Multilingual Text-to-Speech Synthesis: The Bell Labs Approach

Richard Sproat (editor)
(Bell Laboratories, Lucent Technologies)

Dordrecht: Kluwer Academic
Publishers, 1998, xxvi+300 pp;
hardbound, ISBN 0-7923-8027-4,
$110.00, £72.75, Dfl 240.00

Reviewed by
Douglas O'Shaughnessy
Université du Québec

This collection is a very welcome addition to the literature on automatic speech synthesis, also known as "text-to-speech" (TTS). It has been more than a decade since a comprehensive, edited collection of chapters on this topic has been published. Because much has changed in TTS research over the last ten years, this text will prove very useful to workers in the field. Along with Dutoit's recent book (*An Introduction to Text-to-Speech Synthesis* [1997], also published by Kluwer), it is essential reading for anyone serious about TTS research. (Other recent works have included chapters on TTS, but their main focus has been on other aspects of speech processing, and thus the TTS details there are far fewer. While Bailly and Benoit's collection *Talking Machines* [1992] has synthesis as its sole focus, it derives from numerous presentations at a workshop and suffers accordingly due to its many uneven short chapters.) A more direct comparison can be made to Allen, Hunnicutt, and Klatt's book on MITalk (1987), especially as both that book and the one currently under review describe in detail specific TTS systems developed at two of the major research centres involved in speech synthesis over the years. As the foreword of the present book notes, the MITalk system was largely based on morphological decomposition of words and synthesis via a Klatt formant architecture, while the modular Bell system is distinguished by its regular relations for text analysis, its use of concatenation of diphone units, and its emphasis on the importance of careful selection of texts and recording conditions.

While they cover similar ground, the newer Sproat book is quite different from Dutoit's. It has seven authors, all contributors to the Bell Labs system that is described in detail in the book. Often in multiauthor books, one finds a significant unevenness in coverage and style across chapters due to lack of coordination among the authors. This is much less apparent in Sproat's book, because all authors worked at the same lab and because one author (van Santen) was involved in seven of the chapters; at least one of the two principal authors (Sproat and van Santen) contributed to all nine chapters.

Further distinguishing this book is its emphasis on multilingual TTS: the Bell system exists for ten diverse languages, and the book provides many specific examples of interesting problems in different languages. While several multilingual synthesizers are available commercially, most technical literature has focused on one language at a time. Given that all speech synthesis is based on the same human speech production mechanism and that the world's languages share many aspects of phonetics, it is prudent to examine how a uniform methodology can be applied to many different languages for TTS. Unlike speech coders, which normally function equally well for all languages without adjustment, speech recognizers and synthesizers necessarily need training for individual languages. One of the foci of this book is minimizing the work to

be repeated when developing TTS for another language. While many examples from different languages shown in the book help to demonstrate the Bell Labs system's flexibility, most of the book focuses on English (indeed, the only evaluation is given for English, but their new "gentex" text analysis is not used for English).

The book begins with a very frank foreword by Louis Pols, noting that the Bell Labs method shares with Dutoit's the concatenative approach to TTS by chaining together stored speech units. Pols points out that such a method is not very flexible in terms of simulating new voices, and may have difficulties adding emotions to synthetic voices. The introductory Chapter 1 notes that an ultimate TTS system, one indistinguishable from a human speaker, still awaits significant research in articulatory synthesis; in the meantime, however, more practical systems have been increasingly following the lines of work described in this book.

The casual reader may find difficulty with Chapter 2, which describes statistical methods and finite-state automata that are needed to understand fully much of the detail in ensuing chapters. I would advise many readers to skip to Section 2.6, and to go back to the rest of Chapter 2 only when needed to better understand notation from later chapters. While many later examples need a good understanding of Chapter 2, most of the book can be read well without dwelling too long here.

