

Instrumental Quality Estimation for Synthesized Speech Signals

Dissertation

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

(Dr.-Ing.)

der Technischen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Christoph Robert Norrenbrock

Kiel 2013

1. Berichtersteller: Prof. Dr.-Ing. habil. Ulrich Heute
2. Berichtersteller: Prof. Dr.-Ing. habil. Sebastian Möller

Datum der mündlichen Prüfung: 23.01.2014

Arbeiten über Digitale Signalverarbeitung

Band 38

Christoph Robert Norrenbrock

**Instrumental Quality Estimation
for Synthesized Speech Signals**

Shaker Verlag
Aachen 2014

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Kiel, Univ., Diss., 2014

Copyright Shaker Verlag 2014

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-2669-6

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Danksagung

Die vorliegende Dissertation habe ich während meiner Tätigkeit als wissenschaftlicher Mitarbeiter in der Arbeitsgruppe Digitale Signalverarbeitung und Systemtheorie (DSS), vormals Lehrstuhl für Netzwerk- und Systemtheorie (LNS), an der Christian-Albrechts-Universität zu Kiel angefertigt. Sie entstand im Wesentlichen im Rahmen des Drittmittelprojekts „Instrumentelle Schätzung der Qualität synthetisierter Sprachsignale“, welches durch die Deutsche Forschungsgemeinschaft (DFG) finanziert worden ist.

Mein besonderer Dank gilt meinem Doktorvater Herrn Prof. Dr.-Ing. Ulrich Heute. Er hat mir durch die Möglichkeit zur Mitarbeit in seiner Forschungsgruppe großes Vertrauen entgegengebracht. Seine außerordentliche Unterstützung und Diskussionsbereitschaft, gepaart mit seinem unbefangenen Blick auf Inhalt und orthographische Form, haben wesentlich zu einem glücklichen Arbeitsverhältnis beigetragen. Dabei habe ich den mir gewährten Freiraum stets sehr zu schätzen gewusst.

Des Weiteren möchte ich Herrn Prof. Dr.-Ing. Sebastian Möller dafür danken, dass er als Gutachter meiner Dissertation fungiert hat. Die durch Engagement, Professionalität und Freundschaft geprägte Zusammenarbeit im Rahmen des genannten DFG-Projekts hat mir stets viel Freude bereitet. In diesem Zuge möchte ich mich herzlich bei meinem lieben Doktorandenkollegen Florian Hinterleitner bedanken. Die enge Zusammenarbeit, dokumentiert durch die gemeinsamen Veröffentlichungen, Konferenzreisen, Projekttreffen, und nicht zuletzt die zahlreichen Telefonate und Diskussionen, hat mir sehr geholfen meinen Weg durch das Thema „Qualität“ zu bahnen, und dabei die eigenen Sichtweisen und Lösungen regelmäßig zu hinterfragen und zu verbessern.

Herzlich bedanken möchte ich mich bei meinen ehemaligen Studenten, die durch ihre Abschlussarbeit oder ihre wissenschaftliche Hilfstätigkeit zu der vorliegenden Arbeit beigetragen haben, namentlich Dipl.-Wirtsch.-Ing. Frank Heydasch, Dipl.-Wirtsch.-Ing. Friedemann Köster und M.Sc. Befkadu Temesgen Gebru.

Allen aktuellen und ehemaligen Kollegen des LNS, der Arbeitsgruppe DSS und der Arbeitsgruppe Informations- und Codierungstheorie (ICT) möchte ich für die

schöne Zeit als Teil dieses Kollegiums danken. Insbesondere möchte ich mich bei Dipl.-Ing. Rebecca Adam, Dipl.-Ing. Roman Kreimeyer und Dipl.-Wirtsch.-Ing. Viet Duc Nguyen bedanken für die vielen Stunden ausdauernden Diskutierens und Korrigierens. Obwohl das Promovieren selbst letzten Endes durch einen begrenzten zwischenmenschlichen Interaktionsgrad gekennzeichnet ist, so ist das Arbeitsumfeld zweifelsohne von besonderer Bedeutung, um die gemeinsame Freude an der Forschung zu zelebrieren, sich in gegenseitigem Respekt zu üben, den hohen thematischen Ansprüchen gerecht zu werden, und nicht zuletzt um die Motivations- und Willenskraft zu fördern und aufrecht zu erhalten. Die großartige LNS/DSS/ICT-Truppe hat hierzu die vorzüglichsten Eigenschaften an den Tag gelegt und ich bereue es keinen Augenblick mich in diesem angenehmen Umfeld, und mit frischem Espresso versorgt, dem Abenteuer Promotion gewidmet zu haben.

Ein ganz besonderer Dank gilt meinen Freundinnen und Freunden. Die vielfältigen gemeinsamen Aktivitäten und Erlebnisse – deren Erläuterung hier den Rahmen sprengen würde – haben maßgeblich dazu beigetragen, dass ich die Zeit an der Kieler Förde in bester Erinnerung behalten werde.

