Processing View of Information

The signal processing view of information is relatively easy to understand and leads to a simple calculation of information rate in the speech signal. The first thing we need to decide is the bandwidth of the speech signal as used in a communication system. There are two standard values of speech bandwidth, namely 4 kHz for telephone quality speech (actually the telephone bandwidth of speech is closer to 3.2 kHz but generally we sample telephone speech at a rate of 8000 times/second, so we consider 4 kHz to be the nominal bandwidth of telephone speech), and 8 kHz for wideband hi-fidelity speech as might be used in an AM or FM radio broadcast. The appropriate sampling rates for telephone-quality and wideband speech are 8000 and 16,000 times per second. Although we generally use linear quantizers for sampling waveforms, we will show later in this book that a logarithmic quantizer (with about 8 log encoded bits per sample) is more appropriate for speech, and thus from a signal processing or communications point of view, we consider the information rate of speech as being 64 Kbps (8000 samples per second times 8 bits per sample) for telephone quality speech, and 128 Kbps (16,000 samples per second times 8 bits per sample) for wideband quality speech.

From the above discussion we see that there is a factor of between 1000 and 2000 times difference in information rate of speech from the message view and from the signal processing view. Hence the overarching task in speech processing is to somehow begin with the high information rate speech waveform, and extract information or compress the signal, by more than three orders to magnitude, yet still preserve the message content in an appropriate manner. We have yet to achieve this goal and it may never be realized in practice–but at least it sets the standard to which speech research has aspired for the past many decades.

1.11 Speech Signal Processing Model

The discussion of the preceding section shows that there are many ways of looking at the basic information model for speech–from the information source, which produces symbols at a rate on the order of 60 bps, to the information sink (the human listener or the machine that has to process the speech signal), which must handle about 64-128 Kbps of information flow. A fairly generic model of information manipulation and processing from the speech message to the human listener or the machine is depicted in Figure 1.14. The human speaker is the

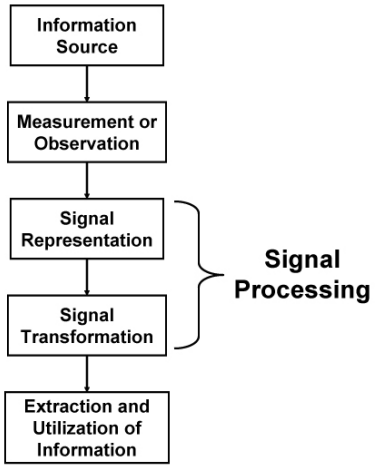

Figure 1.14: General view of information manipulation and processing

basic information source and the measurement or observation is generally the acoustic waveform (although it could equally well be a set of positions of the articulators (over time), or even measurements of the neural control signals for the articulators.

Signal processing involves first obtaining a representation of the signal based on a given model and then applying 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 by either human listeners or 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 unknown speaker. For this example, the information in the signal is the identity of the speaker.

Thus we see that 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.

1.12 The Role of Digital Signal Processing in Speech 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 will 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 1960s 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. Modern computers (circa 2000) are so fast relative to the required processing speeds for most speech processing algorithms, that computation has essentially become a non-issue for all but the most extensive simulations, and virtually all speech processing systems of interest run in fractions of real time on low cost processors.

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 that the tremendous number of computations required in many signal processing functions can be implemented in a fraction of 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 com-

munications 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 communications problems [1].