The longest chapter at 50 pages, Chapter 3 compares the traditional TTS approach (e.g., MITalk) with the Bell Labs method, emphasizing the advantages of the new approach especially for a modular and multilingual system. Very specific and detailed examples from seven languages are given. The important issue of rule-based versus automatic learning is addressed: as in speech recognition (but less so for TTS), there has been a significant trend toward use of automatic training, as opposed to labor-intensive development of system rules by experts. The authors correctly conclude that purely automatic learning systems are doomed to inferior performance; as in recognition, a combination of phonetic and linguistic structure with proper use of statistics will eventually lead to improved results in synthesis.

Chapter 4 covers issues of text analysis dealing with accentuation, disambiguation of homographs, and prediction of boundaries of prosodic phrases. Chapters 5 and 6 cover the assignment of timing and intonation for TTS. It is assumed that one can obtain a reasonably sized inventory of acoustic units to concatenate (e.g., no more than several thousand units, since individual speakers would be stressed to utter more at a time), from which natural-sounding speech of any text is feasible. Issues of discovering which linguistic factors affect intonation and their interactions are discussed in detail. The tone sequence approach (e.g., from Pierrehumbert [1981]) is compared to superposition (e.g., from Fujisaki [Fujisaki, Ljungqvist, and Murata 1993]). The latter is adopted and modified, although the authors admit that applying it to tone languages would be a challenge.

Ways to concatenate acoustic units—from rule-based (e.g., MITalk) to concatenative (e.g., PSOLA [Hamon, Moulines, and Charpentier 1989])—are treated in Chapter 7. The major difficulty with the latter, adopted, method is the large number of units needed to handle all the variations of coarticulation. Proper generation of acoustic unit inventories requires great care in token selection and excision, which is again described in detail.

Formal evaluation methods are examined in Chapter 8, which present problems due to the many variables involved in judging intelligibility and naturalness for TTS with many parameters. The book ends with a useful discussion of the limitations of current technology and directions where future research needs to make progress.

The book has ample indices, both of subjects and authors, and an extensive list of references. Its physical layout makes the book easy to read, and many figures and ta-

bles also help the reader. An accompanying Web site (*http://www.bell-labs.com/project/tts*) provides access to audio versions of the synthetic speech. (Technically, the book's only flaw is an occasional problem in spacing words equally.) In summary, I recommend this book strongly to anyone wanting a detailed description of current ways to do text-to-speech synthesis. It will be most useful to TTS researchers, but people knowledgeable about other technical aspects of speech communication will also find it useful. For a course on speech synthesis, I would recommend a combination of the Sproat and Dutoit books. The Sproat book is more detailed on issues of natural language processing, while Dutoit covers acoustic aspects more comprehensively. These two synthesis books will provide the standards in TTS for the next several years.

References

Allen, Jonathan, M. Sharon Hunnicutt, and Dennis H. Klatt (with Robert C. Armstrong and David Pisoni). 1987. *From Text to Speech: The MITalk System.* Cambridge University Press.

Bailly, Gérard and Christian Benoit (editors). 1992. *Talking Machines: Theories, Models, and Designs.* North-Holland, Amsterdam.

Dutoit, Thierry. 1997. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, Dordrecht.

Fujisaki, Hiroya, Mats Ljungqvist, and H. Murata. 1993. Analysis and modeling of word accent and sentence intonation in Swedish. *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing,* volume II, pages 211–214, Glasgow.

Hamon, C., Eric Moulines, and Francis Charpentier. 1989. A diphone synthesis system based on time-domain prosodic modifications of speech. *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing,* pages 238–241, Glasgow.

Pierrehumbert, Janet. 1981. Synthesizing intonation. *Journal of the Acoustical Society of America,* 70: 985–995.

Douglas O'Shaughnessy has been a professor at INRS–Télécommunications (Université du Québec) for 20 years, working on aspects of speech synthesis, coding, and recognition. A fellow of the Acoustical Society of America and an associate editor of *IEEE Transactions on Speech and Audio Processing,* he is also the author of a textbook entitled *Speech Communications.* O'Shaughnessy's address is: INRS–Télécom, 16 Place du Commerce, Nun's Island, Quebec H3E 1H6, Canada; e-mail: dougo@inrs-telecom.uquebec.ca