Preface

Speech signal processing has been a dynamic and constantly developing field for more than 70 years. The earliest speech processing systems were analog systems. They included, for example, the Voder (voice demonstration recorder) for synthesizing speech by manual controls, developed by Homer Dudley and colleagues at Bell Labs in the 1930s and demonstrated at the 1939 New York World's Fair; the channel vocoder or voice coder, also developed in the 1930s by Homer Dudley at Bell Labs; the sound spectrograph, a system for displaying time-varying speech patterns in time and frequency, developed by Koenig and his colleagues in the 1940s at Bell Labs; and early systems for recognizing spoken words, developed in research labs throughout the world in the 1950s.

Speech processing was the driving force for many of the early developments in the broader field of digital signal processing (DSP) whose roots began to take hold in the 1960s. During this period, pioneering researchers such as Ben Gold and Charlie Rader at MIT Lincoln Labs, and Jim Flanagan, Roger Golden, and Jim Kaiser at Bell Labs began to study methods for the design and application of digital filters for use in simulations of speech processing systems. With the technical disclosure of the fast Fourier transform (FFT) algorithm by Jim Cooley and John Tukey in 1965, and its subsequent widespread application in the areas of fast convolution and spectral analysis, the shackles and limitations of analog technology rapidly fell away and the field of digital speech processing emerged and took on a clear identity. The authors of the present book (LRR and RWS) worked closely together at Bell Labs during the period from 1968 to 1974, when many fundamental advances in DSP occurred. When RWS left Bell Labs in 1975 for an academic position at Georgia Tech, the field of digital speech processing had developed so much that we decided that it was an appropriate time to write a textbook on methods and systems for digital processing of speech signals. We were confident that the theory of digital speech processing had developed sufficiently (by 1976) that a carefully written textbook would effectively serve as both a textbook for teaching the fundamentals of digital speech processing, and as a reference textbook for practical system design of speech processing systems for the foreseeable future. The resulting textbook, entitled "Digital Processing of Speech Signals" was published by Prentice-Hall Inc. in 1978. In his new position in academia, RWS was able to create one of the first graduate courses in digital speech processing based on this textbook, while LRR continued basic digital speech processing research at Bell Labs. (LRR had a 40-year career with AT&T Bell Labs and AT&T Labs Research after which he also joined academia, jointly teaching at both Rutgers University and the University of California in Santa Barbara, in 2002. RWS had a 30-year career at Georgia Tech and he joined Hewlett Packard Research Labs in 2004.)

The goals of the 1978 textbook were to present the fundamental science of speech together with a range of digital speech processing methods that could be used to create powerful speech signal processing systems. To a large extent, our initial goals were met. The original textbook has served as intended for more than 30 years, and, to our delight, it is still widely used today in teaching undergraduate and graduate courses in digital speech processing. However, as we have learned from our personal experiences in teaching speech processing courses over the past two decades, while its fundamentals remain sound, much of the material in the original volume is greatly out-of-touch with modern speech signal processing systems, and entire areas of current interest are completely missing. The present book is our attempt to correct these weaknesses. In approaching the daunting task of unifying current theory and practice of digital speech processing, we found that much of the original book remained true and relevant, so we had a good starting point for this new textbook. Furthermore, we learned from both practical experience in research in speech processing and from our teaching experiences that the organization of the material in the 1978 volume, although basically sound, was just not suitable for understanding modern speech processing systems. With these weaknesses in mind, we adopted a new framework for presenting the material in this new book, with two major changes to the original framework. First we embraced the concept of the existence of a hierarchy of knowledge about the subject of digital speech processing. This hierarchy has a base level of fundamental scientific and engineering speech knowledge. The second level of the hierarchy focuses on representations of the speech signal. The original book focused primarily on the bottom two levels but even then some key topics were missing from the presentations. At the third level of the hierarchy are algorithms for manipulating, processing, and extracting information from the speech signal that are based on technology and science from the two lower layers. At the top of the hierarchy (i.e., the fourth level) are the applications of the speech processing algorithms, along with techniques for handling problems in speech communication systems. We have made every attempt to follow this new hierarchy (called the speech stack

