Neural Text-to-Speech Synthesis
Artificial Intelligence: Foundations, Theory and Algorithms fosters the dissemination of knowledge, technologies and methodologies that advance developments in artificial intelligence (AI) and its broad applications. It brings together the latest developments in all areas of this multidisciplinary topic, ranging from theories and algorithms to various important applications. The intended readership includes research students and researchers in computer science, computer engineering, electrical engineering, data science, and related areas seeking a convenient way to track the latest findings on the foundations, methodologies, and key applications of artificial intelligence.

This series provides a publication and communication platform for all AI topics, including but not limited to:

- Knowledge representation
- Automated reasoning and inference
- Reasoning under uncertainty
- Planning, scheduling, and problem solving
- Cognition and AI
- Search
- Diagnosis
- Constraint processing
- Multi-agent systems
- Game theory in AI
- Machine learning
- Deep learning
- Reinforcement learning
- Data mining
- Natural language processing
- Computer vision
- Human interfaces
- Intelligent robotics
- Explanation generation
- Ethics in AI
- Fairness, accountability, and transparency in AI

This series includes monographs, introductory and advanced textbooks, state-of-the-art collections, and handbooks. Furthermore, it supports Open Access publication mode.
Xu Tan

Neural Text-to-Speech Synthesis

Springer
I would like to dedicate this book to my family, especially my grandfather and grandmother.
Foreword by Dong Yu

Text-to-speech (TTS) synthesis is an artificial intelligence (AI) technique that renders a preferably naturally sounding speech given an arbitrary text. It is a key technological component in many important applications, including virtual assistants, AI-generated audiobooks, speech-to-speech translation, AI news reporters, audible driving guidance, and digital humans. In the past decade, we have observed significant progress made in TTS. These new developments are mainly attributed to deep learning techniques and are usually referred to as neural TTS. Many neural TTS systems have achieved human quality for the tasks they are designed for.

Although many TTS books have been published, this book is the first of its kind that provides a comprehensive introduction to neural TTS, including but not limited to the key components such as text analysis, acoustic model, and vocoder, the key milestone models such as Tacotron, DeepVoice, FastSpeech, and the more advanced techniques such as expressive and controllable TTS, robust TTS, and efficient TTS. Xu Tan, the author of this book, has contributed significantly to the recent advances in TTS. He has developed several impactful neural TTS systems such as FastSpeech 1/2, DelightfulTTS, and NaturalSpeech, the latter of which has achieved human parity on the TTS benchmark dataset. His knowledge of the domain and his first-hand experience with the topic allow him to organize the contents effectively and make them more accessible to readers, and to describe the key concepts, the basic methods, and the state-of-the-art techniques and their relationships in detail and clearly. I am very glad that he introduced and clarified many key concepts and background knowledge at the beginning of this book so that people with little or no knowledge of TTS can also read and understand the book effectively.

This is a very well-written book and certainly one that provides useful and thoughtful information to readers at various levels. I believe this book is a great reference book for all researchers, practitioners, and students who are interested in
quickly grasping the history, the state-of-the-art, and the future directions of speech synthesis or are interested in gaining insightful ideas on the development of TTS.

ACM/IEEE/ISCA Fellow
Seattle, WA, USA
October, 2022
For more than 60 years, researchers have studied how to synthesize natural-sounding speech. The focus of speech synthesis research shifted from simulating the physical process of speech production (articulatory speech synthesis) to concatenating basic speech units (concatenative speech synthesis) in the late 1970s. The “unit-selection” method of concatenative synthesis, which was invented in the late 1980s, could synthesize highly natural and intelligible speech on some subset of the sentences. However, because of its “exemplar-based” nature, synthesizing expressive speech, which is essential to achieve human-to-human-level interactions, was still difficult.

In the mid-1990s, a promising new method for speech synthesis, called statistical parametric speech synthesis, was proposed. This method is based on generative models, where statistical models represent the conditional probability distribution of output speech given an input text. Although it offered high intelligibility and flexibility to synthesize a variety of speech, it had limited success until the 2000s due to its inferior naturalness. Thanks to the advancement of signal processing techniques and machine learning in the late 2000s, its quality improved significantly. Furthermore, the introduction of deep learning/neural networks as its generative models in the 2010s completely changed the landscape of text-to-speech research; synthetic speech sounds not only natural and intelligible but also expressive and beyond. Nowadays most of the components in a text-to-speech synthesis system are realized by neural networks. Such text-to-speech synthesis systems are often referred to as “neural text-to-speech”.