1.13 Digital Speech Processing

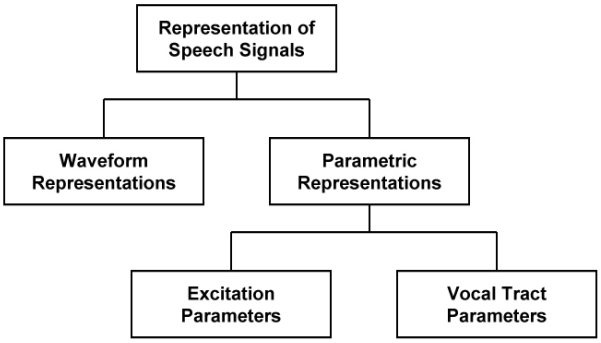

Figure 1.15: Representations of speech signals

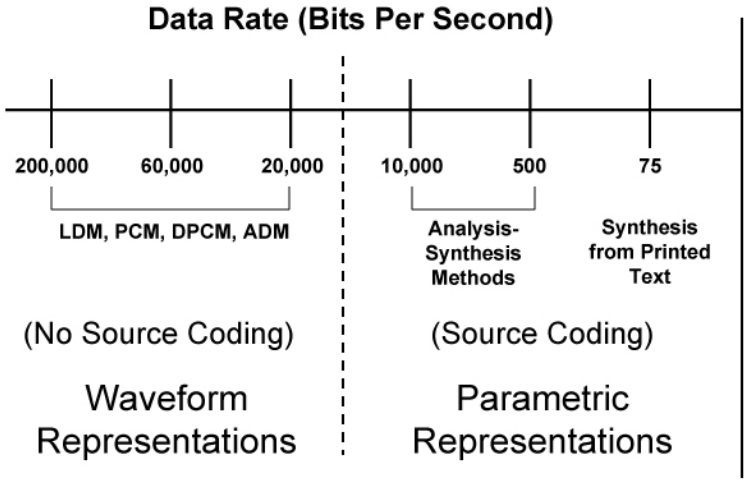

Figure 1.16: Range of bit rates for various types of speech representations. (After Flanagan [1])

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 [2] 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 1.15, these representations can be classified into two broad groups, namely waveform representations and parametric representations. Waveform representations, as 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).[3]

Figure 1.16 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 bps (approximately the basic message information of the text) to data rates upward of 200,000 bps 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 1.17 shows just a few of the many applications areas in speech communications, most of which we have already mentioned in the preceding sections of this chapter.

1.14 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 se-

[3]Chapter 5 provides a detailed discussion of parametric speech models.

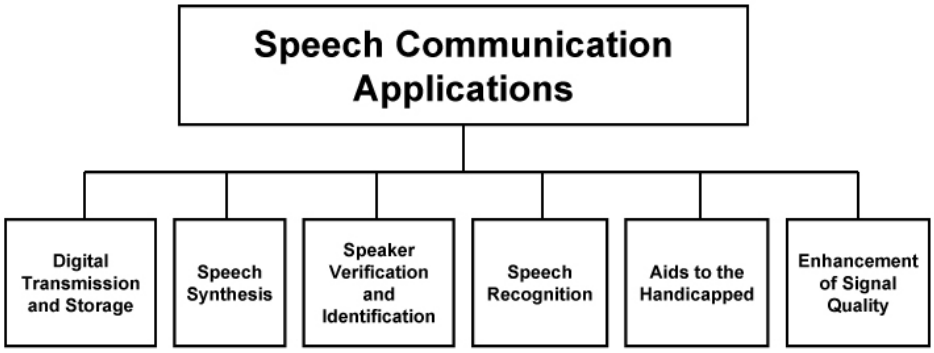

Figure 1.17: Some typical speech communications applications.

lected 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 1.15. Alternatively, a book on speech processing could be written to emphasize application areas based on the hierarchy of Figure 1.17. A third possibility would be to organize the book with respect to signal processing methods. The approach we have chosen is based on the ideas presented in the Speech Stack (Figure 1.1) where we choose to first discuss the fundamental concepts in speech production and perception, followed by discussions of the four signal processing representations of speech, followed by a chapter on speech measurement and parameter estimation systems, and ending with chapters on the major speech applications. As such, the remaining chapters of this book discuss the following topics:

- **Chapter 2** – review of digital signal processing with emphasis on techniques crucial to digital speech processing systems
- **Chapter 3** – the fundamentals of speech generation/production in humans
- **Chapter 4** – the fundamentals of speech perception in humans
- **Chapter 5** – the theory of sound propagation in the human vocal tract
- **Chapter 6** – time-domain methods for speech processing
- **Chapter 7** – frequency-domain methods for speech processing and the short-time Fourier transform
- **Chapter 8** – homomorphic methods for speech processing and cepstral analysis
- **Chapter 9** – linear predictive coding methods for speech processing
- **Chapter 10** – algorithms for measuring and estimating fundamental properties and attributes of speech signals

- **Chapter 11** – digital coding of speech signals
- **Chapter 12** – frequency-domain coding of speech and audio
- **Chapter 13** – digital methods of synthesizing speech from text
- **Chapter 14** – digital methods for recognizing and understanding speech

It is anticipated that the reader will be able to follow the material presented in this book as it follows the structure of the speech stack as closely as possible, and each chapter naturally builds on ideas and concepts introduced in earlier chapters. There are a number of excellent books on speech processing listed in the bibliography at the end of this section and each of them is recommended to the interested reader for additional insights into digital speech processing.