in Chapter 1) in presenting the material in this new book. To that end, in Chapters 2–5, we concentrate on the process of building a firm foundation at the bottom layer of the stack, including topics such as the basics of speech production and perception, a review of DSP fundamentals, and discussions of acoustic–phonetics, linguistics, speech perception, and sound propagation in the human vocal tract. In Chapters 6–9 we develop an understanding of how various digital speech (short-time) representations arise from basic signal processing principles (forming the second layer of the speech stack). In Chapter 10 we show how to design speech algorithms that are both reliable and robust for estimating a range of speech parameters of interest (forming the basis for the third layer of the speech stack). Finally, in Chapters 11–14 we show how our knowledge from the lower layers of the speech stack can be used to enable design and implementation of a range of speech applications (forming the fourth layer of the speech stack). The second major change in structure and presentation of the new book is in the realization that, for maximal impact in teaching, we had to present the material with an equal focus on three areas of learning of new ideas, namely theory, concept, and implementation. Thus for each fundamental concept introduced in this book, the theory is explained in terms of well-understood DSP concepts; similarly the understanding of each new concept is enhanced by providing simple interpretations of the mathematics and by illustrating the basic concepts using carefully-explained examples and associated graphics; and finally, the implementation of new concepts based on understanding of fundamentals is taught by reference to MATLAB code for specific speech processing operations (often included within the individual chapters), along with extensive and thoroughly documented MATLAB exercises in the (expanded) homework problem section of each chapter. We also provide a course website with all the material needed to solve the various MATLAB exercises, including specialized MATLAB code, access to simple databases, access to a range of speech files, etc. Finally we provide several audio demonstrations of the results of a range of speech processing systems. In this manner, the reader can get a sense of the resulting quality of the processed speech for a range of operations on the speech signal. More specifically, the organization of this new book is as follows. Chapter 1 provides a broad brush introduction to the area of speech processing, and gives a brief discussion of application areas that are directly related to topics discussed throughout the book. Chapter 2 provides a brief review of DSP with emphasis on a few key concepts that are pervasive in speech processing systems:

1. conversion from the time domain to the frequency domain (via discrete-time Fourier transform methods)
2. understanding the impact of sampling in the frequency domain (i.e., aliasing in the time domain)
3. understanding the impact of sampling (both downsampling and upsampling) in the time domain, and the resulting aliasing or imaging in the frequency domain Following the review of the basics of DSP technology, we move on to a discussion of the fundamentals of speech production and perception in Chapters 3 and 4. These chapters, together with Chapters 2 and 5, comprise the bottom layer of the speech stack. 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 communication with an analysis of speech perception processes, beginning with a discussion of how speech is processed in the ear, and ending with a discussion of methods for the transduction of sound to neural signals in the auditory neural pathways leading to the brain. We briefly discuss several possible ways of embedding knowledge of speech perception into an auditory model that can be utilized in speech processing applications. Next, in Chapter 5, we complete our discussion of fundamentals with a discussion of issues of sound propagation in the human vocal tract. We show that uniform lossless tube approximations to the vocal tract have resonance structures elucidating the resonant (formant) frequencies of speech. 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 a specified radiation characteristic for transmission of sound at the lips.
We devote the next four chapters of the book to digital speech signal representations (the second layer in the speech stack), with separate chapters on each of the four major representations. We begin, in Chapter 6, with the temporal model of speech production and show how we can estimate basic time-varying properties of the speech model from simple time-domain measurements. In Chapter 7 we show how the concept of short-time Fourier analysis can be applied to speech signals in a simple and consistent manner such that a completely 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 used in a wide range of applications, depending on the nature of the information that is required for further processing. In Chapter 8

we describe a homomorphic (cepstrum) representation of speech where we use the property that a convolutional signal (such as speech) can be transformed into a set of additive components. With the understanding that a speech signal can be well represented as the convolution of an excitation signal and a vocal tract system, we see that the speech signal is well suited to such an analysis. 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 linear predictor (with minimized mean-squared error) that optimally matches the speech signal over a given time duration. Chapter 10, which represents the third layer in the speech stack, deals with using the signal processing representations and fundamental knowledge of the speech signal, presented in the preceding chapters, as a basis for measuring or estimating properties and attributes of the speech signal. Here we show how the measurements of short-time (log) energy, short-time zero crossing rates, and short-time autocorrelation can be used to estimate basic speech attributes such as whether the signal section under analysis is speech or silence (background signal), whether a segment of speech represents voiced or unvoiced speech, the pitch period (or pitch frequency) for segments of voiced speech signals, the formants (vocal tract resonances) for speech segments, etc. For many of the speech attributes, we show how each of the four speech representations can be used as the basis of an effective and efficient algorithm for estimating the desired attributes. Similarly we show how to estimate formants based on measurements from two of the four speech representations. Chapters 11–14, representing the top layer in the speech stack (speech applications), deal with several of the major applications of speech and audio signal processing technology. These applications are the payoffs of understanding speech and audio technology, and they represent decades of research on how best to integrate various speech representations and measurements to give the best performance for each speech 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 perceptual masking criteria; Chapter 13 deals with building text-to-speech 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. Our goal in these chapters is to provide up-to-date examples but not to be exhaustive in our coverage. Entire textbooks have been written on each of these application areas.