This book, Neural Text-to-Speech Synthesis, is for people who want to understand how modern deep learning/neural network-based text-to-speech synthesis systems are implemented and how they have progressed from traditional concatenative and statistical parametric speech synthesis systems to recent integrated, neural end-to-end text-to-speech systems. Xu Tan, who is one of the leading researchers in neural text-to-speech synthesis, has put together a book based on the progress in neural text-to-speech synthesis over the past decade. The first three chapters of the book give the basics of spoken language processing and deep learning for text-to-speech synthesis. The next three chapters address the problem of converting an input text to a speech waveform in neural text-to-speech synthesis, including both a
cascaded approach consisting of three key modules (text analysis, acoustic models, and vocoders) and an integrated, end-to-end approach. The following four chapters provide a review of advanced topics, which are important to deploy neural text-to-speech synthesis systems to real-world applications, such as data and computational efficiency, controllability, expressiveness, and robustness. This is followed by a chapter describing the relationship with the three related speech synthesis areas. The last chapter gives the concluding remarks and possible future research directions.

This book will be a great help for readers to understand the landscape of neural text-to-speech synthesis and to explore new frontiers in text-to-speech synthesis research.

ISCA Fellow 
Tokyo, Japan
December, 2022

Heiga Zen
Text-to-speech synthesis (TTS) enables machines to speak naturally and express human emotions. With the advent of deep learning, the quality of synthesized speech has improved by leaps and bounds.

This is a book like no other. It is the first to provide a comprehensive overview that covers the breadth and depth of neural TTS. It was written by my research friend, Xu Tan, the architect of the widely adopted FastSpeech TTS system. Xu Tan is known as a prominent researcher, a prolific author, and a hands-on engineer. I am fortunate to be the first to read this book, and I must say that this book has fulfilled my longstanding wish of having a more accessible guidebook for neural TTS.

This book provides a unique, historical, and technological perspective on the recent development of neural TTS, that I resonate with. It is very timely given the increasing interest in the research community. It is a book that research students, TTS beginners, and practitioners cannot miss.

IEEE/ISCA Fellow
Shenzhen, China
November, 2022

Haizhou Li
Preface

Speaking is one of the most important language capabilities (the others being listening, reading, and writing) of human beings. Text-to-speech synthesis (TTS for short), which aims to generate intelligible and natural speech from text, plays a key role to enable machines to speak and is an important task in artificial intelligence and natural language/speech processing.

With the development of deep learning and artificial intelligence, neural network-based TTS has significantly improved the quality of synthesized speech in recent years. Considering neural TTS involves multiple disciplines such as speech signal processing and deep learning, and there are abundant of literature in this area. It is very challenging for TTS practitioners especially beginners to understand the landscape of neural TTS. Thus, there is a growing demand for a book dedicated for neural text-to-speech synthesis.

I started my research on neural TTS several years ago. During this period, I was deeply impressed by the rapid progress of neural TTS brought by whole speech synthesis community. Then I started to plan and prepare for this book two years ago. However, it is not an easy thing considering the diverse methodologies and abundant literature in this area. Thus, I divide this difficult job into multiple stages: (1) first give a tutorial on neural TTS at ISCSLP 2021 conference; (2) then write a survey paper on neural speech synthesis based on this tutorial; (3) further enrich the previous tutorial and survey paper gradually and give tutorials at IJCAI 2021, ICASSP 2022, and INTERSPEECH 2022 conferences; (4) and finally write this book based on these survey and tutorials.¹

This book gives a comprehensive introduction to neural TTS, aiming to provide a good understanding of its basic methods, current research, and future trends. I first introduce the background of speech synthesis and the history of TTS technologies. I then introduce some preliminary knowledge of neural TTS in the first part of this

¹ Please find all these survey and tutorials in this page: https://github.com/tts-tutorial/. Readers can use this page to check updates and initiate discussions on this book: https://github.com/tts-tutorial/book.
book, including the basics of spoken language processing and deep learning. In the second part, I introduce the key components of neural TTS, including text analyses, acoustic models, and vocoders. I further introduce several advanced topics of neural TTS in the third part, including expressive and controllable TTS, robust TTS, model-efficient TTS, data-efficient TTS, and some tasks beyond TTS. At last, I summarize this book and discuss future research directions. I also list some resources related to TTS (e.g., TTS tutorials and talks, open-source implementations, and datasets) in the appendix.

