SpringerBriefs in Speech Technology

Studies in Speech Signal Processing, Natural Language Understanding, and Machine Learning

Series Editor:
Amy Neustein

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, standardized manuscript preparation and formatting guidelines, and expedited production schedules.

The goal of the **SpringerBriefs in Speech Technology** series is to serve as an important reference guide for speech developers, system designers, speech engineers and other professionals in academia, government and the private sector. To accomplish this task, the series will showcase the latest findings in speech technology, ranging from a comparative analysis of contemporary methods of speech parameterization to recent advances in commercial deployment of spoken dialog systems.

More information about this series at http://www.springer.com/series/10043

K. Sreenivasa Rao • N. P. Narendra

Source Modeling Techniques for Quality Enhancement in Statistical Parametric Speech Synthesis

Springer

K. Sreenivasa Rao
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal, India

N. P. Narendra
Aalto University
Espoo, Finland

ISSN 2191-737X ISSN 2191-7388 (electronic)
SpringerBriefs in Speech Technology
ISBN 978-3-030-02758-2 ISBN 978-3-030-02759-9 (eBook)
https://doi.org/10.1007/978-3-030-02759-9

Library of Congress Control Number: 2018959748

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Speech is the most natural way for humans to communicate with each other. Synthesis of artificial human speech provides efficient human-computer communication. Nowadays, the speech synthesis systems are widely used in various applications such as screen readers for visually challenged people, speech interface for mobile devices, navigation, and personal guidance gadgets. As humans are very sensitive in perceiving even the slightest distortions in the speech signal, speech synthesizers with suboptimal quality make them unfit for usage in commercial applications. The main goal of this book is to improve the quality of statistical parametric speech synthesis (SPSS) by efficiently modeling the source or excitation signal. The excitation signal used in synthesis should preserve all natural variations so that the synthesized speech is close to natural quality. The work presented in this book confines its scope to the (1) accurate estimation of pitch (F_0) and (2) precise modeling of excitation signal. For modeling the excitation signal, both parametric and hybrid approaches are explored. In this work, creaky voice has been synthesized at appropriate places by proposing appropriate methods and models.

The contents of the book are useful for both researchers and system developers. For researchers, the book will be useful for knowing the current state-of-the-art excitation source models for SPSS and further refining the source models to incorporate the realistic semantics present in the text. For system developers, the book will be useful to integrate the sophisticated excitation source models mentioned in the book to the latest models of mobile/smart phones. The book has been organized as follows:

- Chapter 1 introduces the topic of text-to-speech synthesis. Different speech synthesis approaches are briefly discussed.
- Chapter 2 provides a review of the state-of-the-art methods for F0 estimation and parametric and hybrid source modeling approaches.
- Chapter 3 discusses the design of a voicing detection and F_0 estimation method by adaptively choosing appropriate window size for zero-frequency filtering method.
- Chapter 4 presents two parametric source modeling methods.

- Chapter 5 describes two proposed hybrid methods of modeling the excitation signal.
- Chapter 6 deals with the generation of creaky voice by addressing two main issues, namely, automatic detection of creaky voice and source modeling of creaky voice.
- Chapter 7 summarizes the contributions of the book along with some important conclusions. Directions toward the scope for possible future work are also discussed.

We would especially like to thank all professors of the Department of Computer Science and Engineering, IIT Kharagpur, for their consistent support during the course of editing and organization of the book. Special thanks to our colleagues at Indian Institute of Technology, Kharagpur, India, for their cooperation to carry out the work. We are grateful to our parents and family members for their constant support and encouragement. Finally, we thank all our friends and well-wishers.

Kharagpur, India K. Sreenivasa Rao
Espoo, Finland N. P. Narendra

Contents

Acronyms

AC	AutoCorrelation
AMDF	Average Magnitude Difference Function
APP	Aperiodicity, Periodicity and Pitch
CART	Classification And Regression Tree
CC	Cross Correlation
CD	Continuous probability Distribution
CMOS	Comparative Mean Opinion Scores
CRD	Cumulative Relative Dispersion
CSTR	Centre for Speech Technology Research
CW	Characteristic Waveform
DNN	Deep Neural Networks
DSM	Deterministic plus Stochastic Model
EGG	ElectroGlottoGraph
EM	Expectation-Maximization
ESPS	Entropic Signal Processing System
F	Female
FPE	Fine Pitch Error
FPR	False Positive Rate
GCI	Glottal Closure Instants
GPE	Gross Pitch Error
HE	Hilbert Envelope
HMM	Hidden Markov Model
HNR	Harmonic-to-Noise Ratio
HTS	HMM-based Speech Synthesis System
Hz	Hertz
IAIF	Iterative Adaptive Inverse Filtering
IFAS	Instantaneous Frequency Amplitude Spectrum
IFP	IntraFrame Periodicity
IPS	InterPulse Similarity
KB	KiloByte
KD	Kane-Drugman

kHz	Kilo Hertz
LF	Liljencrants-Fant
LP	Linear Prediction
LPC	Linear Predictive Coding
LSD	Log-Spectral Distance
LSF	Line-Spectral Frequencies
M	Male
MB	MegaByte
MBROLA	Multi-Band Resynthesis OverLap and Add
MDL	Minimum Description Length
MELP	Mixed Excitation Linear Prediction
MFCC	Mel Frequency Cepstral Coefficients
MGC	Mel-Generalized Cepstrum
MGLSA	Mel-Generalized Log Spectral Approximation
MLSA	Mel Log Spectral Approximation
ms	Milli Seconds
MSD	Multi-Space prabability Distribution
MSE	Maximum Strength of Excitation
NN	Neural Networks
NPI	Next Phone Identity
PCA	Principal Component Analysis
PI	Phone Identity
PP	Position of Phrase
PPI	Previous Phone Identity
PS	Position of Syllable
PSOLA	Pitch Synchronous OverLap and Add
PW	Position of Word
RAPT	Robust Algorithm for Pitch Tracking
RTSE	Relative Time Squared Error
SHS	SubHarmonic Summation
SPSS	Statistical Parametric Speech Synthesis
SRH	Summation of Residual Harmonics
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
TD-PSOLA	Time-Domain Pitch Synchronous Overlap and Add
TPR	True Positive Rate
UV	Unvoicing
V	Voicing
VDE	Voicing Decision Error
WI	Waveform Interpolation
ZFF	Zero-Frequency Filtering
ZFFHE	ZFF with Hilbert Envelope
ZFFUW	ZFF with Uniform Window
ZFR	Zero-Frequency Resonator