Book Notice
Taylor, Paul. 2009. *Text-to-Speech Synthesis.* Cambridge: Cambridge University Press.*

Rania Al-Sabbagh
University of Illinois at Urbana-Champaign
alsabba1@illinois.edu

Text-to-Speech synthesis is the artificial production of human speech typically used for assistive technology applications like screen readers and voice output communication aids, in addition to entertainment productions such as games, anime and similar applications. Taylor's (2009) *Text-to-Speech Synthesis* provides a complete, end-to-end account of this speech synthesis technology with in-depth explanations of all its aspects, both theoretical—chapters 1 to 12—and technical—chapters 13 to 17. In addition, this volume sheds light on future directions for text-to-speech synthesis technology.

Chapter 1 is an overview of the book defining text-to-speech systems (hereafter TTS) and highlighting the dual goals of any TTS system, namely intelligibility and naturalness. Although the book primarily focuses on the practical, engineering aspects of TTS, it gives necessary theoretical background in the field and is therefore suitable for both academic and commercial audiences.

Chapter 2 lays out the theoretical foundations set by the writer, defining the concept of communication in the framework of TTS and what the components of language are. It discusses the basic properties of language, the nature of signal, form meaning and the four main processes of generating, encoding, decoding and understanding. In the field of TTS, communication is a semiotic system of three levels of representation: meaning, form and signal; communication can be further classified into affective, iconic and symbolic communication with the latter being the main concern of TTS. This is because symbolic communication is eligible for discrete representations and is thus computationally manageable. From a TTS perspective, the two main components of language are the verbal component and the prosodic component. The verbal component is comprised of phonemes, words and sentences, represented by a discrete symbolic system as a finite number of units that can generate an enormous number of messages. The prosodic component is by contrast continuous,

* 626 pp. Hardback (ISBN-13: 978-0-521-89927-7), $99.00

Studies in the Linguistic Sciences: Illinois Working Papers **2010**: 33-36

and is used either to express speakers' emotion (affective prosody) or to disambiguate and reinforce the verbal component (augmentative prosody).

Chapter 3 focuses on the text-to-speech problem in more detail and the mapping between the written form and the spoken form of the text. The relation between both forms is not a one-to-one relation: the written message, for instance, is meant to be read silently with more focus on the message itself; written messages, therefore, do not encode some features that are essential for a human-like speech synthesizer, namely prosodic features. Challenges of this type are the main concern of chapter 3.

Extracting linguistic information from text input is the main concern of Chapters 4, 5 and 6. Chapter 4 is basically about preprocessing issues like how to find whole sentences in running text and how to handle markup or control information. Chapter 5 focuses on the main processes of text analysis itself with special focus on resolving homograph ambiguity. Finally, Chapter 6 describes how prosodic information can be predicted from an improvised text input given the absence of such information from the written text form. This is among the major challenges of TTS given that there are not clear criteria for evaluating prosodic prediction.

Chapters 7, 8 and 9 are more solidly situated in linguistics, providing basic information about phonetics, phonology and prosody. For readers with a background in linguistics, these chapters might serve as a refreshing review. For commercial readers and software developers with little or no linguistic knowledge, however, the chapters are a satisfactory introduction to the linguistic notions of phonemes, vowels, consonants, phonological features and stress, syllables, phonotactics, intonation, prosody, pitch, timing and the grapheme-to-phoneme conversion process wherein a word's pronunciation is predicted based on its spelling.

The techniques of signals and filters together with basic terminology and core equations for computing periodic signals, the frequency domain, digital signals, transforms and digital filters are all introduced in chapter 10; a more formal quantitative model of speech production is given in chapter 11. Lossless tubes are the primary model under discussion, with the vocal organs being represented through a set of discrete interconnected components in a system where each component functions as either a source component or a filter.

Although the book primarily deals with speech synthesis, Chapter 12 deals with speech analysis because many recent techniques for speech synthesis rely on an analysis phase. Three main problems of speech analysis are discussed: eliminating phase, separating source and filter, and

transforming the representation into a space that has more desirable properties. In addition to explaining each problem in detail, the chapter notes recent trends in resolving such problems.

Chapter 13 contains a historical background on the "first-generation" techniques that dominated the field in the 1980s. Discussing these gives better understanding of why today's systems are configured the way they are. This in turn provides the reader with the sufficient background to understand why, as an example, today's dominant technique is unit-selection rather than the more basic approach which generates speech waveforms from scratch. These historical techniques are also of some use today for applications that require small footprints and low processing cost. These techniques mainly include formant synthesis, classical linear prediction, and articulatory synthesis. The basic limitations of these older approaches are their inability to generate natural-sounding speech and the method of determining which parameters to use for a given synthesis specification by hand-written rules. Although manually-crafted rules can produce fairly intelligible speech, they are computationally expensive for real-world applications.

To overcome these limitations, Chapter 14 introduces a recent set of techniques; these are known as the "second-generation" synthesis systems. The main difference between the two generations is that the second-generation systems rely more heavily on data-driven techniques. The standard set-up in a second-generation system is to have a specification composed of a sequence of items such that each item contains a phone description, an F0 and duration for that phone. The phone description is then matched to the data and various techniques are used to modify its pitch and duration. These second-generation systems do encounter a new set of major challenges: modifying the pitch and timing without introducing any unwanted side effects and the problem of data sparseness where it is never possible collect enough data to cover all the effects to be synthesized.

Data-driven approaches are divided into concatenative and statistical, machine-learning approaches. Concatenative approaches are limited to re-creation where the only available option is reordering the original data. Thus, instead of *memorizing* the data as in concatenative approaches, statistical, machine learning approaches *learn* from the data. The most commonly used statistical, machine learning techniques within the framework of TTS are the Hidden Markov Models (HMMs)—covered in Chapter 15—and the unit-selection techniques covered in Chapter 16. These two approaches are typically referred to as the "third-generation techniques".

A number of diverse issues are summarized in Chapter 17, which deals with available databases, evaluation methodologies, generating emotion and audio-visual synthesis. In brief, the chapter surveys available resources for TTS and highlights still-unresolved challenges. Future directions for such challenges are more thoroughly discussed in the concluding chapter, Chapter 18.

In summary, the book is indispensable reading for students of linguistics, electrical engineering and computer science who are interested in text-to-speech synthesis. It requires no specialized prior knowledge of the field, although it assumes a fundamental level of computer literacy and programming experience including the concepts of algorithm, variable, loop and so forth.

REFERENCES

Taylor, Paul. 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.