This book is written for researchers, industry practitioners, and graduate/undergraduate students in speech synthesis, speech/language processing, and artificial intelligence.

Beijing, China
October, 2022

Xu Tan
Acknowledgements

This book would not have been possible without the contributions of many people.

I would like to thank my colleagues and interns at Microsoft, who have been working together with me on the topic of neural text-to-speech synthesis, including Tao Qin, Sheng Zhao, Tie-Yan Liu, Lei He, Frank Soong, Hsiao-Wuen Hon, Lidong Zhou, Jiang Bian, Qiang Huo, Jun-Wei Gan, Yanqing Liu, Bohan Li, Yi Ren, Renqian Luo, Rui Wang, Kaitao Song, Xi Wang, Gang Wang, Jinzhu Li, Yuanhao Yi, Ruiqing Xue, Runnan Li, Dongxu Han, Xianghao Tang, Yuchao Zhang, Peter Pan, Chen Zhang, Jie Ding, Yangjun Ruan, Chenxu Hu, Jin Xu, Mingjian Chen, Hao Sun, Yichong Leng, Kai Shen, Zeqian Ju, Sang-gil Lee, Zehua Chen, Haohe Liu, Jian Cong, Jiawei Chen, Yuzi Yan, Guangyan Zhang, Yihan Wu, Jian Luan, Peiling Lu, Junliang Guo, Chen Zhang, Qi Meng, and Chang Liu. I would also like to thank my external collaborators including Zhou Zhao, Ruihua Song, Jian Li, Kejun Zhang, Yuan Shen, Wei-Qiang Zhang, Tan Lee, Guihua Wen, Sungroh Yoon, and Danilo Mandic.

I would like to thank Heiga Zen, Haizhou Li, and Dong Yu for providing suggestions and forewords to this book. I want to make special thanks to Heiga Zen for giving me so much help to improve this book. I also want to thank those who gave me permission to directly use or reproduce images/figures from their publications.

At last, I want to thank the people in the whole speech synthesis community for pushing forward TTS technologies.
Contents

1 Introduction ................................................................. 1
  1.1 Motivation ............................................................. 1
  1.2 History of TTS Technology ......................................... 2
    1.2.1 Articulatory Synthesis ..................................... 3
    1.2.2 Formant Synthesis .......................................... 3
    1.2.3 Concatenative Synthesis .................................... 3
    1.2.4 Statistical Parametric Synthesis ......................... 4
  1.3 Overview of Neural TTS ........................................... 5
    1.3.1 TTS in the Era of Deep Learning ....................... 5
    1.3.2 Key Components of TTS ................................... 5
    1.3.3 Advanced Topics in TTS .................................. 6
    1.3.4 Other Taxonomies of TTS .................................. 6
    1.3.5 Evolution of Neural TTS ................................... 7
  1.4 Organization of this Book ......................................... 9

References ..................................................................... 10

Part I Preliminary

2 Basics of Spoken Language Processing .............................. 17
  2.1 Overview of Linguistics ............................................ 18
    2.1.1 Phonetics and Phonology .................................. 18
    2.1.2 Morphology and Syntax ................................... 19
    2.1.3 Semantics and Pragmatics ................................. 20
  2.2 Speech Chain .......................................................... 20
    2.2.1 Speech Production and Articulatory Phonetics .......... 21
    2.2.2 Speech Transmission and Acoustic Phonetics .......... 24
    2.2.3 Speech Perception and Auditory Phonetics .......... 26
  2.3 Speech Signal Processing .......................................... 28
    2.3.1 Analog-to-Digital Conversion .............................. 28
    2.3.2 Time to Frequency Domain Transformation .......... 30
    2.3.3 Cepstral Analysis .......................................... 31
### 3 Basics of Deep Learning