The material in this book can be taught in a one-semester course in speech processing, assuming that students have taken a basic course on DSP. In our own teaching, we emphasize the material in Chapters 3–11, with selected portions of the material in the remaining chapters being taught to give students a sense of the issues in audio coding, speech synthesis, and speech recognition systems. To aid in the teaching process, each chapter contains a set of representative homework problems that are intended to reinforce the ideas discussed in each chapter. Successful completion of a reasonable percentage of these homework problems is essential for a good understanding of the mathematical and theoretical concepts of speech processing, as discussed earlier. However, as the reader will see, much of speech processing is, by its very nature, empirical. Hence we have chosen to include a series of MATLAB exercises in each chapter (either within the text or as part of the set of homework problems) so as to reinforce the student's understanding of the basic concepts of speech processing. We have also provided on the course website (http://www.pearsonhighered/Rabiner.com) which will be updated with new material from time to time—the required speech files, databases, and MATLAB code required to solve the MATLAB exercises, along with a series of demonstrations of a range of speech processing concepts.

ACKNOWLEDGEMENTS

Throughout our careers in speech processing, we have both been extremely fortunate in our affiliations with outstanding research and academic institutions that have provided stimulating research environments and also encouraged the sharing of knowledge. For LRR, these have been Bell Laboratories, AT&T Laboratories, Rutgers University, and the University of California at Santa Barbara, and for RWS, it was Bell Laboratories, Georgia Tech ECE, and Hewlett-Packard Laboratories. Without the support and encouragement of our colleagues and the enlightened managers of these fine institutions, this book would not exist. Many people have had a significant impact, both directly and indirectly, on the material presented in this book, but our biggest debt of gratitude is to Dr. James L. Flanagan, who served as both supervisor and mentor for both of us at various points in our careers. Jim has provided us with an inspiring model of how to conduct research and how to present the results of research in a clear and meaningful way. His

influence on both this book, and our respective careers, has been profound. Other people with whom we have had the good fortune to collaborate and learn from include our mentors Prof. Alan Oppenheim and Prof. Kenneth Stevens of MIT and our colleagues Tom Barnwell, Mark Clements, Chin Lee, Fred Juang, Jim McClellan and Russ Mersereau, all professors at Georgia Tech.All of these individuals have been both our teachers and our colleagues, and we are grateful to them for their wisdom and their guidance over many years. Colleagues who have been involved directly with the preparation of this book include Dr. Bishnu Atal, Prof. Victor Zue, Prof. Jim Glass, and Prof. Peter Noll, each of whom provided insight and technical results that strongly impacted a range of the material presented in this book. Other individuals who provided permission to use one or more figures and tables from their publications include Alex Acero, Joe Campbell, Raymond Chen, Eric Cosatto, Rich Cox, Ron Crochiere, Thierry Dutoit, OdedGhitza, Al Gorin, Hynek Hermansky, Nelson Kiang, Rich Lippman, Dick Lyon, Marion Macchi, John Makhoul, Mehryar Mohri, Joern Ostermann, David Pallett, Roberto Pieraccini, Tom Quatieri, Juergen Schroeter, Stephanie Seneff, Malcolm Slaney, Peter Vary, and Vishu Viswanathan. We would also like to acknowledge the support provided by Lucent-Alcatel, IEEE, the Acoustical Society of America, and the House-Ear Institute for granting permission to use several figures and tables reprinted from published or archival material. Also, we wish to acknowledge the individuals at Pearson Prentice Hall who helped make this book come to fruition. These include Andrew Gilfillan, acquisitions editor, Clare Romeo, production editor, William Opaluch, editorial assistant. I would also like to acknowledge Maheswari PonSaravanan of TexTech International who was the lead individual in the copy-editing and page proof process. Finally, we thank our spouses Suzanne and Dorothy for their love, patience, and support during the seemingly never ending task of writing of this book.

Lawrence R. Rabiner and RonaldW. Schafer