1

Introduction

1.1 WHAT IS SPEECH COMMUNICATION?

Speech communication is the transfer of information from one person to another via *speech*, which consists of variations in pressure coming from the mouth of a speaker. Such pressure changes propagate as *waves* through air and enter the ears of listeners, who decipher the waves into a received message. Human communication often includes gestures, which are not not part of speech, such as head and hand movements; such gestures, though normally part of face-to-face communication, are not considered in this book. The chain of events from the concept of a message in a speaker's brain to the arrival of the message in a listener's brain is called the *speech chain* [1]. The chain consists of a speech production mechanism in the speaker, transmission through a medium (e.g., air), and a speech perception process in the ears and brain of a listener.

In many applications of speech processing (italicized below), part of the chain is implemented by a simulation device. Automatic *synthesis* or generation of speech by algorithm (by computer) can simulate the speaker's role, except for generation of the original message, which usually comes in the form of a text (furnished by a computer user). In automatic speech or speaker *recognition*, an algorithm plays the listener's role in decoding speech into either an estimate of the underlying textual message or a hypothesis concerning the speaker's identity. Speech *coders* allow replacing the analog transmission medium (e.g., air or telephone lines) with a digital version, modifying the representation of the signal; in this way, speech can be efficiently stored and transmitted, often without noise problems and with enhanced security.

1.2 DEVELOPMENTS IN SPEECH COMMUNICATION

While many of the developments affecting speech communication have occurred in the last few decades, the basic tools for speech analysis are founded in mathematics, e.g., Fourier analysis, developed many decades ago. A basic understanding of how we produce speech has

existed for hundreds of years (e.g., mechanical speech synthesizers existed in the 1700s), but detailed knowledge of audio perception is fairly recent (e.g., Békésy's experiments on the basilar membrane in the 1940s). Modern speech research started in the 1930s, when the practical digital transmission method of *pulse-code modulation* (PCM) was developed and when a mechanical synthesizer called the Voder was demonstrated. The invention of the *sound spectrograph* in 1946 spurred much speech analysis work, since it allowed practical displays of the acoustic output of the vocal tract.

Viewing individual sounds or *phonemes* as composed of discrete, distinctive features originated in the 1950s and spurred development of electronic speech synthesizers, e.g., the Pattern Playback. More efficient digital speech coding in the form of delta modulation was developed at this time as well. Fant's benchmark work on speech acoustics appeared in 1960 [2], beginning a decade of much speech research, during which speech was first synthesized by computer and the important analysis techniques of the *cepstrum* and *linear prediction* were introduced.

The early 1970s saw development of time-adaptive speech coding as well as a big increase in speech recognition work, including the Advanced Research Projects Agency (ARPA) speech understanding project [3] and the use of dynamic programming in matching templates from different speech signals. *Digital signal processing* as a discipline saw much development [4]. In the late 1970s, more complex speech systems such as *subband* and *adaptive transform* coders appeared. Large-scale integrated circuits made their appearance in the form of one-chip speech synthesizers, and stochastic methods became accepted for speech recognition (e.g., *hidden Markov models*).

The major developments of the 1980s included single-chip digital signal processors, the use of *vector quantization* for low-rate speech coding, the search for better excitation models for speech synthesis (e.g., *multipulse excitation*), the use of auditory models in speech applications based on hearing experiments using speechlike stimuli, and the use of language models to aid speech recognition.

The 1990s have seen widespread acceptance of speech coders, synthesizers and recognizers, as computational power has continued to increase substantially while costs decrease. While the pace of major breakthroughs has slowed in recent years (e.g., the last significant new paradigm introduced in speech processing was the use of neural networks in the late 1980s), research continues unabated, because current speech products are far from ideal.

1.3 OUTLINE OF THE BOOK

The book is divided into two main parts: the first deals with the way **humans** generate and interpret speech, and the second examines how **machines** simulate human speech performance and how they code speech for efficient transmission. The first part is further subdivided into chapters on speech production, general audition, and speech perception. The second part has six chapters, three describing analysis, coding and enhancement of speech, one examining speech synthesis, and two on recognition (speech and speaker). The following sections briefly describe the different domains of scientific study found in speech communication research.