3.1 Machine Learning Basics ........................................ 37
3.1.1 Learning Paradigms ......................................... 37
3.1.2 Key Components of Machine Learning .................... 40
3.2 Deep Learning Basics ........................................... 40
3.2.1 Model Structures: DNN/CNN/RNN/Self-attention ...... 40
3.2.2 Model Frameworks: Encoder/Decoder/
    Encoder-Decoder .............................................. 43
3.3 Deep Generative Models ......................................... 45
3.3.1 Autoregressive Models ...................................... 46
3.3.2 Normalizing Flows ........................................... 47
3.3.3 Variational Auto-encoders .................................. 49
3.3.4 Denoising Diffusion Probabilistic Models ............... 51
3.3.5 Score Matching with Langevin Dynamics, SDEs, and ODEs ........................................... 53
3.3.6 Generative Adversarial Networks .......................... 55
3.3.7 Comparisons of Deep Generative Models ................. 56

References .................................................................. 59

---

### 4 Text Analyses

4.1 Text Processing .................................................... 67
4.1.1 Document Structure Detection ............................... 68
4.1.2 Text Normalization ......................................... 68
4.1.3 Linguistic Analysis ........................................... 69
4.2 Phonetic Analysis .................................................. 70
4.2.1 Polyphone Disambiguation ................................ 71
4.2.2 Grapheme-to-Phoneme Conversion ....................... 71
4.3 Prosodic Analysis .................................................. 71
4.3.1 Pause, Stress, and Intonation .............................. 72
4.3.2 Pitch, Duration, and Loudness .......................... 72
4.4 Text Analysis from a Historic Perspective ................. 73
4.4.1 Text Analysis in SPSS ..................................... 73
4.4.2 Text Analysis in Neural TTS .............................. 74

References .................................................................. 74

---

### 5 Acoustic Models

5.1 Acoustic Models from a Historic Perspective ............. 81
5.1.1 Acoustic Models in SPSS .................................... 82
5.1.2 Acoustic Models in Neural TTS ........................... 83
8.1.4 Recording Devices or Noise Environments .......................... 127
8.2 Modeling Variation Information for Expressive Synthesis ........ 127
  8.2.1 Explicit or Implicit Modeling ..................................... 127
  8.2.2 Modeling in Different Granularities ............................. 129
8.3 Modeling Variation Information for Controllable Synthesis ......... 129
  8.3.1 Disentangling for Control ......................................... 130
  8.3.2 Improving Controllability ....................................... 130
  8.3.3 Transferring with Control ...................................... 131
References ........................................................................ 131

9 Robust TTS .................................................................. 141
  9.1 Improving Generalization Ability ................................. 142
  9.2 Improving Text-Speech Alignment ................................. 143
    9.2.1 Enhancing Attention ............................................ 143
    9.2.2 Replacing Attention with Duration Prediction ............ 145
  9.3 Improving Autoregressive Generation ............................ 146
    9.3.1 Enhancing AR Generation ..................................... 147
    9.3.2 Replacing AR Generation with NAR Generation ......... 147
References ........................................................................ 148

10 Model-Efficient TTS .................................................... 153
  10.1 Parallel Generation .................................................. 154
    10.1.1 Non-Autoregressive Generation with CNN or
           Transformer ...................................................... 154
    10.1.2 Non-Autoregressive Generation with GAN, VAE, or
           Flow .............................................................. 155
    10.1.3 Iterative Generation with Diffusion ......................... 155
  10.2 Lightweight Modeling .............................................. 156
    10.2.1 Model Compression ............................................ 156
    10.2.2 Neural Architecture Search ................................ 156
    10.2.3 Other Technologies ........................................... 157
  10.3 Efficient Modeling with Domain Knowledge .................... 157
    10.3.1 Linear Prediction .............................................. 157
    10.3.2 Multiband Modeling .......................................... 157
    10.3.3 Subscale Prediction ........................................... 157
    10.3.4 Multi-Frame Prediction ....................................... 158
    10.3.5 Streaming or Chunk-Wise Synthesis ....................... 158
    10.3.6 Other Technologies ........................................... 158
References ........................................................................ 158

11 Data-Efficient TTS ..................................................... 163
  11.1 Language-Level Data-Efficient TTS ............................. 164
    11.1.1 Self-Supervised Training .................................... 164
    11.1.2 Cross-Lingual Transfer ...................................... 165
    11.1.3 Semi-Supervised Training .................................. 165
    11.1.4 Mining Dataset in the Wild ................................. 165
## Contents