Abschließend möchte ich mich bei meiner Familie bedanken, die mich in allen Lebenslagen so umfassend unterstützt hat, wie ich es mir nicht besser hätte wünschen können. Ein besonderes Andenken bewahre ich meinem Alten Herrn und geliebten Vater Dr. med. Peter Norrenbrock, verstorben am 18. Januar 2014.

Kiel, im Februar 2014

Christoph R. Norrenbrock

Contents

Acronyms and Notation	v
1 Introduction	1
1.1 Motivation	1
1.2 Thesis outline	3
2 Fundamentals of Text-to-Speech Synthesis	5
2.1 Overview of artificial speech generation	5
2.2 Overview of text-to-speech synthesis	6
2.2.1 Procedural perspective	6
2.2.2 Analytic perspective	7
2.3 TTS-system types	9
2.3.1 Formant synthesis	10
2.3.2 Concatenative synthesis	10
2.3.3 Unit-selection synthesis	12
2.3.4 HMM synthesis	13
2.3.5 Hybrid synthesis	14
2.4 Summary of Chapter 2	14
3 Quality Assessment of Synthesized Speech	15
3.1 Fundamentals of speech quality	15
3.1.1 Quality perception	15
3.1.2 From individual perception to objective speech quality	17
3.1.3 Quality reference	18
3.2 Taxonomy of speech-quality assessment for synthesized speech	20
3.3 Auditory methods	22
3.4 Instrumental methods	23

3.4.1	Discontinuities and acoustic-distance join costs	24
3.4.2	Prosodic quality	26
3.4.3	Intelligibility	26
3.4.4	Integral quality	27
3.5	Discussion and research positions	30
3.6	Summary of Chapter 3	31
4	Concepts of Non-intrusive Quality Assessment (NiQA)	33
4.1	General modelling approach	33
4.2	Evaluation of a NiQA model	34
4.2.1	Plain maximum-likelihood approach	37
4.2.2	Bayesian approach	38
4.2.3	Measurand-consistency approach	38
4.2.4	Hierarchical approach	39
4.2.5	Multidimensional approach	39
4.2.6	Discussion	40
4.3	Regular-perception approach	40
4.3.1	Two-stage approach for NiQA	41
4.3.2	Theory of regular perception	42
4.3.3	Perceptual regularization	45
4.4	Perceptual regularization for synthetic speech	47
4.5	Summary of Chapter 4	49
5	Synthetic Speech Material and Perceptual Analysis	51
5.1	Synthetic speech databases	51
5.1.1	Test I	51
5.1.2	Test II	53
5.1.3	Test III	54
5.1.4	Audiobook test	55
5.1.5	Natural reference signals	55
5.1.6	Scales and scaling	56
5.2	Quality space of TTS	56
5.2.1	Quality dimensions	57
5.2.2	Signal examples	58
5.2.3	Discussion	61

5.3	Summary of Chapter 5	62
6	Signal-based Properties of Synthetic-Speech Quality	63
6.1	Remarks on speech-signal analysis	63
6.2	Prosodic properties	64
6.2.1	Intonation	65
6.2.2	Voice	67
6.2.3	Rhythm	70
6.3	Articulatory properties	71
6.3.1	MFCCs	71
6.3.2	Vocal-tract approximation	73
6.3.3	Modulation spectrum	77
6.4	Summary of Chapter 6	80
7	Prediction Models: Development, Regression, and Assessment	81
7.1	Perceptual-regularization modes	82
7.2	Regression models	83
7.2.1	Remarks on regression modelling	83
7.2.2	Discussion	86
7.2.3	Partial-least-squares (PLS) regression	87
7.2.4	Support-vector regression (SVR)	87
7.2.5	Regular perception model (RPM)	89
7.3	Model assessment	90
7.3.1	General cross-validation setup	91
7.3.2	Leave-one-test-out CV (LOTO)	92
7.3.3	Leave-one-configuration-out CV (LOCO)	92
7.3.4	Feature normalization	93
7.3.5	Feature selection	94
7.4	Discussion	96
7.5	Summary of Chapter 7	97
8	Prediction Models: Results	99
8.1	Overview of the results	99
8.2	Significance testing	100
8.3	General trends	100
8.3.1	Influence of dimension	100

8.3.2	Influence of voice gender	101
8.3.3	Influence of perceptual regularization	102
8.3.4	Influence of property group	102
8.3.5	Influence of model type	103
8.3.6	Results for LOCO CV	104
8.3.7	Top measurands	105
8.4	Discussion	106
8.4.1	CV setups	106
8.4.2	Linear vs. nonlinear	106
8.4.3	Perceptual reference	107
8.4.4	Towards the assessment of time-variant quality	108
8.4.5	Validity and scope of application	109
8.5	Summary of Chapter 8	110
9	Summary and Future Work	121
A	Regression Algorithms	125
A.1	Partial-least-squares regression	125
A.2	Support-vector regression	126
	Bibliography	129