1.3.1 Production of Speech

Chapter 3 examines the first link in the speech chain: production. After a brief introduction, Section 3.2 discusses the organs of the *vocal tract* that a speaker uses to produce speech, from the viewpoint of their functions. The relationships between the positioning and motion of these organs and the sounds of a language (using English as the basic example) are the subject of Section 3.3. Since the speech wave or *acoustic signal* is of prime importance in all practical applications, Section 3.4 discusses how each sound can be simply described by a set of acoustic features observable in the speech signal's waveform or frequency *spectrum*.

Modeling aspects of human behavior during speech communication is of major interest, both to understand the speech process better and to suggest ways of simulating human speech tasks by machine. Thus, Sections 3.5 and 3.6 model vocal tract behavior in relation to time and frequency aspects of the speech signal. First, the tract is *modeled* by means of electrical circuits and transmission lines, to understand how the reflection of waves inside the tract causes different speech frequencies to be amplified or attenuated, depending on the shape of the tract. Then practical digital filters used in speech synthesis are introduced. The final two sections of Chapter 3 examine two major sources of difficulties in speech analysis: *coarticulation* and *intonation*; the former refers to variations in speech sounds due to context, and the latter denotes variations in the tone, length, and intensity of sounds.

1.3.2 Sound Perception

Aspects of human speech perception are divided into two chapters. Chapter 4 deals with the conversion of speech waves (and other sounds) into auditory nerve patterns, including elementary *auditory psychophysics*. Chapter 5 examines the relationships between the acoustic features of sounds and what listeners perceive. After a brief introduction, Section 4.2 discusses how the organs of the ear convert acoustic speech waves into electrical signals on the *auditory nerve*. Questions of how intense and long a sound must be to be heard are discussed in Section 4.3, along with aspects of *pitch* perception and the perceptual effects of one sound on another. While this section deals with simple sounds (e.g., clicks and tones), Section 4.4 extends these ideas to examine auditory responses to speech signals.

High-quality synthetic speech requires accurate modeling of aspects of the speech signal that are important perceptually. Since the identity of a sound in natural speech is signaled in a complex way through many redundant cues, Section 5.2 examines the difficulty of determining perceptually-important speech features when using synthetic speech stimuli. Section 5.3, which analyzes different speech perception models and theories, is followed by a summary of the results of perception experiments for vowels (Section 5.4) and consonants (Section 5.5). How intonation is used to segment continuous speech and highlight specific words is the subject of Section 5.7, preceded by a discussion of the utility of timing in speech. Miscellaneous aspects of speech perception (e.g., issues of adaptation, dichotic listening, distortions) end Chapter 5.

1.3.3 Speech Analysis

Whereas Chapters 3–5 focus on how humans utilize speech, the remaining chapters address applications of speech communication that typically involve digital computers. Chapter 6, an introduction to the application chapters, notes the key elements of automatic

speech analysis. Chapter 7 investigates how speech signals can be coded for efficient transmission, Chapter 8 explores ways to improve distorted speech, Chapter 9 examines how speech is generated synthetically, and Chapters 10 and 11 consider techniques for extraction of the message and of the speaker's identity, respectively, from a speech signal.

The basic representation of a speech signal in a digital computer requires limiting the spectral bandwidth of the signal (e.g., 0–4 kHz), sampling it at a certain corresponding rate (e.g., 8000 samples/s), and storing each sample with an adequate resolution, e.g., 12 bits (*binary digits*) each. Eliminating frequencies in speech above, say, 4 kHz causes a slight degradation in speech quality or naturalness, but has little effect on information content. For many communication applications, preserving the intelligibility of the speech is paramount, and some degradations in quality are acceptable. Besides *intelligibility* (being able to understand the speech message), certain information about the speaker (e.g., identity and mood) is usually important to retain in coded speech; often, low-rate coders preserve intelligibility while sacrificing such speaker information.

Digital signal representations of sufficient bit rate can be converted back into speech without significant loss of quality (other than the unavoidable loss of high frequencies in the original bandlimiting). From the perspective of information transmission, however, using almost 100,000 bits/s of speech is very wasteful because each second of speech typically contains 12 distinct sounds, from a inventory of about $32 = 2^5$ linguistic units called *phonemes*, which suggests an actual information rate of about 60 bit/s. Speech contains information other than the simple sequence of sounds, however, since listeners can infer speaker identity and emotion as well as assign a linguistic structure to each utterance. Nonetheless, simple speech coding can theoretically be improved by a factor of about 1000 by extracting appropriate features from the signal rather than using elementary sampling.

