COMPUTER SPEECH PROCESSING

Fallside, Frank and Woods, William A. (editors)

London: Prentice-Hall International, 1985,
xxi+506 pp.
ISBN 0-13-163841-6

Reviewed by
John C. Thomas
NYNEX Science and Technology

There are three major speech technologies: *Speech recognition* involves having the computer translate from human voice to data (typically text). *Speech synthesis* involves translating from data (typically text) to speech. *Speech coding* deals with the efficient storage and transmission of voice communication. Using the computer to process speech in these ways shows promise of considerable economic impact. For instance, perfection of these technologies means that every telephone worldwide potentially becomes a terminal to every computer database. In addition, the attempt to perfect these technologies provides tremendous insight into human speech. This is so even if one takes a purely "engineering" approach; that is, even when no explicit attempt is made to model how people accomplish these tasks.

Computer Speech Processing, edited by Frank Fallside and William A. Woods, is a useful book in a field that should be of considerable interest to computational linguists. It contains several extremely strong chapters. Peter Ladefoged's article should be required reading for every engineer who might someday work in speech technology. Bishnu Atal does an excellent job of covering linear predictive coding of speech including his own recent important advance, multipulse. In this type of technical material, though, errors become very confusing. For instance, "The rms prediction error decreases fast for small values of p but decreases [more slowly?] for higher values of p" (p. 92).

The chapter by Jonathan Allen is a good conceptual introduction to speech synthesis, and the chapter by Stella broadens our understanding of the issues and presents a historical perspective.

We may roughly divide the speech synthesis problem into two parts: translating from text into a series of control signals; and, given a "hand-built" set of control signals, building a mechanism for producing perfectly intelligible, natural-sounding speech. John Holmes has essentially solved the second half of this problem, and his chapter reveals many details of his excellent system. The first half of the problem is more oriented toward linguistic problems and remains unsolved. In fact, a *complete* solution probably subsumes the solution to the general artificial intelligence problem of comprehension. For example, how can one perfectly apply correct emphatic stress without comprehending? Issues of intent, content, and anaphora quickly become involved when one attempts to generate perfect prosody. However, the more modest goal of reading a sentence as a native speaker might who was unaware both of the meanings of the open-class words and of the context, is slowly being approached. *Computer Speech Processing* provides a good introduction to the issues, which, I believe, should be of considerable interest to readers of this journal.

The book as a whole also provides a good introduction to the issues of speech recognition. However, the story here is much less clear to the reader. Approaches to speech recognition vary greatly in how much they try to do things "the way people do". In fact, various approaches are not just on a continuum, because there are a number of separate decision points where one can model, to a greater or lesser degree, human mechanisms. At the "front-end", there are a number of ways to do the preliminary analysis of the speech signal. Fallside's chapter illustrates some of the methods based on frequency analysis. I find the chapter problematic, however. Most fundamentally, it is not clear who the intended reader is, nor the purpose. I do not believe it is sufficient for the digital signal processing engineer who wants to build a state-of-the-art "front-end". Yet, I find it hard to believe that even a mathematically sophisticated linguist who does not already specifically know digital signal processing will be able to learn these techniques from the presentation. This is further complicated by some apparent errors, e.g., the captions in Figures 3.11 and 3.12. As an introduction, I also believe the chapter needs some heavier caveats. One can easily gain the impression that these methods work universally; e.g., that "pitch-trackers" are always accurate. In fact, there is increasing use of "ear models" in speech recognition front-ends. Bladon's chapter gives one example of how normalization based on an ear model simplifies the task of dealing with sex differences in vowel space. The more linguistically oriented Chapters 11 through 16 will be of substantial interest to readers of this journal and deal primarily with the "back-end" of speech recognition, namely using higher-order knowledge to improve recognition.

In summary, the book, like most edited volumes, is somewhat uneven as to audience, goals, and style but contains much that is worthwhile in a field of substantial practical and theoretical importance.

John Thomas received his PhD in experimental psychology from the University of Michigan in 1971. He is director of the Artificial Intelligence Laboratory at NYNEX. His R&D activities include speech technology, expert systems, human-computer interaction, and machine vision. Thomas's address is: NYNEX Science and Technology, 500 Westchester Avenue, White Plains, NY 10604.