Bibliography

[1] J. L. Flanagan, Computers That Talk and Listen: Man-Machine Communication by Voice, *Proc. IEEE* Vol. 64, No. 4, pp. 416-422, Apr. 1976.

[2] H. Nyquist, Certain Topics in Telegraph Transmission Theory, *Trans. AIEE* Vol. 47, pp. 617-644, Feb. 1928.

[3] C. E. Shannon, A Mathematical Theory of Communication, *Bell System Tech. J.* Vol. 27, pp. 623-656, Oct. 1948.

General Speech Processing References

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

[5] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, 1976.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Inc., 1978.

[7] R. W. Schafer and J. D. Markel, editors, *Speech Analysis*, IEEE Press Selected Reprint Series, 1979.

[8] D. O'Shaughnessy, *Speech Communication, Human and Machine*, Addison-Wesley, 1987.

[9] S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, Marcel Dekker Inc., 1991.

[10] P. B. Denes and E. N. Pinson, *The Speech Chain*, second edition, W. H. Freeman and Co., 1993.

[11] J. Deller, Jr., J. G. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing, 1993, Wiley-IEEE Press, Classic Reissue, 1999.

[12] K. N. Stevens, *Acoustic Phonetics*, MIT Press, 1998.

[13] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley and Sons, 2000.

[14] S. Furui, editor, *Digital Speech Processing, Synthesis and Recognition*, second edition, Marcel Dekker Inc., New York, 2001.

[15] T. F. Quatieri, *Principles of Discrete-Time Speech Processing*, Prentice-Hall Inc., 2002.

[16] L. Deng and D. O'Shaughnessy, *Speech Processing, A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Inc., 2003.

[17] J. Benesty, M. M. Sondhi and Y. Huang, editors, *Springer Handbook of Speech Processing and Speech Communication*, Springer, 2008.

Speech Coding References

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

[19] P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice-Hall Inc., 1984.

[20] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[21] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*, Elsevier, 1995.

[22] T. P. Barnwell and K. Nayebi, *Speech Coding, A Computer laboratory Textbook*, John Wiley and Sons, 1996.

[23] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*, CRC Press, 2000.

[24] W. C. Chu, *Speech Coding Algorithms*, John Wiley and Sons, 2003.

[25] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, second edition, John Wiley and Sons, 2004.

Speech Synthesis References

[26] J. Allen, S. Hunnicutt, and D. Klatt, *From Text to Speech*, Cambridge University Press, 1987.

[27] J. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English*, Springer-Verlag, 1993.

[28] Y. Sagisaka, N. Campbell, and N. Higuchi, *Computing Prosody*, Springer-Verlag, 1996.

[29] J. VanSanten, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, Springer-Verlag, 1996.

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

[31] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley and Sons, 1999.

[32] S. Narayanan and A. Alwan, editors, *Text to Speech Synthesis: New Paradigms and Advances*, Prentice-Hall Inc., 2004

[33] P. Taylor, *Text-to-Speech Synthesis*, Univ. of Cambridge, 2008.

Speech Recognition and Natural Language References

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

[35] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition-A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[36] C. H. Lee, F. K. Soong, and K. K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, Kluwer Academic Publishers, 1996.

[37] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1998.

[38] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[39] X. D. Huang, A. Acero, and H-W Hon, *Spoken Language Processing*, Prentice-Hall Inc., 2000

[40] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, Prentice-Hall Inc., 2000.

[41] S. E. Levinson, *Mathematical Models for Speech Technology*, John Wiley and Sons, 2005.

Speech Enhancement References

[42] P. Vary and R. Martin, *Digital Speech Transmission, Enhancement, Coding and Error Concealment*, John Wiley and Sons, 2006.

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

Audio Processing References

[44] H. Kars and K. Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, 1998.

[45] A. Spanias, T. Painter and V. Atti, *Audio Signal Processing and Coding*, John Wiley and Sons, 2006.