

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

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

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 F_0 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
Espoo, Finland

K. Sreenivasa Rao
N. P. Narendra

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Speech Synthesis Methods	2
1.2.1	Formant Synthesis	2
1.2.2	Articulatory Synthesis	3
1.2.3	Concatenative Synthesis	3
1.2.4	Statistical Parametric Speech Synthesis	4
1.2.5	Hybrid Synthesis Methods	5
1.3	Objectives and Scope of the Work	5
1.4	Contributions of the Book	6
1.4.1	Robust Voicing Detection and F_0 Estimation Method	6
1.4.2	Parametric Approach of Modeling the Excitation Signal	7
1.4.3	Hybrid Approach of Modeling the Excitation Signal	7
1.4.4	Generation of Creaky Voice	7
1.5	Organization of the Book	8
	References	9
2	Background and Literature Review	11
2.1	HMM-Based Speech Synthesis	11
2.1.1	Hidden Markov Model	12
2.1.2	System Overview	14
2.1.3	Duration Modeling	15
2.1.4	Decision Tree-Based Context Clustering	15
2.1.5	Synthesis	16
2.2	Voicing Detection and F_0 Estimation: A Review	16
2.3	Source Modeling Approaches: A Review	19
2.4	Generation of Creaky Voice: A Review	22
2.5	Summary	24
	References	25

3	Robust Voicing Detection and F_0 Estimation Method	29
3.1	F_0 Modeling and Generation in HTS	29
3.2	Proposed Method for Voicing Detection and F_0 Estimation	30
3.2.1	Zero-Frequency Filtering Method for Detecting the Instants of Significant Excitation	31
3.2.2	Voicing Detection	32
3.2.3	Influence of Window Size on the Strength of Excitation	34
3.2.4	F_0 Estimation	38
3.2.5	Performance Evaluation	40
3.3	Implementation of the Proposed Voicing Detection and F_0 Extraction in HTS Framework	43
3.4	Evaluation	45
3.4.1	Evaluation of Voicing Detection	46
3.4.2	Subjective Evaluation	48
3.5	Summary	50
	References	51
4	Parametric Approach of Modeling the Source Signal	53
4.1	Parametric Source Modeling Method Based on Principal Component Analysis	53
4.1.1	Generation of Pitch-Synchronous Residual Frames	53
4.1.2	Parameterization of Residual Frame Using PCA	55
4.1.3	Speech Synthesis Using the Proposed PCA-Based Parametric Source Model	57
4.1.4	Evaluation	58
4.2	Parametric Source Modeling Method Based on the Deterministic and Noise Components of Residual Frames	60
4.2.1	Analysis of Characteristics of Residual Frames	60
4.2.2	Overview of Proposed Parametric Source Model	62
4.2.3	Parameterization of Deterministic Component	63
4.2.4	Parameterization of Noise Component	64
4.2.5	Speech Synthesis Using the Proposed Deterministic and Noise Component-Based Parametric Source Model	66
4.2.6	Evaluation	67
4.3	Summary	71
	References	73
5	Hybrid Approach of Modeling the Source Signal	75
5.1	Optimal Residual Frame-Based Hybrid Source Modeling Method	75
5.1.1	Computation of Optimal Residual Frame for a Phone	75
5.1.2	Clustering the Optimal Residual Frames	77
5.1.3	Speech Synthesis Using the Proposed Optimal Residual Frame-Based Hybrid Source Model	78
5.1.4	Evaluation	80

5.2	Time-Domain Deterministic Plus Noise Model-Based Hybrid Source Model	82
5.2.1	Decomposition of Deterministic and Noise Components	88
5.2.2	Clustering the Deterministic Components	88
5.2.3	Storing Natural Instance of Noise Signal Along with the Deterministic Component	89
5.2.4	Parameterizing the Noise Components	90
5.2.5	Speech Synthesis Using the Proposed Time-Domain Deterministic Plus Noise Model-Based Hybrid Source Model	91
5.2.6	Evaluation	93
5.2.7	Discussion	100
5.3	Summary	102
	References	103
6	Generation of Creaky Voice	105
6.1	HMM-Based Speech Synthesis System for Generating Modal and Creaky Voices	105
6.2	Automatic Detection of Creaky Voice	107
6.2.1	Analysis of Epoch Parameters	108
6.2.2	Computation of Variance of Epoch Parameters	110
6.2.3	Classification Using Variance of Epoch Parameters	111
6.2.4	Performance Evaluation	111
6.3	Hybrid Source Model for Generating Creaky Excitation	115
6.3.1	Generation of Creaky Residual Frames	117
6.3.2	Deterministic Plus Noise Decomposition for Every Phonetic Class	118
6.3.3	Synthesis of Creaky Voice Using Proposed Hybrid Source Model	119
6.3.4	Evaluation	120
6.3.5	Discussion	122
6.4	Summary	123
	References	123
7	Summary and Conclusions	125
7.1	Summary of the Book	125
7.2	Contributions of the Book	127
7.3	Directions for Future Work	128
	References	128
	Index	131

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 probability 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