### 11.1.5 Purely Unsupervised Learning

165

### 11.2 Speaker-Level Data-Efficient TTS

166

11.2.1 Improving Generalization

167

11.2.2 Cross-Domain Adaptation

167

11.2.3 Few-Data Adaptation

168

11.2.4 Few-Parameter Adaptation

168

11.2.5 Zero-Shot Adaptation

169

References

169

### 12 Beyond Text-to-Speech Synthesis

175

12.1 Singing Voice Synthesis

175

12.1.1 Challenges in Singing Voice Synthesis

176

12.1.2 Representative Models for Singing Voice Synthesis

176

12.2 Voice Conversion

177

12.2.1 Brief Overview of Voice Conversion

177

12.2.2 Representative Methods for Voice Conversion

177

12.3 Speech Enhancement/Separation

178

References

178

### Part IV Summary and Outlook

183

### 13 Summary and Outlook

183

13.1 Summary

183

13.2 Future Directions

183

13.2.1 High-Quality Speech Synthesis

184

13.2.2 Efficient Speech Synthesis

185

References

185

### A Resources of TTS

187

### B TTS Model List

191

References

194
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Autoregressive Flow</td>
</tr>
<tr>
<td>AM</td>
<td>Acoustic Model</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>BAP</td>
<td>Band Aperiodicities</td>
</tr>
<tr>
<td>BFCC</td>
<td>Bark-Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>CMOS</td>
<td>Comparative Mean Opinion Score</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>CTC</td>
<td>Connectionist Temporal Classification</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>Diffusion or DDPM</td>
<td>Denoising Diffusion Probabilistic Model</td>
</tr>
<tr>
<td>DNN</td>
<td>Dense Neural Network</td>
</tr>
<tr>
<td>DTFT</td>
<td>Discrete-Time Fourier Transform</td>
</tr>
<tr>
<td>DTW</td>
<td>Dynamic Time Warping</td>
</tr>
<tr>
<td>E2E</td>
<td>End-to-End</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental Frequency</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>Flow</td>
<td>Normalizing Flow</td>
</tr>
<tr>
<td>G2P</td>
<td>Grapheme to Phoneme</td>
</tr>
<tr>
<td>GAN</td>
<td>Generative Adversarial Network</td>
</tr>
<tr>
<td>GRU</td>
<td>Gated Recurrent Unit</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IAF</td>
<td>Inverse Autoregressive Flow</td>
</tr>
<tr>
<td>IPA</td>
<td>International Phonetic Alphabet</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
<tr>
<td>LSP</td>
<td>Line Spectral Pairs</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long Short-Term Memory</td>
</tr>
<tr>
<td>MCC</td>
<td>Mel-Cepstral Coefficients</td>
</tr>
<tr>
<td>MCD</td>
<td>Mel-Cepstral Distortion</td>
</tr>
<tr>
<td>MFA</td>
<td>Montreal Forced Alignment</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>MGC</td>
<td>Mel-Generalized Coefficients</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>NAR</td>
<td>Non-Autoregressive</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse-Code Modulation</td>
</tr>
<tr>
<td>PESQ</td>
<td>Perceptual Evaluation of Speech Quality</td>
</tr>
<tr>
<td>POS</td>
<td>Part of Speech</td>
</tr>
<tr>
<td>PPG</td>
<td>Phonetic Posteriorgrams</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>SDR</td>
<td>Signal-to-Distortion Ratio</td>
</tr>
<tr>
<td>SMOS</td>
<td>Similarity Mean Opinion Score</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Parametric Speech Synthesis</td>
</tr>
<tr>
<td>STFT</td>
<td>Short-Time Fourier Transform</td>
</tr>
<tr>
<td>STOI</td>
<td>Short-Time Objective Intelligibility</td>
</tr>
<tr>
<td>SVS</td>
<td>Singing Voice Synthesis</td>
</tr>
<tr>
<td>TN</td>
<td>Text Normalization</td>
</tr>
<tr>
<td>ToBI</td>
<td>Tones and Break Indices</td>
</tr>
<tr>
<td>TTS</td>
<td>Text-to-Speech Synthesis</td>
</tr>
<tr>
<td>VAE</td>
<td>Variational Auto-Encoder</td>
</tr>
<tr>
<td>VC</td>
<td>Voice Conversion</td>
</tr>
<tr>
<td>ZCR</td>
<td>Zero-Crossing Rate</td>
</tr>
</tbody>
</table>