An Introduction to Text-to-Speech Synthesis

Thierry Dutoit
(Faculté Polytechnique de Mons)

Dordrecht: Kluwer Academic
Publishers (Text, Speech and Language
Technology series, edited by Nancy Ide
and Jean Véronis, volume 3), 1997,
xxv+285 pp; hardbound, ISBN
0-7923-4498-7, $99.00, Dfl 170.00, £63.00

Reviewed by
Eileen Fitzpatrick
Montclair State University and AT&T Labs—Research

In his final chapter, Dutoit claims that the practical goal of a good textbook is to present a "sort of book-sized classification tree." By this measure, Dutoit has had great success. The book explores every component in the design of a text-to-speech (TTS) system, every design alternative for each component, and the advantages and disadvantages of each alternative. To accomplish this goal in 285 pages is no small task, and it comes at a cost. The presentation is dense; it is not a text for a student audience. It is intended "for those, engineers or linguists, who are trying to develop a complete TTS system" or trying to understand TTS synthesis. Even for this (rather disparate) audience, however, some background in parsing and digital signal processing is assumed.

After a brief introduction to speech production and analysis, and a consideration of the cognitive processing involved when humans read aloud, the book is about equally divided between the natural language processing (NLP) components of a TTS system and the digital signal processing (DSP) components. The NLP chapters include "Grammars, inference, parsing and transduction," "NLP architectures for TTS," "Morpho-syntactic analysis," "Automatic phoneticization," and "Automatic prosody generation." The DSP chapters include "Synthesis strategies," a comparison of rule-based and concatenation-based synthesizers, "Linear prediction synthesis," "Hybrid harmonic/stochastic synthesis," and "Time-domain algorithms." A final summary chapter contains two tables that outline the tasks and costs involved in each NLP and DSP approach considered. The concise summarization of all the alternatives, their advantages and weaknesses, is a real gift to the reader who is using this book to make design choices.

Because the book is aimed at system designers rather than end users, it is devoted exclusively to laboratory systems. The reader should not expect to find a comparison of commercial TTS systems here. As the author notes, industrial products do not allow full control of their speech parameters and the research behind these systems is not in the public domain.

Since the book lays out the strengths and weaknesses of each NLP and DSP alternative within TTS, it seems fair to do the same here. First, the strengths. The book is good at breaking up each area of research into possible alternatives, e.g., different types of parsers, alternative intonation models, alternative synthesis algorithms. It is also excellent at maintaining consistent evaluation criteria by which to compare the alternatives. Each synthesis algorithm, for example, is compared in terms of linear interpolation and smoothing properties, data compression ratios, prosody matching capabilities, computational load, and segmental quality. The consistency would enable

a naive reader to appreciate why a certain algorithm would be advantageous given certain constraints, even if the details required more background in signal processing. This comparison of alternatives is supported in each chapter by an up-to-date bibliography; the book could function as a literature review for TTS.

The main drawback of the book is its density of presentation. There is little space given to an appreciation of the linguistic problems that have generated the alternative solutions. Also, examples are rarely used to illustrate a problem. This makes it difficult to understand the form that alternative solutions take. For instance, while more than a chapter is devoted to parsing, it is never made clear why a full parse might be desirable for TTS if "the prosodic structure of a sentence is flatter than its syntactic one" (p. 148).

There are a couple of areas that the book might have covered to advantage. Deterministic (Marcus) parsing is not considered as a parsing alternative, although it has been used in one of the rule-based prosodic phrasing algorithms discussed here. There is also a noticeable gap regarding the evaluation of naturalness in speech, a quality that is primarily associated here with NLP performance. Naturalness and intelligibility are the two goals that the book sets for TTS. Intelligibility is, of course, the *sine qua non* of TTS and the author includes a good discussion of the tests used to evaluate intelligibility. Naturalness as a goal, on the other hand, deserves some discussion. Is naturalness, for example, an aesthetic criterion or does its absence degrade the listener's ability to understand the message? How do we evaluate naturalness in any way that is not subjective? Does the lack of naturalness serve as a cue to the listener to give the system some slack? Questions like these deserve consideration when putting together a TTS system.

A final quibble: The text would have benefited from more dogged proof-reading. There are a fair number of typos and, more worrisome, some confusing cases of subcategorization and lexical choice errors in English.

One half of the book is going to be rough going for either half of its intended audience, the engineers or the computational linguists. Nevertheless, the determined reader will be well rewarded with a comprehensive overview of the TTS literature and an understanding of how the TTS pieces fit together.

Eileen Fitzpatrick's address is: Department of Linguistics, Montclair State University, Upper Montclair, NJ 07043; e-mail: fitzpatricke@montclair.edu.