Data-Driven Techniques in Speech Synthesis

R. I. Damper (editor)
(University of Southampton)

Boston: Kluwer Academic Publishers,
2001, xviii+316 pp; hardbound, ISBN
0-412-81750-0, $145.00, €148.00, £100.00

Reviewed by
Thierry Dutoit
Faculté Polytechnique de Mons

Never say "never." In 1997, most experts would have sworn that text-to-speech (TTS) synthesis technologies had reached a plateau, from which it would be very hard to leave. Five years later, speech synthesis has been widely and unexpectedly revolutionized by data-driven techniques. Wherever handcrafted rule-based systems were chosen for their incremental design and analytic controllability, machine learning (ML) techniques are now increasingly used for their scalability and genericity, key elements for the design of multilingual, and possibly embedded, TTS systems. The established, "linguist-friendly" paradigm ("if you don't get a substantial quality increase with ML, stick to expert systems") is thus being turned into a more pragmatic strategy ("even if it brings a small percentage of error increase, go for ML"). This 316-page book, edited by Robert I. Damper and written by top specialists in the field, addresses such recent advances in data-driven techniques for speech synthesis, with a very strong emphasis on the use of ML techniques for natural language processing issues (and even more specifically for automatic phonetization).

After Damper's introduction to the architecture of TTS systems in chapter 1, Ghulum Bakiri and Thomas G. Dietterich open a series of seven chapters devoted to automatic grapheme-to-phoneme (GTP) transcription. Their chapter 2, "Constructing High-Accuracy Letter-to-Phoneme Rules with Machine Learning," examines extensions to NetTalk and NetSpeak, the pioneering (but rather deceiving) work of Sejnowski and Rosenberg. They point out how, by modifying the original multilayer perceptron, it is possible to reach better transcription rates than those possible using established rule-based systems. In chapter 3, "Analogy, the Corpus and Pronunciation," Kirk P. H. Sullivan presents the idea of pronunciation-by-analogy and its relation to a psychological model of oral reading. The chapter ends with a (somewhat confused) discussion of an implementation of the Sullivan and Damper method for English, Maori, and German. Helen Meng examines the use of probabilistic formal grammars for phonetizing words in chapter 4, "A Hierarchical Lexical Representation for Pronunciation Generation." Based on a multilevel linguistic description of words that is obtained with a handcrafted context-free grammar, the method attaches probabilities to sibling-sibling transitions in the rules of the parser. Chapter 5, "English Letter–Phoneme Conversion by Stochastic Transducers," by Robert W. P. Luk and Robert I. Damper, is devoted to the use of stochastic finite-state transducers for GTP conversion in English, a hot but complex topic. After a discussion on maximum-likelihood transduction and on possible ways of achieving automatic GTP alignment (a prerequisite for most GTP transcription systems), it is shown that the best results are obtained when a priori linguistic information is used for alignment. This chapter is dense and thus not truly self-contained.

Sabine Deligne, François Yvon, and Frédéric Bimbot focus on their multigram approach in chapter 6, "Selection of Multiphone Synthesis Units and Grapheme-to-

Phoneme Transcription Using Variable-Length Modeling of Strings," for estimating the probability of a string seen as the concatenation of (automatically derived) independent variable-length sequences of symbols. After presenting the classical multigram approach and its extension to joint multigrams (i.e., on several nonsynchronized streams of symbols), the authors propose two applications for TTS synthesis: that of deriving the set of most frequently needed multiphone units for the design of a concatenative speech synthesis system (which obviously deserves further investigation) and that of performing joint multigram-based GTP conversion. Lazy, or memory-based learning is the subject of chapter 7, "TREETALK: Memory-Based Word Phonemisation," by Walter Daelemans and Antal Van den Bosch. The authors present "normal" lazy learning (IB1-IG), their information-theoretic IGTree-building technique, and a hybrid TRIBL method for optimizing transcription speed while maintaining low error rates. The chapter ends with an analytic discussion on the use of monolithic versus modular GTP systems and surprisingly shows that the best results are obtained when the intermediate levels are left implicit. Chapter 8, "Learnable Phonetic Representations in a Connectionist TTS system—Text to Phonetics," by Andrew D. Cohen, concludes this GTP-oriented part of the book, with a journey into the land of nonsegmental phonology. Departing from the traditionally phoneme-oriented interface between GTP and speech synthesis, a more phonetic interface is examined, which is moreover obtained in an unsupervised way by training a combination of neural networks on a database composed of words in their written and oral forms. The machine itself proposes phonetic units, in the form of attractor basins in a self-organizing map. This chapter, together with chapter 12 by the same author, is certainly one of the most complex and experimental of the book (together they constitute a dense summary of the author's doctoral dissertation).