Chapter 6 examines the basic techniques of *parameter* and *feature* extraction from speech, which is of direct use to coding and recognition and of indirect use to synthesis. Since a speech signal changes its characteristics for each new sound, speech analysis must be performed on short *windowed* segments for which the vocal tract is assumed in most cases to be essentially fixed (Section 6.2). Certain relevant features (e.g., energy and periodicity) can be observed directly in the time-domain display of a speech signal (Section 6.3). To accurately distinguish sounds, however, requires a spectral analysis (Section 6.4).

The important technique of *linear predictive* (LP) analysis is the subject of a detailed Section 6.5, including: (a) the basic LP model in terms of the two traditional block analysis methods (autocorrelation and covariance), (b) the relationship of spectral modeling resolution to the order of the LP model, (c) adaptive and lattice filters, and (d) the effects of the size of the analysis window.

Cepstral analysis, a general signal processing technique with specific application to speech (mostly recognition), is examined in Section 6.6. Other recent developments in analysis (e.g., wavelets) are explored in Section 6.7. The difficult task of pitch estimation is described in Section 6.8. Issues of *robustness* against distortions are the subject of Section 6.9. Chapter 6 ends with a discussion of how extracted features can be smoothed in time, to represent the speech signal more efficiently while still permitting adequate reconstruction of the speech.

1.3.4 Speech Coding

Solving the problem of reducing the bit rate of a speech representation, while preserving the quality of speech reconstructed from such a representation, continues in

Chapter 7. It builds on the fundamental procedures of Chapter 6 and addresses the tradeoffs of quality, coding rate, and algorithmic complexity. Fundamentals of linear and logarithmic *quantization* of speech are described in Section 7.2, noting how coding rate can be reduced by using basic amplitude statistics of speech signals. Section 7.3 gives an overview of the aspects of speech that are exploited in coders that reconstruct the signal sample by sample. Section 7.4 describes measures to evaluate speech quality. Sections 7.5–7.8 describe waveform coders that operate directly on the time signal. Various coding schemes exploit different properties of speech: (a) its average intensity changes slowly with respect to the sampling rate (Section 7.5), (b) it has primarily low-frequency energy and is adequately parameterized in the short term by simple models (Section 7.6), (c) it is often periodic (Section 7.7), and (d) its spectral components vary in perceptual importance (Section 7.8).

Section 7.9 deals with many aspects of linear predictive coding (LPC), ranging from simple differential PCM to the standard LPC *vocoding* method, which separates two sources of information in the speech signal (the *excitation* and the *frequency response* of the vocal tract) for efficient manipulation and transmission. Among the topics covered are: how the basic LPC parameters can be transformed into equivalent but more efficient representations, how the nonstationarity of speech affects the transmission rate of LPC, how the basic *all-pole* LPC model can be enhanced at the cost of extra complexity, and how we can trade transmission rate for quality via use of more complex excitation models.

Section 7.10 describes filtering approaches to coding (where more perceptually important frequency ranges are assigned more bits), including a *frequency-transform* method (which directly codes a spectrum) and methods that code speech harmonics directly. While LPC dominates vocoding methods, alternative vocoders (e.g., channel vocoding) are noted in Section 7.11.

Most speech coders transmit time or frequency samples as independent (scalar) parameters, but coding efficiency can be enhanced by eliminating redundant information within blocks of parameters and transmitting a single index code to represent the entire block, i.e., *vector quantization* (Section 7.12). Section 7.13 examines network aspects of speech transmission: when many speech signals are mixed with data and sent over a network where traffic varies with time, tradeoffs of speech quality and network availability arise.

1.3.5 Speech Enhancement

Chapter 8 examines how to increase the quality of degraded speech signals. Speech is often distorted by background noise (or other speech) and/or by poor transmission conditions. Various filtering or other processing techniques can reduce the distortion effects, rendering the signal easier to listen to. If noisy speech can be captured via several microphones, its intelligibility can even be raised. Sections 8.1 and 8.2 give an introduction to the problem, and Section 8.3 describes the types of interfering sounds we must deal with. The major enhancement methods are summarized in Section 8.4, and then discussed in more detail: (a) subtracting estimates of noise from noisy spectral amplitudes (Section 8.5), (b) filtering out the distortion (e.g., *adaptive noise cancellation*) (Section 8.6), (c) suppressing energy between speech harmonics (*comb filtering*) (Section 8.7), and (d) resynthesis of the speech after vocoder modeling (Section 8.8).

