Foreword

Recognition and understanding of spontaneous unrehearsed speech remains an elusive goal. To understand speech, a human considers not only the specific information conveyed to the ear, but also the context in which the information is being discussed. For this reason, people can understand spoken language even when the speech signal is corrupted by noise. However, understanding the context of speech is, in turn, based on a broad knowledge of the world. And this has been the source of the difficulty and over forty years of research.

It is difficult to develop computer programs that are sufficiently sophisticated to understand continuous speech by a random speaker. Only when programmers simplify the problem—by isolating words, limiting the vocabulary or number of speakers, or constraining the way in which sentences may be formed—is speech recognition by computer possible.

Since the early 1970s, researchers at ATT, BBN, CMU, IBM, Lincoln Labs, MIT, and SRI have made major contributions in Spoken Language Understanding Research. In 1971, the Defense Advanced Research Projects Agency (Darpa) initiated an ambitious five-year, $15 million, multisite effort to develop speech-understanding systems. The goals were to develop systems that would accept continuous speech from many speakers, with minimal speaker adaptation, and operate on a 1000-word vocabulary, artificial syntax, and a constrained task domain. Two of the systems, Harpy and Hearsay-II, both developed at Camegie-Mellon University, achieved the original goals and in some instances surpassed them.

During the last three decades I have been at Carnegie Mellon, I have been very fortunate to be able to work with many brilliant students and researchers. Xuedong Huang, Alex Acero and Hsiao-Wuen Hon were arguably among the outstanding researchers in the speech group at CMU. Since then they have moved to Microsoft and have put together a world-class team at Microsoft Research. Over the years, they have contributed with standards for building spoken language understanding systems with Microsoft's SAPI/SDK family of products, and pushed the technologies forward with the rest of the community. Today, they continue to play a premier leadership role in both the research community and in industry.

The new book "Spoken Language Processing" by Huang, Acero and Hon represents a welcome addition to the technical literature on this increasingly important emerging area of Information Technology. As we move from desktop PCs to personal digital assistants (PDAs), wearable computers, and Internet cell phones, speech becomes a central, if not the only, means of communication between the human and machine! Huang, Acero, and Hon have undertaken a commendable task of creating a comprehensive reference manuscript covering theoretical, algorithmic and systems aspects of spoken language tasks of recognition, synthesis and understanding.

The task of spoken language communication requires a system to recognize, interpret, execute and respond to a spoken query. This task is complicated by the fact that the speech signal is corrupted by many sources: noise in the background, characteristics of the microphone, vocal tract characteristics of the speakers, and differences in pronunciation. In addition the system has to cope with non-grammaticality of spoken communication and ambiguity of language. To solve the problem, an effective system must strive to utilize all the available sources of knowledge, i.e., acoustics, phonetics and phonology, lexical, syntactic and semantic structure of language, and task specific context dependent information.

Speech is based on a sequence of discrete sound segments that are linked in time. These segments, called phonemes, are assumed to have unique articulatory and acoustic characteristics. While the human vocal apparatus can produce an almost infinite number of articulatory gestures, the number of phonemes is limited. English as spoken in the United States, for example, contains 16 vowel and 24 consonant sounds. Each phoneme has distinguishable acoustic characteristics and, in combination with other phonemes, forms larger units such as syllables and words. Knowledge about the acoustic differences among these sound units is essential to distinguish one word from another, say "bit" from "pit."

When speech sounds are connected to form larger linguistic units, the acoustic characteristics of a given phoneme will change as a function of its immediate phonetic environment because of the interaction among various anatomical structures (such as the tongue, lips, and vocal chords) and their different degrees of sluggishness. The result is an overlap of phonemic information in the acoustic signal from one segment to the other. For example, the same underlying phoneme "t" can have drastically different acoustic characteristics in different words, say, in "tea," "tree," "city," "beaten." and "steep." This effect, known as coarticulation, can occur within a given word or across a word boundary. Thus, the word "this" will have very different acoustic properties in phrases such as "this car" and "this ship."

This manuscript is self-contained for those who wish to familiarize themselves with the current state of spoken language systems technology. However a researcher or a professional in the field will benefit from a thorough grounding in a number of disciplines such as:

- *signal processing*: Fourier Transforms, DFT, and FFT.
- *acoustics*: Physics of sounds and speech, models of vocal tract.
- *pattern recognition*: clustering and pattern matching techniques.
- *artificial intelligence*: knowledge representation and search, natural language processing.
- *computer science*: hardware, parallel systems, algorithm optimization.
- *statistics*: probability theory, hidden Morkov models, dynamic programming and
- *linguistics*: acoustic phonetics, lexical representation, syntax, and semantics.

A newcomer to this field, easily overwhelmed by the vast number of different algorithms scattered across many conference proceedings, can find in this book a set of techniques that the Huang, Acero and Hon have found to work well in practice. This book is unique in that it includes both the theory and implementation details necessary to build spoken language systems. If you were able to assemble all of the individual material that are covered in the book and put it on a shelf it would be several times larger than this volume, and yet you would be missing vital information. You would not have the material that is in this book that threads it all into one story, one context. If you need additional resources, the authors include references to get that additional detail. This makes it very appealing both as a textbook as well as a reference book for practicing engineers. Some readers familiar with some topic may decide to skip a few chapters; others may want to focus in other chapters. As such, this is not a book that you will pick up and read from cover to cover, but one you will keep near you as long as you work in this field.

Raj Reddy