The four last chapters explore, although to a much lesser extent, the use of data-driven approaches for prosody generation and speech signal synthesis. Chapter 9, "Using the Tilt Intonation Model," by Alan W. Black, Kurt E. Dusterhoff, and Paul A. Taylor, summarizes the authors' Tilt model of intonation. After presenting the easy F0-to-Tilt and Tilt-to-F0 pathways, it is shown that classification and regression trees (CARTs) can do a good job when asked to decide the value of Tilt parameters, using a linguistic prediction feature set. In Chapter 10, "Estimation of Parameters for the Klatt Synthesizer from a Speech Database," John Coleman and Andrew Slater provide a "Klatt synthesizer primer" in which they show how to synthesize high-quality, formant-based English sounds by using automatic acoustic analysis of real speech combined with "tricks of the trade." In Chapter 11, "Training Accent and Phrasing Assignment on Large Corpora," Julia Hirschberg summarizes the use of CART techniques for predicting accent and phrasing assignment (a prerequisite for intonation and duration generation); the method is based on the Pierrehumbert hierarchical description of intonation. The author gives analytic results on several databases (citation-form sentences, news stories by a single speaker, multispeaker broadcast radio and multispeaker spontaneous speech) and obtains results comparable to those derived from a handcrafted rule-based system. The chapter ends with experiments on using text corpora annotated by native speakers in place of time-consuming speech corpora, which make it possible to train models in a (small) fraction of the time needed in the original speech-based training. The book concludes with a short proposal, chapter 12, for extending the ideas of Cohen's first chapter to concatenative speech signal synthesis itself. Cohen proposes a complex combination of neural networks for producing sequences of linear predictive coding (LPC) coefficients and F0 values from the output of his unsupervised GTP system.

I read this book with great pleasure and undoubtedly learned from it. I have no doubt that postgraduate students and researchers in the area will benefit from its

reading. It should be clear, however, that prior exposure to neural networks, statistical language modeling, and finite-state models is required to take full advantage of the book, especially for chapters 5–8 and 12. Although most of the material presented in this book appears elsewhere (the authors of each chapter are also their main protagonists and have thus already published their work in various journal papers), it has been given a compact and comprehensive form here.

The book inevitably suffers from "edited book syndrome." The introductions of the first seven chapters tend to have strong overlaps, and the chapters in general contain only few cross-references. Not all chapters are of equal interest for the same person. Researchers will be more interested in chapters 3, 5, and 6, whereas system designers will probably prefer chapters 7, 9, and 11. On the other hand, chapters can be read in virtually any order (except for chapter 1, which should be read first, and chapter 12, which assumes prior reading of chapter 8).

The reader always wants more: One would certainly have loved to get test data, and example training and testing scripts in an included CD-ROM, especially since the authors discuss their own work. More comparative results (possibly as an "add-on" chapter) would have been welcome too. But as judiciously mentioned by several authors, it is not easy to compare technologies with different training hypotheses and testing procedures.

This raises an additional, and maybe broader, question (in the sense that it addresses the field of data-driven GTP in general): Is speech synthesis (and most particularly GTP conversion) seen as a test bed for ML techniques, or is it considered the problem to solve? When comparing systems, most authors emphasize the pros and cons of the underlying technologies (and comment on their possible extensions to various areas), whereas the title of the book somehow suggests a task-oriented approach. Readers who expect the book to provide keys to designing a full data-driven TTS system will be disappointed by the more scientific and prospective considerations they will find. Those interested in having a clearer picture of ML techniques, tested here on speech synthesis problems, will be rewarded.

One last but important caveat: This book surprisingly contains only partial information on data-driven prosody generation and very little information on what seems to be the hottest topic in the TTS industry these last years: data-driven concatenative speech signal synthesis (sometimes referred to as nonuniform unit (NUU) synthesis). Maybe the title is misleading in that respect: The book is actually strongly biased toward language modeling and even more toward GTP conversion.

Summarizing, this book is clearly a must for post-graduate students and researchers in the area of data-driven phonetization. It is the first to propose in-depth, state-of-the-art information on the topic and to offer a comparative view of the underlying technologies. It therefore brings a fresh perspective to this quickly moving field. It can also be used as a pointer to other aspects of data-driven speech synthesis (namely, prosody and speech signal synthesis), although the reader should be aware that these are only very incompletely covered.

Thierry Dutoit has been a professor of circuit theory, signal processing, and speech processing at Faculté Polytechnique de Mons in Belgium since 1993. Between 1996 and 1998, he spent 16 months at AT&T–Bell Labs in New Jersey. He is the initiator of the MBROLA speech synthesis project, the author of a reference book on speech synthesis (in English), and the coauthor of a book on speech processing (in French). He has written or cowritten about 60 papers on speech processing and software engineering. Dutoit is also involved in industrial activities as a consultant. Dutoit's address is Faculté Polytechnique de Mons, MULTITEL-TCTS Lab, Initialis Scientific Park, Avenue Copernic, B-7000 Mons, Belgium; e-mail: thierry.dutoit@fpms.ac.be; URL: tcts.fpms.ac.be/∼dutoit.