1.3.6 Speech Synthesis

Chapter 9 examines automatic generation of speech. Section 9.2 introduces the major aspects of speech synthesis: the size of the stored speech unit to be concatenated, the method of synthesis (which usually follows a coding method from Chapter 7), and the difference in output quality between *voice response* systems of very limited vocabulary and *text-to-speech* systems that accept unrestricted text input. Section 9.3 represents the bulk of Chapter 9, giving details for articulatory, formant, and LPC synthesizers. The three major reasons for limits on synthetic speech quality are discussed: simplistic models for excitation, intonation, and spectral time behavior. Section 9.4 examines the difficulty of simulating natural intonation. Section 9.5 discusses how to simulate different speaking voices, and Section 9.6 notes research progress for languages other than English. Performance in speech recognizers and in many speech coders can be measured objectively (i.e., via the percentage of words recognized correctly, or via signal-to-noise ratios); however, there is no objective way to evaluate speech synthesizers (Section 9.7). Chapter 9 ends with a section on specialized hardware for synthesis.

1.3.7 Speech and Speaker Recognition

The other side of human–machine communication via speech is automatic recognition, where either the textual message (Chapter 10) or the speaker's identity (Chapter 11) is extracted or verified from the speech signal. Chapter 10 starts with the view of speech identification as a *pattern recognition* task, where the input signal is reduced to a set of parameters or features, which in turn are compared to *templates* or models in memory to find the one with the best match. Sections 10.3 and 10.4 detail the initial stages of recognition: normalizing the signal, and extracting parameters and features from the data. When comparing an N-parameter model of an unknown input utterance (of, for example, a word) with a stored model of a known utterance, *distance measures* should reflect how well each parameter separates models for different words in N-dimensional space (Section 10.4).

Section 10.5 looks at the problems of comparing utterance representations to evaluate their similarity. A lengthy Section 10.6 examines the many sources of variability in speech, and how recognizers accommodate this variability. As an example, consider speech spoken with different durations or speaking rates: linearly normalizing all templates to the same duration often leads to poor comparisons, so nonlinear normalization through hidden Markov models (HMMs) or *dynamic time warping* is often used. State networks are commonly used in recognition as models of the sequence of acoustic events in speech. Each state in a network represents an acoustic event (e.g., a sound, a word), and the transition from state A to state B is labeled with the probability that event B follows event A in the sentences or words of the vocabulary. Section 10.7 continues the discussion of speech variability, in noting how recognizers can *adapt* to variability due to different speakers and recording conditions.

The use of statistics of sequences of words in text occurs in *language models* for speech recognition (Section 10.8). As constraints are relaxed on speakers (e.g., allowing use of wider vocabularies or speaking without frequent pauses), issues of computation time and memory become significant. This leads to ways to optimize the large *search space* in speech recognition.

Discriminating between phonetically similar words is often difficult for many recognizers. Nonlinear techniques using *artificial neural networks*, based on simple models of the human brain, are capable of more precise discrimination than HMMs, but do poorly on

handling temporal variations (Section 10.10). Future systems will likely integrate the current stochastic recognition methods (e.g., HMMs) with *expert system* approaches (Section 10.11). Chapter 10 ends with a brief section describing available commercial recognizers.

Speaker identification is the subject of Chapter 11, whose first two sections briefly introduce the problem and distinguish between *recognition* (selecting one out of N known speakers) and *verification* (deciding whether a speaker is who he claims to be). The methodology of speaker identification, examined in Section 11.3, has many similarities and some differences compared with speech recognition techniques. If the speaker utters a text known to the system, the methods can be very similar; however, if the training and testing utterances use different texts, simple template matching is not possible. Section 11.4 describes common features extracted from speech signals that help distinguish speakers; some are based on physiological traits of vocal tracts, while others measure dynamic variations such as speaking style. Section 11.5 investigates the design of speaker identification systems, e.g., how data are collected and how the use of telephone speech affects system performance. Section 11.6 describes the related tasks of identifying the language being spoken and the accent of the speaker. The chapter ends with a comparison of how well humans and machines can identify speakers.

1.4 OTHER TOPICS

While the book surveys most major aspects of speech communication, of necessity certain areas are emphasized at the expense of others. For example, the developmental aspects of speech production and perception are not discussed. The interested reader is referred to reviews on biological development [5] as well as on phonetic and linguistic development [6, 7]. Also omitted for space reasons are discussions of impaired production and perception, that is, how the human speech mechanisms function in people with speech and hearing organs that are abnormal due to disease or injury [8].