Review of "Data-Driven Techniques in Speech Synthesis"

(R.I. Damper, ed.)

By T. Dutoit, Faculté Polytechnique de Mons

Never say « never ».
In 1997, most experts would have sworn that text-to-speech synthesis (TTS) technologies had come to a plateau, from which it would be very hard to depart. Five years later, speech synthesis have 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").

The 316 pp. book edited by R.I. Damper and co-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 (yet even more specifically for automatic phonetization).

After a necessary introduction to the architecture of TTS systems in Chapter One, G. Bakiri and T. Dietterich open a series of seven chapters devoted to automatic grapheme-to-phoneme (GTP) transcription. Their Chapter Two examines extensions to NetTalk and NetSpeak, the pioneering (but rather deceiving) work of Sejnowski and Rosenberg. They point out how, by modifying the original multi-layer perceptronit is possible to reach better transcription rates than established rule-based systems. In Chapter Three, K. Sullivan exposes the pronunciation-by-analogy ideas, and their 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. H. Meng examines the use of probabilistic formal grammars for phonetizing words, in Chapter Four. Based on a multi-level linguistic description of words, obtained with a handcrafted context-free grammar, the method attaches probabilities to sibling-sibling transitions in the rules of the parser. Chapter Five, by Luk and 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 pre-requisite for most GTP transcription systems), it is shown that the best results are obtained when *a priori* linguistic (VC) information is used for alignment. This chapter is dense, thus not truly self-contained. S. Deligne and F. Bimbot focus on their multigram approach in Chapter Six, for estimating the probability of a string seen as the concatenation of (automatically derived) independent variable-length sequences of symbols. After exposing the classical multigram approach and its extension to joint-multigrams (i.e., on several non-synchronized 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 Seven, by W. Daelemans and A. Van den Bosch. The authors expose 'normal' lazy learning (IB1-IG), their information-theoretic IGTree building technique based on work by Quinlan, Lucassen and Mercer, 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 vs. modular GTP systems, and surprisingly shows that the best results are obtained when the intermediate levels are left implicit.
Chapter Eight, by A. Cohen, concludes this GTP-oriented part of the book, with a journey in non-segmental phonology land. 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 (after Kohonen's work). This chapter, together with Chapter 12 by the same author is certainly one of the most complex and experimental of the book (it is a dense summary of A. Cohen's PhD).

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 Nine, by A. Black, K.

Dusterhoff and P. Taylor, summarizes the authors' Tilt model of intonation. After exposing the easy F0-to-Tilt and Tilt-to-F0 pathways, it is shown that classification and regression tress (CARTs) can do a good job when asked to decide for the value of Tilt parameters, based on a linguistic prediction feature set. In Chapter Ten, J. Coleman and A. 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". J. Hirschberg summarizes, in Chapter Eleven, the use of CART techniques for predicting accent and phrasing assignment (a prerequisite for intonation and duration generation), 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, multi-speaker 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 in Chapter Twelve, with a short proposal for extending the ideas of Chapter 8 to concatenative speech signal synthesis itself. A. Cohen proposes a complex combination of neural networks for producing sequences of 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 post-graduate 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 Five to Eight, and Twelve.

Although most of the material exposed here 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 in this book.

The book inevitably suffers, from the "edited book syndrome". The introductions of the seven first chapters tend to have strong overlaps, while 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 Three, Five and Six, while system designers will probably prefer chapters Seven, Nine, and Eleven. On the other hand, chapters can be read in virtually any order (except for Chapter One, 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 a bonus CDROM, especially since the authors discuss their own work. More comparative results (possibly as an additional "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 brings 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 as the problem to solve? When comparing systems, most authors emphasize on the pro's and con's of the underlying technologies (and comment on their possible extensions to various areas), while 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: that of data driven concatenative speech signal synthesis (sometimes referred to as "non-uniform unit, or NUU, synthesis)". Maybe the title is misleading in that respect: the book is actually strongly biased towards language modeling, and even more towards 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 partially covered.

T. Dutoit is a professor of Circuit Theory, Signal Processing, and Speech Processing at Faculté Polytechnique de Mons, Belgium, since 1993. Between 1996 and 1998, he spent 16 months at AT&T-Bell Labs, in Murray Hill (NJ) and Florham Park (NJ). He is the initiator of the MBROLA speech synthesis project, the author of a reference book on speech synthesis (in English), and the co-author of a book on speech processing (in French). He wrote or co-wrote about 60 papers on speech processing and software engineering. T. Dutoit is also involved in industrial activities as a consultant for Babel Technologies, S.A., Multitel ASBL, and IT-Optics, S.A.