Book Reviews

Progress in Speech Synthesis

Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg (editors)
(AT&T Labs—Research)

New York, Springer-Verlag, 1997,
xxii+598 pp and CD-ROM; hardbound,
ISBN 0-387-94701-9, $129.00

Reviewed by
Briony Williams
University of Edinburgh

1. Overview

This is a weighty book in more senses than one. At nearly 600 pages (plus an accompanying CD-ROM), it has the space to range very widely over the field of speech synthesis. The chapters comprise a selection of the papers presented at the Second ESCA/IEEE/AAAI Workshop on Text-to-Speech Synthesis in 1994 at Mohonk, New Jersey, U.S.A. (which was organized by the editors of this book). The work represented covers many laboratories in several countries. This makes for a snapshot of the state of speech synthesis research across the globe rather than at only one site.

There are eight sections, each preceded by one or two "section introduction" chapters. Including the latter, there are a total of 46 chapters. The eight sections are as follows:

1. Signal processing and source modelling
2. Linguistic analysis
3. Articulatory synthesis and visual speech
4. Concatenative synthesis and automated segmentation
5. Prosodic analysis of natural speech
6. Synthesis of prosody
7. Evaluation and perception
8. Systems and applications

The fact that a CD-ROM is included with the book greatly increases the usefulness of the text, since the speech samples referred to can be played out as one reads. This is preferable even to making speech files available at a website, since as everyone knows, download times can be long and servers can be temporarily unavailable.

The standard of editing at times leaves something to be desired. There are various slips in punctuation and spelling, and also some infelicities of style, due no doubt to the fact that many papers are written by speakers of a language other than English. In

such a large volume, a few slips may be expected, but the overall impression is that not enough attention has been paid to the details of spelling and style.

2. Detailed Comments

The chapters reflect the fact that most research in speech synthesis is based on American English. However, many chapters describe aspects of systems developed for other languages or varieties, such as German, Japanese, French, Swedish, Italian, Mandarin, and British English. In addition, there is some discussion of modules or whole systems for language-independent use, such as the Bell Labs multilingual system architecture (Sproat and Olive), and data-oriented grapheme-to-phoneme conversion (Daelemans et al.). The inclusion of a range of languages is to be welcomed, not least because the use of new languages has the potential to introduce interesting new research problems to the field.

The growing subfield of prosody and its application to speech synthesis is rightly given a prominent place, with four section-introduction chapters (for two sections) and ten contributed chapters. In addition, the chapter entitled "Speech models and speech synthesis" by Mary Beckman (in the section "Articulatory synthesis and visual speech") provides a very lucid and thorough presentation of the symbiotic relationship between basic research in speech science and developments in speech synthesis research, with particular reference to F0 and temporal parameters. Many readers with a primary interest in prosody for speech synthesis might find this the best chapter to begin their reading with, presenting as it does the necessary overview within which individual chapters may be placed.

Concatenative synthesis receives much more attention than formant synthesis, reflecting the current direction of the field. The section on this theme ("Concatenated synthesis and automated segmentation") would provide a graduate student with an excellent insight into current research problems that are still far from solved. As the section introduction by Joseph Olive notes about the concatenative method: "So far, this has not produced a consensus, but rather a diversity of approaches." This comment could well apply to the book as a whole, which succeeds admirably in displaying the range of approaches on offer.

Occasional chapters are a little too skimpy and leave the reader unsatisfied. This is probably unavoidable in a book that seeks to cover the entire range of speech synthesis research in one volume. For example, the chapter "A structured way of looking at the performance of text-to-speech systems" (Pols and Jekosch) presents the method in brief outline, and might have benefited from more detail and from a description of an implementation of the method. Likewise, the chapter "A modular architecture for multilingual text-to-speech" (Sproat and Olive) provides little more than a quick tour through the structure of a working multilingual synthesis system. In the case of the latter, much fuller details are available elsewhere (in Sproat 1998) and its relatively parsimonious appearance here reflects the fact that the present volume aims to cover a wide spread of systems rather than concentrating on one alone.

3. Summary

The work reported covers many different laboratories, countries, and theoretical interests. For this reason, it is useful as a general introduction to the major research issues for a reader who has a little background in linguistics, computing, or signal processing. It therefore has a different readership from the more recent comparable book *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (Sproat 1998). The lat-

ter centres on research at one laboratory in some detail, and focuses in particular on the Bell Labs multilingual speech synthesizer (where common algorithms are used for multiple languages). The material in the present book (which derives from a 1994 workshop) is of less-recent origin than that of Sproat. However, this is not necessarily a disadvantage for the reader whose aim is to learn about the background to current research issues and the various kinds of method used to address them.

In short, this book would be a valuable addition to the library of any researcher in speech synthesis. It would also be of use to a theoretical linguist seeking to learn how phonological theories have been implemented in working systems, and the nature of the problems involved. The inclusion of a CD-ROM with speech examples is a great asset in a field where output speech quality is the measure of success. The section introductions alone would form a useful quick tutorial for those needing only a taste of the field. But the individual chapters offer an astonishingly wide range of subjects, presented by active researchers, and will repay close study.

Reference

Sproat, Richard, editor. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Dordrecht: Kluwer.

Briony Williams has been active in speech technology research since 1983, with an initial training in phonetics and linguistics. Her interests lie mainly in concatenative speech synthesis and in speech database annotation and analysis. She has worked on both English and Welsh speech systems, and developed the first Welsh speech synthesizer. Williams's address is: Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, UK; e-mail: briony@cstr.ed.ac.uk.